Big Data Analytics: A New Perspective*

A. Chudik
Federal Reserve Bank of Dallas

G. Kapetanios
King's College London, University of London

M. Hashem Pesaran
USC Dornsife INET, and Trinity College, Cambridge

February 11, 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 has a number of advantages over the 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 computationally simple and considerably faster, and it performs better in small samples for almost all of the five different sets of experiments considered in this paper. Despite its simplicity, the theory behind the proposed approach is quite complicated. We provide extensive theoretical and Monte Carlo results in support of adding the proposed OCMT model selection procedure to the toolbox of applied researchers.

Keywords: One covariate at a time, multiple testing, model selection, high dimensionality, penalised regressions, boosting, Monte Carlo experiments
JEL Classifications: C52, C55

*We are grateful to Zemin Zheng for providing us with the Matlab codes for the Lasso, Sica and Hard thresholding penalised regression methods used in this paper. We have also benefited from helpful comments by Jinchi Lv and Yingying Fan. The views expressed in this paper are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Dallas or the Federal Reserve System.

1 Introduction

The problem of correctly specifying a model has been a major and 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.

The recent advent of large datasets has made this specification task much harder. In particular, the reality of having datasets where the number of potential regressors for a given regression model can be of the same or larger order of magnitude compared to the number of observations, has spurred considerable advances in statistical and econometric methodology. Large datasets are becoming increasingly available in a number of areas. In macroeconomics, an ever-increasing set of indicators and surveys are used to inform policy makers in central banks and other policy-making institutions. In microeconomics, data sets cover thousands of firms or individuals observed over space and time and across many different characteristics. Even when the number of available covariates is relatively small, researchers rarely know the exact functional form with which these variables enter the regression model, and they might be interested in including non-linear transformations of the available covariates, such as interaction terms, which lead to a much larger set of covariates to be considered. A general discussion of high-dimensional data and their use in microeconomic analysis can be found in Belloni, Chernozhukov, and Hansen (2014a).

Model selection and estimation in this high-dimensional regression setting has largely settled around a set of methods collectively known as penalised 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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \left[\sum_{t=1}^{T} (y_t - \boldsymbol{x}_{nt}' \boldsymbol{\beta})^2 + P(\boldsymbol{\beta}, \boldsymbol{\lambda}) \right],$$

in which $P(\boldsymbol{\beta}, \boldsymbol{\lambda})$ 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 $P(\boldsymbol{\beta}, \boldsymbol{\lambda})$ 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, 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 Frank and Friedman (1993), 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 with a particular focus on the desirability of its convexity and the choice of the tuning parameter(s). 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 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) and Friedman (2001)). Boosting constructs a regression function by considering all regressors one by one in a simple regression setting and successively selecting the best fitting ones. More details on boosting algorithms for linear models and their theoretical properties can be found in Buhlmann (2006).

Boosting may also be viewed within the context of stepwise regression methods which are methods that overlap, and to some extent predate, greedy methods. In stepwise regression the choice of regressors is based on an automatic testing procedure. Two main approaches are common: Forward selection involves successively adding variables based on which variable has the highest t-statistic in absolute value when added to the regression, while backward elimination starts with a model that contains all variables and successively removes variables based again on the relevant t-statistics. An early reference is Hocking (1976). Stepwise regression does not seem to have been rigorously analysed, as it has mainly been used in practical and empirical contexts.

Related to stepwise regression, recent work by David Hendry and various co-authors has used a variant of backward elimination for model specification. This is referred to as the 'General-to-Specific' model specification methodology, see Hendry and Krolzig (2005). This methodology has been applied to a variety of problems. More recently, it has been applied to break detection as detailed in Doornik, Hendry, and Pretis (2013) and Hendry, Johansen, and Santos (2008). Again, the approach does not seem to have been rigorously examined

from a statistical point of view especially when the number of available regressors is allowed to diverge.

A further approach that has a number of common elements with our proposal and combines penalised regression with greedy algorithms has been put forward by Fan and Lv (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 selects 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. In a second step, all statistically significant covariates are included as joint determinants of y_t in a multiple regression setting. In some exceptional cases it might also be required to iterate on 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. But, it is shown that asymptotically the number of such additional iterations will be less than the number of true covariates explaining y_t. Whilst the initial regressions of our procedure are common to boosting and to the screening approach of Fan and Lv (2008), the multiple testing element provides a powerful stopping rule without needing to resort to model selection or penalised regression subsequently.

In short, instead of considering all or sub-sets of the covariates together, we consider 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. In addition to its theoretical properties which we shall discuss below, OCMT is computationally simple and fast even for extremely large datasets, unlike penalised regression which presents some computational challenges in such cases. The method is extremely effective 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 but, unlike the penalised regressions, 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.

Despite its simplicity, the theory behind the proposed approach is quite complicated. We provide extensive theoretical results for the proposed OCMT procedure under assumptions that compare favourably in terms of their general applicability to those made in the analysis of penalised regressions. In particular, we do not assume either a fixed design or time series independence for $\boldsymbol{x}_{nt}$ but consider a milder martingale difference condition. While the martingale difference condition is our maintained assumption, we also provide theoretical arguments for alternative variants of the main method that allow the covariates to follow mixing processes that include autoregressive schemes as special cases.

We establish conditions under which the pseudo-true model (to be defined below) is selected with probability approaching 1 and derive oracle type properties for Euclidean norms of the estimated coefficients of the selected model and its in-sample errors. Under slightly milder conditions, we also establish the consistency of the variable selection procedure in consistently recovering the support of the true regression model. More specifically, we establish conditions under which True Positive Rate and False Positive Rate of our proposed variable selection procedure are 1 and 0, respectively, with probabilities tending to 1.

We also 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 five different data generating schemes. The results clearly highlight the advantages of the OCMT procedure as compared to penalised regressions, with convex and nonconvex penalty functions, as well as to boosting techniques. We also show that the OCMT approach is reasonably robust to non-Gaussian innovations and, to a lesser extent, to serially correlated covariates. Finally, we provide some evidence on the relative computational time of the different methods considered and show that the proposed procedure is about 10^2 and 10^4 times faster than penalised regressions with convex and nonconvex penalty functions, respectively, and about 50 times faster than boosting.

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 concludes and technical proofs are relegated to appendices. Two online supplements provide additional theoretical results and Monte Carlo results for all the experiments conducted.

2 The Variable Selection Problem

Suppose that the target variable, y_t, is generated from the following standard sparse linear regression equation, to be referred to as the DGP (data generating process)

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

where k is small relative to T, u_t is an error term whose properties will be specified below, and $0 < |\beta_i| \leq C < \infty$, for $i = 1, 2, ..., k$. However, 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 $i = k+1, k+2, ..., n$. We do not require the regressors to be normalised, in contrast with penalised regression, where normalisation of regressors affects the selection outcome. 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

$$y_t = a + \sum_{i=1}^{n} I(\beta_i \neq 0)\beta_i x_{it} + u_t. \tag{2}$$

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

$$\theta_i = \sum_{j=1}^{n} I(\beta_j \neq 0)\beta_j \sigma_{ji} = \sum_{j=1}^{k} \beta_j \sigma_{ji}, \tag{3}$$

where $\sigma_{ji} = cov\left(x_{jt}, x_{it}\right)$. 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 $I(\beta_i \neq 0)\beta_i$ 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 $\beta_i = -\sum_{j=1,j\neq i}^{k} \beta_j \sigma_{ji}$. On the other hand, the third case of noise variables with nonzero net effect is quite likely.

For the noise variables, we require their net effects on the target variable to be bounded, which can be formalized by the absolute summability condition, $\sum_{j=k+1}^{n} |\theta_j| < K < \infty$. However, such a condition is too generic to be of use for deriving results and is specialised in a few ways. The first and main assumption is that there exist further k^* variables for which $\theta_i \neq 0$. We shall refer to these noise variables as "pseudo-signal" variables since they are correlated with the signal variables and hence 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. The remaining $n - k - k^*$ variables will be assumed to have $\theta_i = 0$ and 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 $k^*/n \to 0$, at a sufficiently slow rate. 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.

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 $\theta_i = K_i \varrho^i$, $|\varrho| < 1$, $i = 1, 2, ..., n$. A second specification modifies the decay rate and assumes that $\theta_i = K_i i^{-\gamma}$, for some $\gamma > 0$. In both specifications $\max_{1\leq i\leq n} |K_i| < K < \infty$. These specifications allow for various decays in the way noise variables are correlated with the signals. This cases are of technical interest and cover the autoregressive type designs considered in the literature (Zhang (2010) and Belloni, Chernozhukov, and Hansen (2014b)) to model the correlations across the covariates.

As discussed in the Introduction, 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 introduce our alternative approach which is loosely inspired by the multiple testing literature, although here we focus on correct identification of the signal variables rather than controlling the size of the union of the multiple tests that are being carried out.

Notation

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 $n \times \ell$ 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, $\boldsymbol{\tau}_T = (1, 1, ..., 1)'$. $O(.)$ and $o(.)$ denote the Big O and Little o notations, respectively. 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 $\sup_{n \geq N_0}(f_n/g_n) \leq C_1$. Notation $\to_p$ denotes convergence in probability, and $\to_d$ denotes convergence in distribution.

3 A Multiple Testing Approach

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

$$y_t = c_i + \phi_i x_{it} + e_{it},\ t = 1, 2, ..., T. \tag{4}$$

Denote the t-ratio of ϕ_i in this simple regression by $t_{\hat{\phi}_i}$, and note that

$$t_{\hat{\phi}_i} = \frac{\hat{\phi}_i}{s.e.\left(\hat{\phi}_i\right)} = \frac{T^{-1/2}\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{y}}{\hat{\sigma}_i\sqrt{\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{x}_i/T}}, \tag{5}$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})'$, $\boldsymbol{y} = (y_1, y_2, ..., y_T)'$, $\hat{\phi}_i = (\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{x}_i)^{-1}\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{y}$, $\hat{\sigma}_i^2 = \boldsymbol{e}_i'\boldsymbol{e}_i/T$, $\boldsymbol{e}_i = \boldsymbol{M}_{i,\tau}\boldsymbol{y}$, $\boldsymbol{M}_{i,\tau} = \boldsymbol{I}_T - \boldsymbol{X}_{i,\tau}(\boldsymbol{X}_{i,\tau}'\boldsymbol{X}_{i,\tau})^{-1}\boldsymbol{X}_{i,\tau}'$, $\boldsymbol{X}_{i,\tau} = (\boldsymbol{x}_i, \boldsymbol{\tau}_T)$, $\boldsymbol{M}_\tau = \boldsymbol{I}_T - \boldsymbol{\tau}_T\boldsymbol{\tau}_T'/T$, and $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of ones.

Remark 1 *If other deterministic terms, besides the constant, were considered they would be included in the definition of the orthogonal projection matrix $\boldsymbol{M}_\tau$ that filters out these effects. Similarly, if some variables were a priori known to be signals, then they would also be included in the definition of $\boldsymbol{M}_\tau$. The multiple testing method can easily accommodate both possibilities.*

The multiple testing estimator of $I(\beta_i \neq 0)$ is given by

$$\widehat{I(\beta_i \neq 0)} = I\left[\left|t_{\hat{\phi}_i}\right| > c_p(n)\right], \text{ for } i = 1, 2, ..., n, \tag{6}$$

where $c_p(n)$ is a "critical value function" defined by

$$c_p(n) = \Phi^{-1}\left(1 - \frac{p}{2f(n)}\right). \tag{7}$$

$\Phi^{-1}(.)$ is the inverse function of the cumulative standard normal distribution. $f(n)$ can take a variety of forms depending on modelling needs but we will consider mainly

$$f(n) = n^{\delta}, \tag{8}$$

for $0 < \delta < \infty$. p $(0 < p < 1)$ is the nominal size of the individual tests to be set by the investigator.

Remark 2 *The choice of the critical value function, $c_p(n)$, given by (7)-(8), is important since it allows the investigator to relate the size and power of the selection procedure to the inferential problem in the classical statistics, with this 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 $n = O(T^{\kappa})$, for some $\kappa > 0$. This follows from result (i) of Lemma 1, which establishes that $c_p(n) = O\left\{[\ln(n)]^{1/2}\right\}$. Note also that $c_p(n) = o\left(T^{C_0}\right)$, for all $C_0 > 0$, if there exists $\kappa > 0$ such that $n = O(T^{\kappa})$.*

Covariates for which $\widehat{I(\beta_i \neq 0)} = 1$ are selected as signals or pseudo-signals. Denote the number of selected covariates by $\hat{k} = \sum_{i=1}^{n} \widehat{I(\beta_i \neq 0)}$. 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

$$\tilde{\beta}_i = \begin{cases} \hat{\beta}_i^{(\hat{k})}, & \text{if } \widehat{I(\beta_i \neq 0)} = 1 \\ 0, & \text{otherwise} \end{cases}, \text{ for } i = 1, 2, ..., n, \tag{9}$$

where $\hat{\beta}_i^{(\hat{k})}$ 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.

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 the true positive rate (TPR), and the false positive rate (FPR) defined by

$$TPR_{n,T} = \frac{\sum_{i=1}^{n} I\left[\widehat{I(\beta_i \neq 0)} = 1 \text{ and } \beta_i \neq 0\right]}{\sum_{i=1}^{n} I(\beta_i \neq 0)}, \tag{10}$$

$$FPR_{n,T} = \frac{\sum_{i=1}^{n} I\left[\widehat{I(\beta_i \neq 0)} = 1, \text{ and } \beta_i = 0\right]}{\sum_{i=1}^{n} I(\beta_i = 0)}. \tag{11}$$

We also consider the Euclidean norms of the parameter estimation errors, $\tilde{\beta}_i - \beta_i$, and the in-sample regression errors defined by

$$E\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| = E\sqrt{\sum_{i=1}^{n}\left(\tilde{\beta}_i - \beta_i\right)^2},$$

and

$$F_{\tilde{u}} = E\left(\frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2\right),$$

where

$$\tilde{u}_t = y_t - \hat{a}^{(k)} - \textstyle\sum_{i=1}^{n}\tilde{\beta}_i x_{it} = y_t - \hat{a}^{(k)} - \tilde{\boldsymbol{\beta}}_n' \boldsymbol{x}_{nt}, \tag{12}$$

$\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)'$, $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_n)'$, and $\hat{a}^{(k)}$ is the OLS estimator of the constant term in the final regression.

We consider the following assumptions:

Assumption 1 *(a) The error term in DGP (1), u_t, is a martingale difference process with respect to $\mathcal{F}_{t-1}^u = \sigma(u_{t-1}, u_{t-2}, ...,)$. In addition, u_t has zero mean and a constant variance, $0 < \sigma^2 < C < \infty$. (b) 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 2 *(a) Slope coefficients of the true regressors in DGP (1), β_i, for $i = 1, 2, ..., k$, are bounded constants different from zero. (b) Net effect coefficients, θ_i, defined by (3) are nonzero for $i = 1, 2, ..., k$.*

Assumption 3 *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*

$$\sup_{i,t}\Pr(|x_{it}| > \alpha) \leq C_0 \exp(-C_1\alpha^{s_x}), \text{ for all } \alpha > 0, \tag{13}$$

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

$$\sup_{t}\Pr(|u_t| > \alpha) \leq C_2 \exp(-C_3\alpha^{s_u}), \text{ for all } \alpha > 0. \tag{14}$$

Assumption 4 *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 $\mathcal{F}_t^{xn} = \cup_{j=k+k^*+1}^{n}\mathcal{F}_{jt}^x$, $\mathcal{F}_t^{xs} = \cup_{i=1}^{k+k^*}\mathcal{F}_{jt}^x$, and $\mathcal{F}_t^x = \mathcal{F}_t^{xn} \cup \mathcal{F}_t^{xs}$. 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 $i = 1, 2, ..., k + k^*$, $j = k + k^* + 1, ..., n$, 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 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*

$$u_{x,t,T} =: u_{x,t} = x_t - \boldsymbol{\gamma}'_{qx,T}\boldsymbol{q}_{\cdot t}. \tag{15}$$

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

Under Assumption 1(b), the net effect coefficient, θ_i, defined in (3), can be equivalently written as

$$\theta_i = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{X}_k \boldsymbol{\beta}_k\right) = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{y}/T\right) = \sum_{j=1}^{k}\beta_j\sigma_{ji}, \tag{16}$$

where

$$\boldsymbol{y} = a\boldsymbol{\tau}_T + \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{u}, \tag{17}$$

is the DGP, (1), written in matrix form, in which as before $\boldsymbol{\tau}_T$ 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 signal variables, $\boldsymbol{\beta}_k = (\beta_1, \beta_2, ..., \beta_k)'$ 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.

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.

Remark 3 *The signal and (pure) noise variables are allowed to be correlated amongst themselves; namely, no restrictions are imposed on* $\rho_{ij} = E(x_{it}x_{jt})$ *for* $i, j = 1, 2, ..., k$, *and on* ρ_{ij} *for* $i, j = k + k^* + 1, k + k^* + 2, ..., n$. *Also signal and pseudo-signal variables are allowed to be correlated; namely,* ρ_{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 4,* $E[x_{it} - E(x_{it})|x_{jt}] = 0$ *for* $i = 1, 2, ..., k$ *and* $j = k + k^* + 1, ..., n$. *This implies that, if there are common factors, they cannot be shared between signal/pseudo-signal variables and noise variables.*

Remark 4 *The exponential bounds in Assumption 3 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).*

Remark 5 *Assumption 2 is a set of regularity conditions. Condition (a) is needed for obvious reasons. In our setting, it is assumed that $\beta_i \neq 0$, for $i = 1, 2, ..., k$, and zero otherwise. Theorem 3 can be used to extend this framework to small but nonzero β_i as discussed in Remark 13. Assumption 2.b is needed to preclude the possibility that the linear combination $\sum_{j=1}^{k} \beta_j \rho_{ji}$ is exactly zero despite a non-zero β_i. This assumption can be relaxed as discussed in Section 4.6.*

Remark 6 *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 $u_{x,t,T} = x_t - \mu_{x,T}$.*

Remark 7 *It is important to contrast our assumptions to those in the literature. In most analyses of alternative methods, such as penalised regression, it is usual to assume that either x_{it} is deterministic or, in a more general setting, iid. See, for example, Buhlmann and van de Geer (2011) or Zheng, Fan, and Lv (2014) for a more recent contribution. Our martingale difference assumption compares favourably to the iid assumption. Further, in Section 4.7 we relax this assumption in a variety of ways. See also Remark 20, on the need to assume that noise variables are martingale difference processes.*

Remark 8 *It is also important to consider how our assumptions on the correlation between signal and pseudo-signal covariates compare to those made in the literature. We allow for noise variables to have a common factor, and do not require the covariance matrix of $\boldsymbol{x}_{nt}$ to be sparse, in contrast with the existing large-dataset literature, where sparsity of the covariance matrix of the n potential regressor variables is a common assumption.*

Remark 9 *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. As our results will indicate, the OCMT approach can identify the $k + k^*$ signal and 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 $k^* = o(T^{1/4})$.[1] In contrast, a number of crucial*

[1]The rate $O\left(T^{1/4}\right)$ is more restrictive than the rate $O\left(T^{1/3}\right)$ commonly derived in the literature that deals with an increasing number of regressors, see Berk (1974), Said and Dickey (1984) or Chudik and Pesaran (2013). The difference comes from the assumption on the norm of the covariance matrix of regressors and its inverse. The cited literature considers an increasing number of lags of a stationary variable as regressors and, consequently, this norm is bounded in the number of regressors. In contrast, our analysis allows for the presence of strong cross-sectional dependence among the regressors and, therefore, the norm of their covariance matrix is no longer bounded in the number of regressors.

issues arise in the context of Lasso, or more generally when L_q, $0 \leq q \leq 1$, penalty functions are used. Firstly, it is customary to assume a restrictive framework of fixed-design regressor matrices, where in many cases a generalisation to stochastic regressors is not straightforward, such as the spark condition of Donoho and Elad (2003) and Zheng, Fan, and Lv (2014). Secondly, an essentially necessary 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, e.g., Section 7.5 of Buhlmann and van de Geer (2011). This condition is acknowledged to be rather restrictive for a large n.

Remark 10 *A final 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 rarely reported. 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, and correlations among signal and noise variables do not present a real problem.*

4 Theoretical Results

In this section, we present the main theoretical results using a number of lemmas established in the Appendix. The key result which we will be using repeatedly below is Lemma 16. This lemma provides sharp bounds for $\Pr\left(\left|t_{\hat{\phi}_i}\right| > c_p(n)|\theta_i \neq 0\right)$. It is important to appreciate the complex tasks involved in deriving such bounds. These tasks include deriving exponential inequalities for unbounded martingale difference processes (Lemma 9), handling products involving martingale difference processes (Lemma 10), and dealing with the denominator of the $t-$ratio, $t_{\hat{\phi}_i}$, which requires the exponential inequalities derived in Lemma 14. Further, since we wish to accommodate extensions of the procedure for more general forms of time dependence and allowing for the possibility of $\theta_i = 0$ even 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.

4.1 True positive rate ($TPR_{n,T}$)

We first examine the statistical properties of $TPR_{n,T}$ defined by (10), under the assumption that $\theta_i \neq 0$ if $\beta_i \neq 0$. Note that

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

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

$$E\left|TPR_{n,T}\right| = \frac{\sum_{i=1}^{k} \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) | \beta_i \neq 0\right]}{k} = \frac{\sum_{i=1}^{k} \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) | \theta_i \neq 0\right]}{k}.$$

Suppose there exists $\kappa > 0$ such that $n = O\left(T^{\kappa}\right)$. Using (A.87) of Lemma 16, where the matrix $\mathbf{Q}$, referred to in the statement of the Lemma, is set equal to $\boldsymbol{\tau}_T$, and noting that $c_p\left(n\right)$ is given by (7)-(8),

$$1 - \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) \left|\theta_i \neq 0\right.\right] = O\left[\exp\left(-C_2 T^{C_3}\right)\right],$$

for some $C_2, C_3 > 0$, where as defined by (16), $\theta_i = E\left(\boldsymbol{x}_i' \boldsymbol{M}_{\tau} \boldsymbol{y}/T\right)$. 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

$$\Pr\left[\left|t_{\hat{\phi}_i}\right| \leq c_p(n) \left|\theta_i \neq 0\right.\right] = O\left[\exp\left(-C_2 T^{C_3}\right)\right], \tag{18}$$

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

$$k^{-1} \sum_{i=1}^{k} \Pr\left(\left|t_{\hat{\phi}_i}\right| \leq c_p(n) | \beta_i \neq 0\right) = k^{-1} \sum_{i=1}^{k} O\left[\exp\left(-C_2 T^{C_3}\right)\right]. \tag{19}$$

The above arguments lead to the following Theorem:

Theorem 1 *Consider the DGP defined by (1), and suppose that Assumptions 1-4 hold, Assumption 5 holds for the pairs (x_{it}, x_{jt}), $i = 1, 2, ..., k$, $j = k+1, k+2, ..., k+k^*$, $c_p(n)$ is given by (7)-(8) for any positive finite δ and $0 < p < 1$, and $n, T \to \infty$ such that $n = O\left(T^{\kappa}\right)$ for some $\kappa > 0$. Then*

$$E\left|TPR_{n,T}\right| = 1 - O\left[\exp\left(-C_2 T^{C_3}\right)\right], \tag{20}$$

for some $C_2, C_3 > 0$, where $TPR_{n,T}$ is the true positive rate defined by (10) with the OCMT estimator of $I\left(\beta_i \neq 0\right)$ defined by (6).

Proof. (20) directly follows from (19). ■

4.2 False positive rate ($FPR_{n,T}$)

Consider now $FPR_{n,T}$ defined by (11). Again, note that the elements of FPR_n are either 0 or 1 and hence $|FPR_{n,T}| = FPR_{n,T}$. Taking expectations of (11) we have

$$E\left|FPR_{n,T}\right| = \frac{\sum_{i=k+1}^{n} \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) | \beta_i = 0\right]}{n-k}$$
$$= \frac{\sum_{i=k+1}^{k+k^*} \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) | \theta_i \neq 0\right] + \sum_{i=k+k^*+1}^{n} \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) | \theta_i = 0\right]}{n-k}, \quad (21)$$

where, as before, $\theta_i = E\left(\boldsymbol{x}_i' \boldsymbol{M}_\tau \boldsymbol{y}/T\right)$ (see (16)). Recall that under Assumptions 2 and 4, $\theta_i \neq 0$ for $i = 1, 2, ..., k + k^*$ and $\theta_i = 0$ for $i = k + k^* + 1, k + k^* + 2, ..., n$. Using (A.87) of Lemma 16 and assuming there exists $\kappa > 0$ such that $n = O\left(T^\kappa\right)$, we have

$$k^* - \sum_{i=k+1}^{k+k^*} \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) | \theta_i \neq 0\right] = O\left[\exp\left(-C_2 T^{C_3}\right)\right],$$

for some finite positive constants C_2 and C_3. Moreover, (A.86) of Lemma 16 implies that for any $0 < \varkappa < 1$ there exist finite positive constants C_0 and C_1 such that

$$\sum_{i=k+k^*+1}^{n} \Pr\left[\left|t_{\hat{\phi}_i}\right| > c_p(n) | \theta_i = 0\right] = \sum_{i=k+k^*+1}^{n} \left\{\exp\left[\frac{-\varkappa c_p^2(n)}{2}\right] + \exp\left(-C_0 T^{C_1}\right)\right\}. \quad (22)$$

Using these results in (21), overall we obtain

$$E\left|FPR_{n,T}\right| = \left(\frac{k^*}{n-k}\right) + \exp\left[-\frac{\varkappa c_p^2(n)}{2}\right] + O\left[\exp(-C_0 T^{C_1})\right] + O\left[(n-k)^{-1} \exp\left(-C_2 T^{C_3}\right)\right], \quad (23)$$

which establishes the following Theorem:

Theorem 2 *Consider the DGP defined by (1), suppose that Assumptions 1, 3 and 4 hold, Assumption 5 holds for the pairs (x_{it}, x_{jt}), $i = 1, 2, ..., k$, $j = k+1, k+2, ..., k+k^*$, $c_p(n)$ is given by (7)-(8) for any positive finite δ and $0 < p < 1$, and there exists $\kappa > 0$ such that $n = O\left(T^\kappa\right)$. Then*

$$E\left|FPR_{n,T}\right| = \left(\frac{k^*}{n-k}\right) + \exp\left[-\frac{\varkappa c_p^2(n)}{2}\right] + O\left[\exp(-C_0 T^{C_1})\right], \quad (24)$$

for some $0 < \varkappa < 1$ and finite positive constants C_0 and C_1, where the false positive rate $FPR_{n,T}$ is defined in (11) with the OCMT estimator of $I\left(\beta_i \neq 0\right)$ defined by (6). Furthermore, assuming in addition that $k^* = o(n)$,

$$FPR_{n,T} \to_p 0, \quad (25)$$

as $n, T \to \infty$ such that $n = O\left(T^\kappa\right)$ for some $\kappa > 0$.

Proof. (24) directly follows from (23). For $k^* = o(n)$ and $n, T \to \infty$ such that $n = O(T^\kappa)$ for some $\kappa > 0$, (24) implies $E|FPR_{n,T}| \to 0$, which is sufficient for (25). ■

Remark 11 *It is clear that the method of analysis that gives rise to (24) can be used for related calculations. A prominent example is the false discovery rate (FDR) defined by*

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

It is easily seen that

$$FDR_{n,T} = \frac{(n-k)\, FPR_{n,T}}{\hat{R}},$$

where $\hat{R} = \sum_{i=1}^{n} I\left[I\left(\widehat{\beta_i \neq 0}\right) = 1\right]$. *Then, it follows that* $p\lim_{n,T\to\infty} \hat{R} = k + k^*$ *and, by (24),*

$$p \lim_{n,T\to\infty} FDR_{n,T} = \frac{k^*}{k + k^*}.$$

If $k^* = 0$, *then* $p\lim_{n,T\to\infty} FDR_{n,T} = 0$. *But if* $k^* > 0$ *then, it is worth noting that, as discussed in Remark 16, the norm of the estimated coefficient,* $\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n$, *will not be adversely affected since in the final multiple regression all estimated coefficients associated with pseudo-signal variables will tend to zero.*

Theorem 2 relates to the first maintained assumption about the pseudo-signal variables where only k^* of them have non-zero θ_i. This result can be extended to the case where potentially all pseudo-signal variables have non-zero θ_i, as long as θ_i are absolutely summable. Two leading cases considered in the literature are to assume that there exists a (possibly unknown) ordering such that

$$\theta_i = K_i \varrho^i, \text{ for } i = 1, 2, ..., n, , \text{ and } |\varrho| < 1, \tag{26}$$

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

$$\theta_i = K_i i^{-\gamma}, \text{ for } i = 1, 2, ..., n, \text{ and for some } \gamma > 0. \tag{27}$$

Then, we have the following extension of Theorem 2.

Theorem 3 *Consider the DGP defined by (1), suppose that Assumptions 1, 3 and 4 hold, Assumption 5 holds for the pairs* (x_{it}, x_{jt}), $i = 1, 2, ..., k$, $j = k+1, k+2, ..., n$, *and instead of Assumption 2(b), condition (26) holds. Moreover, let* $c_p(n)$ *be given by (7)-(8) for any positive finite* δ *and* $0 < p < 1$, *and suppose there exists* $\kappa > 0$ *such that* $n = O(T^\kappa)$. *Then for all* $\zeta > 0$ *we have*

$$E|FPR_{n,T}| = o(n^{\zeta-1}) + O\left[\exp(-C_0 T^{C_1})\right],$$

for some finite positive constants C_0 and C_1, where $FPR_{n,T}$ is defined by (11) with the OCMT estimator of $I(\beta_i \neq 0)$ defined by (6). If condition (27) holds instead of condition (26), then, assuming $\gamma > 1/2\kappa^2$ and $n, T \to \infty$, such that $n = T^\kappa$, for some $\kappa > 0$, we have

$$FPR_{n,T} \to_p 0.$$

Proof. A proof is provided in Section A of the online theory supplement. ■

Remark 12 *$FPR_{n,T}$ can be somewhat controlled by the choice of p. But, by result (ii) of Lemma 1, it follows that $\exp\left[-\varkappa c_p^2(n)/2\right] = O\left(n^{-\delta\varkappa}\right)$, and hence $E\left|FPR_{n,T}\right|$ converges to zero at the rate of $n^{-\min\{1,\delta\varkappa\}}$ so long as the number of pseudo-signal variables is bounded. The main result of Theorem 2 also holds if the number of pseudo-signal variables, k^*, rise with n so long as $k^*/n \to 0$, as $n \to \infty$.*

Remark 13 *Theorem 3 assumes that $\theta_i \neq 0$, for $i = k+1, 2, ..., n$ while $\beta_i = 0$, for $i = k+1, 2, ..., n$. Of course, exactly the same analysis as in the proof of Theorem 3 can be used when $\beta_i \neq 0$, for $i = k+1, 2, ..., n$, to allow an analysis of the ability of OCMT to pick up weak signal variables, since in the proof of the Theorem we explore the probability that $\left|t_{\hat{\phi}_i}\right| > c_p(n)$ when θ_i is small. It is clear that the relationship between $\sqrt{T}\theta_i$ and $c_p(n)$ is crucial . Given (i) of Lemma 1, a variable will be selected if $\ln(n)^{1/2}/\left(\sqrt{T}\theta_i\right) = o(1)$ and so our analysis can easily handle relatively weak signals as long as $\beta_i = \ominus(\theta_i)$.*

4.3 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. Mathematically, the event of choosing the pseudo-true model is given by

$$\mathcal{A}_0 = \left\{\sum_{i=1}^{k} I\widehat{(\beta_i \neq 0)} = k\right\} \cap \left\{\sum_{i=k+k^*+1}^{n} I\widehat{(\beta_i \neq 0)} = 0\right\}. \tag{28}$$

The above definition implies that the probability of not choosing the pseudo-true model is bounded by the following expression

$$\Pr\left(\sum_{i=1}^{k} I\widehat{(\beta_i \neq 0)} < k\right) + \Pr\left(\sum_{i=k+k^*+1}^{n} I\widehat{(\beta_i \neq 0)} > 0\right) = A + B.$$

However

$$A = \Pr\left(\sum_{i=1}^{k} I\widehat{(\beta_i \neq 0)} < k\right) \leq \sum_{i=1}^{k} \Pr\left(I\widehat{(\beta_i \neq 0)} = 0\right)$$
$$= \sum_{i=1}^{k} \Pr\left[\left|t_{\hat{\phi}_i}\right| \leq c_p(n) \,|\theta_i \neq 0\right] \leq k \sup_i \Pr\left[\left|t_{\hat{\phi}_i}\right| \leq c_p(n) \,|\theta_i \neq 0\right],$$

and using (18), assuming that $n = O\left(T^{\kappa}\right)$ for some $\kappa > 0$, we obtain (see also (A.87) of Lemma 16)

$$A \leq \exp\left(-C_2 T^{C_3}\right).$$

for some finite positive constants C_2 and C_3. Similarly, using (22) and result (ii) of Lemma 1, for some $C_0 > 0$,

$$B \leq \sum_{i=k+k^*+1}^{n} \Pr\left(I\widehat{(\beta_i \neq 0)} = 1\right) \leq \frac{C_0 pn}{f(n)}. \tag{29}$$

So, the probability of choosing the pseudo-true model is bounded from below, namely

$$\Pr(\mathcal{A}_0) \geq 1 - C_0 \frac{n}{f(n)} - \exp\left(-C_2 T^{C_3}\right). \tag{30}$$

If, in addition, $\delta > 1$, then $n/f(n) = n^{1-\delta} \to 0$, and

$$\Pr(\mathcal{A}_0) \to 1,$$

as n,$T \to \infty$ such that $n = O\left(T^{\kappa}\right)$ for some $\kappa > 0$. A further result may be obtained by considering $\Pr\left(\hat{k} - k - k^* > j\right)$ where

$$\hat{k} = \textstyle\sum_{i=1}^{n} I\widehat{(\beta_i \neq 0)}. \tag{31}$$

A bound on this probability is obtained in Lemma 17. The results of that Lemma and the above result on the probability of selecting the pseudo-true model are summarised in the Theorem 4 below.

Theorem 4 *Consider the DGP defined by (1), suppose that Assumptions 1-4 hold, Assumption 5 holds for the pairs* (x_{it}, x_{jt}), $i = 1, 2, ..., k$, $j = k+1, k+2, ..., k+k^*$, $c_p(n)$ *is given by (7)-(8) for any positive finite* δ *and* $0 < p < 1$, θ_i, *defined by (16), is zero for* $i = k+k^*+1, k+k^*+2, ..., n$, *there exists* $\kappa > 0$ *such that* $n = O\left(T^{\kappa}\right)$, *and* $k^* = o(n)$. *Then there exist finite positive constants* C_0, C_1 *and* C_2, *such that*

$$\Pr(\mathcal{A}_0) \geq 1 - C_0 \frac{n}{f(n)} - \exp\left(-C_1 T^{C_2}\right), \tag{32}$$

where $\mathcal{A}_0$ is the pseudo-true model defined by (28) with the OCMT estimator of $I(\beta_i \neq 0)$ defined by (6). If, in addition, $f(n) = n^{\delta}$ and $\delta > 1$, then

$$\Pr(\mathcal{A}_0) \to 1, \tag{33}$$

as $n, T \to \infty$ such that $n = O(T^{\kappa})$ for some $\kappa > 0$. Further, there exist $0 < \varkappa < 1$ and finite positive constants C_0, and C_1, such that,

$$\Pr\left(\hat{k} - k - k^* > j\right) = \frac{(n-k-k^*)}{j}\left\{\exp\left[-\frac{\varkappa c_p^2(n)}{2}\right] + O\left[\exp(-C_0 T^{C_1})\right]\right\}, \tag{34}$$

for $j = 1, 2, ..., n-k-k^*$, where $\hat{k}$ is defined by (31) with the OCMT estimator of $I(\beta_i \neq 0)$ defined by (6).

Proof. Lower bound for $\Pr(\mathcal{A}_0)$ is derived in (30), from which (33) easily follows. Result (34) directly follows from Lemma 17, noting that the term $(k+k^*)j^{-1}O\left[\exp\left(-C_2 T^{C_3}\right)\right]$ on the right side of (A.92) is dominated by the remaining terms when $k^* = o(n)$. ■

Remark 14 *The power of the OCMT procedure in selecting the signal x_{it} rises with $\sqrt{T}|\theta_i|/\sigma_{e_i,(T)}\sigma_{x_i,(T)}$, so long as $c_p(n)/\sqrt{T} \to 0$, as n and $T \to \infty$ (see A.87), where $\sigma_{e_i,(T)}$ and $\sigma_{x_i,(T)}$ are defined by (A.84), replacing $\mathbf{e}$, $\mathbf{x}$, and $\mathbf{Q}$ by $\mathbf{e}_i$, $\mathbf{x}_i$, and $\boldsymbol{\tau}_T$, 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.)*

4.4 The norm of the estimated coefficients

In this section, we consider the coefficient norm $E\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\|$, where $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_n)'$, is the vector of the OCMT estimators, $\tilde{\beta}_i$, for $i = 1, 2, ..., n$, defined by (9), and $\boldsymbol{\beta}_n$ is the associated true values. We need to determine whether, and if so at what rates, $E\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\|$ tends to zero. We assume that we only consider models with a maximum of $l_{\max}$ regressors, namely $\hat{k} = \dim\left(\tilde{\boldsymbol{\beta}}_n\right) \leq l_{\max}$. The choice of $l_{\max}$ will follow from our subsequent analysis. To derive this we consider the set of mutually exclusive events given by

$$\mathcal{A}_{i,j} = \left\{\left[\sum_{s=0}^{k+k^*} \widehat{I(\beta_s \neq 0)} = i\right] \cap \left[\sum_{s=k+k^*+1}^{n} \widehat{I(\beta_s \neq 0)} = j\right]\right\}, \; i = 0, ..., k+k^*, \; j = 0, ..., n-k-k^*.$$

Using this decomposition we can proceed to prove the following Theorem.

Theorem 5 *Consider the DGP defined by (1), suppose that Assumptions 1-4, conditions (i)-(ii) of Lemma 19 hold, Assumption 5 holds for the pairs (x_{it}, x_{jt}), $i = 1, 2, ..., k$, $j =$*

$k+1, k+2, ..., k+k^*$, $c_p(n)$ *is given by (7)-(8) for any positive finite* δ *and* $0 < p < 1$, $\theta_i = 0$, *for* $i = k+k^*+1, ..., n$, *and there exists* $\kappa > 0$ *such that* $n = O(T^\kappa)$. *It then follows that*

$$E\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| = O\left[\left(\frac{l_{\max}^4}{T} + l_{\max}\right)\exp\left(-C_1 T^{C_2}\right)\right] + O\left[\left(\frac{l_{\max}^4}{T}\right)\frac{pn}{f(n)}\right], \tag{35}$$

for some finite positive constants C_1 *and* C_2, *where* $l_{\max}$ *defines the maximum number of the selected regressors, the vector of OCMT estimators* $\tilde{\boldsymbol{\beta}}_n$ *is defined in (9),* $\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)'$ *and* $f(n)$ *is given by (8).*

Proof. The proof is provided in Appendix A.1.1. ■

Remark 15 *As can be seen from the statement of the above theorem, result (35) requires stronger conditions than those needed for the proof of the earlier results on the limiting properties of* $TPR_{n,T}$ *and* $FPR_{n,T}$. *In particular, the additional technical conditions (i) and (ii) of Lemma 19 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* $\left(1 - \|\boldsymbol{\Sigma}_{ss}^{-1}\|_F \left\|\hat{\boldsymbol{\Sigma}}_{ss} - \boldsymbol{\Sigma}_{ss}\right\|_F\right)^{-4}$, *which is needed for our derivations. Under our conditions on the number of selected regressors,* $\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F E\left(\left\|\hat{\boldsymbol{\Sigma}}_{ss} - \boldsymbol{\Sigma}_{ss}\right\|_F\right) = o(1)$, *but this is not sufficient for* $E\left[\left(1 - \|\boldsymbol{\Sigma}_{ss}^{-1}\|_F \left\|\hat{\boldsymbol{\Sigma}}_{ss} - \boldsymbol{\Sigma}_{ss}\right\|_F\right)^{-4}\right] = O(1)$, *so an extra technical assumption is needed.*

Remark 16 *It is important to provide intuition on why we can get a consistency result for the Frobenius norm of the estimated regressors even though the selection includes pseudo-signal variables. There are two reasons for this. First, since OCMT procedure selects all the signals with probability one as* n *and* $T \to \infty$, *then the coefficients of the additionally selected regressors (whether pseudo-signal or noise) will tend to zero with* T. *Second, our restriction, that there exist only a finite number of pseudo-signal (or, in an extended analysis, an infinite number of them that grows at a much lower rate than* T*), implies that their inclusion can be accommodated since their estimated coefficients will tend to zero and the variance of these estimated coefficients will be well controlled. Of course, some noise variables will also be selected in small samples, but we restrict their number by using a bound on the number of selected regressors (namely* $l_T \leq l_{\max}$ *in our proofs). In practice, our Monte Carlo evidence suggests that the number of noise variables selected is very well controlled by our multiple testing framework and there is no practical need for enforcing the bound in small samples, in line with (34).*

4.5 The norm of the in-sample errors

Consider the following norm of the in-sample errors

$$F_{\tilde{u}} = E\left(\frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2\right) = \frac{1}{T}E\left(\tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}}\right) = \frac{1}{T}E\left\|\tilde{\boldsymbol{u}}\right\|^2,$$

where $\tilde{\boldsymbol{u}} = (\tilde{u}_1, \tilde{u}_2, ..., \tilde{u}_T)'$, $\tilde{u}_t$ is defined by (12), and $\left\|\tilde{\boldsymbol{u}}\right\|^2 = \tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}}$.

Theorem 6 *Consider the DGP defined by (1), suppose that Assumptions 1-4 hold, Assumption 5 holds for the pairs (x_{it}, x_{jt}), $i = 1, 2, ..., k$, $j = k+1, k+2, ..., k+k^*$, $c_p(n)$ is given by (7)-(8) for any positive finite δ and $0 < p < 1$, and $\theta_i = 0$, for $i = k+k^*+1, k+k^*+2, ..., n$. Then*

$$E\left(\frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2\right) - \sigma^2 \to 0, \tag{36}$$

as $n, T \to \infty$ such that $n = O\left(T^{\kappa}\right)$ for some $\kappa > 0$. Also, if $n/f\left(n\right) = o\left(1/T\right)$, then

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

Proof. The proof is provided in Appendix A.1.2. ■

Remark 17 *This theorem establishes the oracle property of the OCMT procedure for the in-sample fit of the selected regression equation, and does not require the additional technical conditions required for the proof of the Frobenius norm of the estimated coefficients. This is because fitted values are defined even if the sample covariance of the selected regressors is not invertible.*

4.6 Relaxing the assumption of nonzero signal net effects: an iterated multiple testing procedure

Assumption 2(b) states that regressors for which $\beta_i \neq 0$, also satisfy $\theta_i \neq 0$. Clearly, there are circumstances when this condition does not hold. To deal with such a possibility we propose the following iterated version of the multiple testing procedure. Initially, as before, we consider the n bivariate regressions of y_t on a constant and x_{it} for $i = 1, 2, ..., n$ (see (4)),

$$y_t = c_{i,(1)} + \hat{\phi}_{i,(1)}x_{it} + e_{it,(1)},$$

and compute the t-ratios

$$t_{\hat{\phi}_{i,(1)}} = \frac{\hat{\phi}_{i,(1)}}{s.e.\left(\hat{\phi}_{i,(1)}\right)} = \frac{T^{-1/2}\boldsymbol{x}_i'\boldsymbol{M}_{(0)}\boldsymbol{y}}{\hat{\sigma}_{i,(1)}\sqrt{\boldsymbol{x}_i'\boldsymbol{M}_{(0)}\boldsymbol{x}_i}}, \tag{38}$$

where $\hat{\phi}_{i,(1)} = \left(\boldsymbol{x}_i'\boldsymbol{M}_{(0)}\boldsymbol{x}_i\right)^{-1}\boldsymbol{x}_i'\boldsymbol{M}_{(0)}\boldsymbol{y}$, $\hat{\sigma}^2_{i,(1)} = \boldsymbol{e}'_{i,(1)}\boldsymbol{e}_{i,(1)}/T$, $\boldsymbol{e}_{i,(1)} = \boldsymbol{M}_{i,(0)}\boldsymbol{y}$, $\boldsymbol{M}_{i,(0)} = \boldsymbol{I}_T - \boldsymbol{X}_{i,(0)}(\boldsymbol{X}'_{i,(0)}\boldsymbol{X}_{i,(0)})^{-1}\boldsymbol{X}'_{i,(0)}$, $\boldsymbol{X}_{i,(0)} = (\boldsymbol{x}_i, \boldsymbol{\tau}_T)$, and $\boldsymbol{M}_{(0)} = \boldsymbol{I}_T - \boldsymbol{\tau}_T\boldsymbol{\tau}'_T/T$. The first stage multiple testing estimator of $I(\beta_i \neq 0)$ is, similarly to (6), given by

$$\widehat{I(\beta_i \neq 0)} = I\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n)\right],\ i = 1, 2, ..., n,$$

where $c_p(n)$ is given by (7). Regressors for which $\widehat{I(\beta_i \neq 0)} = 1$ are selected as signals in the first stage. Denote the number of variables selected in the first stage by $\hat{k}^s_{(1)}$, the index set of the selected variables by $\mathcal{S}^s_{(1)}$, and the $T \times \hat{k}^s_{(1)}$ matrix of the $\hat{k}^s_{(1)}$ selected variables by $\boldsymbol{X}^s_{(1)}$. Finally, let $\boldsymbol{X}_{(1)} = (\boldsymbol{\tau}_T, \boldsymbol{X}^s_{(1)})$, $\hat{k}_{(1)} = \hat{k}^s_{(1)}$, $\mathcal{S}_{(1)} = \mathcal{S}^s_{(1)}$ and $\mathcal{N}_{(1)} = \{1, 2, ..., n\} \setminus \mathcal{S}_{(1)}$.

In stages $j = 2, 3, ...$, we consider the $n - \hat{k}_{(j-1)}$ 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

$$t_{\hat{\phi}_{i,(j)}} = \frac{\hat{\phi}_{i,(j)}}{s.e.\left(\hat{\phi}_{i,(j)}\right)} = \frac{\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\mathbf{y}}{\hat{\sigma}_{i,(j)}\sqrt{\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{x}_i}},\ \text{for } i \in \mathcal{N}_{(j-1)},\ j = 2, 3, ..., \tag{39}$$

where $\hat{\phi}_{i,(j)} = \left(\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{x}_i\right)^{-1}\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{y}$ denotes the estimated conditional net effect of x_{it} on y_t in stage j, $\hat{\sigma}^2_{i,(j)} = T^{-1}\boldsymbol{e}'_{i,(j)}\boldsymbol{e}_{i,(j)}$, $\boldsymbol{M}_{(j-1)} = \boldsymbol{I}_T - \boldsymbol{X}_{(j-1)}(\boldsymbol{X}'_{(j-1)}\boldsymbol{X}_{(j-1)})^{-1}\boldsymbol{X}'_{(j-1)}$, $\mathbf{e}_{i,(j)} = \boldsymbol{M}_{i,(j-1)}\boldsymbol{y}$ denotes the residual of the regression, $\boldsymbol{M}_{i,(j-1)} = \boldsymbol{I}_T - \boldsymbol{X}_{i,(j-1)}(\boldsymbol{X}'_{i,(j-1)}\boldsymbol{X}_{i,(j-1)})^{-1}\boldsymbol{X}'_{i,(j-1)}$, and $\boldsymbol{X}_{i,(j-1)} = \left(\boldsymbol{x}_i, \boldsymbol{X}_{(j-1)}\right)$. Regressors for which

$$\widehat{I(\beta_i \neq 0)} = I\left[\left|t_{\hat{\phi}_{i,(j)}}\right| > c_p(n)\right] = 1$$

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 $\hat{k}^s_{(j)}$, their index set by $\mathcal{S}^s_{(j)}$, and the $T \times \hat{k}^s_{(j)}$ matrix of the $\hat{k}^s_{(j)}$ selected variables by $\mathbf{X}^s_{(j)}$. Also define $\mathbf{X}_{(j)} = (\mathbf{X}_{(j-1)}, \mathbf{X}^s_{(j)})$, $\hat{k}_{(j)} = \hat{k}^s_{(j)} + \hat{k}_{(j-1)}$, $\mathcal{S}_{(j)} = \mathcal{S}^s_{(j)} \cup \mathcal{S}_{(j-1)}$, and $\mathcal{N}_{(j)} = \{1, 2, ..., n\} \setminus \mathcal{S}_{(j)}$, and then proceed to stage $j + 1$. The procedure stops when no regressors are selected at a given stage, which we denote by stage J.

In this multiple procedure, $\widehat{I(\beta_i \neq 0)} = 1$ as long as $I\left[\left|t_{\hat{\phi}_{i,(j)}}\right| > c_p(n)\right] = 1$ for some $j = 1, 2, ..., J$. We show in Lemma 20 in the Appendix that, when T is sufficiently large, then at least one signal must be selected in each stage of the iterated multiple testing procedure with high probability. Thus, when signal variables are uncorrelated with noise variables, it must be that $J \leq k$. In practice, J is likely to be small, since the specificity of zero signal net effects is less plausible, and all signals with nonzero θ will be picked up (with high probability) in the first stage.

In a final step, the regression model is estimated by running the OLS regression of y_t on all selected variables, namely the regressors x_{it} for which $\widehat{I(\beta_i \neq 0)} = 1$, over all $i = 1, 2, ..., n$. We will continue to use OCMT to refer to this iterated version, which we will implement in the Monte Carlo section below, since the possibility of signal variables with zero net effect cannot

be ruled out in practice. Setting $J = 1$ tends to improve the small sample performance of the OCMT approach marginally when all signal variables have nonzero net effects, namely $\theta_i \neq 0$ for $i = 1, 2, ..., k$. In other words, our small sample evidence in the next section shows that allowing $J > 1$, using the stopping rule defined above, does not significantly deteriorate the small sample performance when $\theta_i \neq 0$ for $i = 1, 2, ..., k$, while it picks-up the signal variables with zero net effects with high probability.[2]

From a theoretical perspective, we note that our Lemmas, and in particular Lemma 16, can provide an exponential inequality for t-statistics of the form (39) as long as the number of regressors contained in $\boldsymbol{X}_{(j-1)}$ is of lower order than $T^{1/3}$. This is a weaker restriction than the restriction on $l_{\max}$ needed for our Frobenius norm result in Theorem 5, which requires that $l_{\max} = o\left(T^{1/4}\right)$. Therefore, it immediately follows that, under the restriction required for the Frobenius norm, the results obtained in Theorems 1-2 hold for the iterated version of the OCMT approach.

It is worth briefly comparing OCMT to a standard version of boosting (B). OCMT selects more than one regressor at each iteration depending on the particular outcome of OCMT in that iteration, while B only selects one regressor at each iteration. OCMT has a clear stopping rule in that at some iteration the OCMT procedure will select no regressors while B requires the specification of a stopping rule. This is the result of the fact that OCMT has a testing component while B simply ranks regressors at each iteration based on some fitting criterion such as R^2. This difference turns out to be particularly important especially given that no fully satisfactory stopping rule seems to be available in the boosting literature.

4.7 Allowing for serial correlation in the covariates

Another 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. Although, this assumption is considerably weaker than those made in the high-dimensional model estimation and selection literature, where it is usually assumed that regressors are either non-stochastic or independently distributed, it is nevertheless restrictive for many economic applications. If we are willing to assume that either u_t is normally distributed or the covariates are deterministic, then a number of powerful results become available. The relevant lemmas for the deterministic case are presented in Section D of

[2]Monte Carlo findings for the OCMT procedure with J set equal to 1 are available upon request.

the online theory Supplement. Alternatively signal/pseudo-signal regressors can be assumed to be exponentially mixing. In this general case, some weak results can still be obtained. These are described in Section B of the online theory Supplement.

5 A Monte Carlo Study

In this section we compare the small sample properties of our proposed estimator to three versions of the penalised regressions and a boosting procedure, across five different sets of Monte Carlo (MC) designs. The designs differ both in terms of the correlation patterns of the covariates and the way net effects coefficients, θ_i, and the partial effects, β_i, are related to one another. (See (3) and (16)). We also investigate the robustness of the OCMT method by considering non-Gaussian errors and serially correlated non-Gaussian covariates, and provide comparisons with the baseline results obtained using Gaussian observations. The designs are described next (Section 5.1), followed by a description of individual variable selection methods (Section 5.2), summary statistics for MC results (Section 5.3), and the MC findings (Section 5.4).

5.1 Data-generating process (DGP)

In line with our theoretical set up, we distinguish between the net effects of the signal variables, namely θ_i for $i = 1, 2, ..., k$ (which we refer to as signal θ), from those of noise variables, namely noise θ's, defined as θ_i for $i = k+1, k+2, ..., n$. Initially, we consider four sets of designs depending on the choices of θ_i associated with signal and noise variables:

	Noise θ's	
Signal θ's	All noise θ's are zero	At least one noise θ is nonzero
All signal θ's are nonzero	Design set I	Design set II
Some signal θ's are zero	Design set III	Design set IV

In the first set of experiments (set I), $\beta_i \neq 0$ if and only if $\theta_i \neq 0$ and the pseudo-true model and the true model coincide. In the second set of experiments (set II), we allow for some noise variables to have nonzero $\theta's$ (i.e. we allow for inclusion of pseudo-signal variables amongst the covariates). In this case, pseudo-signals will be picked up by the OCMT procedure due to their non-zero correlation with the signal variables. In the third set of experiments (set III), we allow for signal variables with zero net effects, namely variables where $\beta_i \neq 0$ but $\theta_i = 0$. In the fourth set of experiments (set IV), we include signal variables with zero net effect as well as pseudo-signals. Design sets I-IV assume the DGP is exactly sparse with a fixed number of signal variables. To investigate the property of the OCMT procedure when

the DGP is approximately sparse, we also consider experiments where k changes with n (set V).

In the case of all five designs, we consider several options in generating the covariates. We allow the covariates to be serially correlated and consider different degrees of correlations across them. As noted earlier, we also consider experiments with Gaussian and non-Gaussian draws.

5.1.1 Designs with zero correlations between signal and noise variables (design set I)

In the first set of experiments, there are no pseudo-signal variables and all signal variables have $\theta_i \neq 0$. y_t is generated as:

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

$$\text{Gaussian: } u_t \sim IIDN(0,1),$$
$$\text{non-Gaussian: } u_t = \left[\chi^2_t(2) - 2\right]/2,$$

where $\chi^2_t(2)$ 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 the vector of variables $\boldsymbol{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$:

DGP-I(a) Temporally uncorrelated and weakly collinear regressors:

$$\text{signal variables: } x_{it} = (\varepsilon_{it} + \nu g_t)/\sqrt{1+\nu^2}, \text{ for } i = 1, 2, 3, 4, \tag{41}$$
$$\text{noise variables: } x_{5t} = \varepsilon_{5t},\ x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2}, \text{ for } i > 5, \tag{42}$$

where g_t and ε_{it} are independent draws either from $N(0,1)$ or from $[\chi^2_t(2) - 2]/2$, 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 according to (41)-(42) with $\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + \sqrt{1-\rho_i^2} e_{it}$, $e_{it} \sim IIDN(0,1)$ or $IID\left[\chi^2_t(2) - 2\right]/2$, and (as before) $g_t \sim IIDN(0,1)$ or $IID\left[\chi^2_t(2) - 2\right]/2$, and $\nu = 1$. We set $\rho_i = 0.5$ for all i.

DGP-I(c) Strongly collinear noise variables due to a persistent unobserved common factor:

$$\text{signal variables: } x_{it} = (\varepsilon_{it} + g_t)/\sqrt{2}, \text{ for } i = 1, 2, 3, 4,$$
$$\text{noise variables: } x_{5t} = (\varepsilon_{5t} + b_i f_t)/\sqrt{3},\ x_{it} = \left[(\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2} + b_i f_t\right]/\sqrt{3}, \text{ for } i > 5,$$

$b_i \sim IIDN(1,1)$, and $f_t = 0.95 f_{t-1} + \sqrt{1-0.95^2} v_t$, where v_t, g_t, and ε_{it} are independent draws from $N(0,1)$ or $\left[\chi^2_t(2) - 2\right]/2$.

DGP-I(d) Low or high pair-wise correlation of signal variables: Regressors are generated according to (41)-(42) where g_t and ε_{it} are independent draws from $N(0,1)$ or $\left[\chi^2_t(2) - 2\right]/2$ (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.

DGP-I(a) is our baseline experiment, which does not feature any pseudo-signals, and the pure noise variables are only weakly collinear. DGP-I(b) departs from the baseline by introducing temporal correlation among variables. As a result, we expect the performance of all methods to deteriorate in DGP-I(b), since a larger T is required to detect spurious collinearity between the signal and noise variables. DGP-I(c) is used to demonstrate that strong collinearity (and high temporal correlation) of pure noise variables does not affect the baseline performance much. In contrast with DGP-I(b), spurious collinearity between the signal and noise variables is not a problem when signal variables are not temporally correlated (this problem occurs only when *both* signal and noise variables are temporally correlated). DGP-I(d) considers low (20%) and high (80%) pair-wise correlation of signal variables to demonstrate the main trade-offs between the OCMT method and penalised regressions. We expect that an increase in collinearity of signal variables improves the performance of the OCMT method. In contrast, we expect the penalised regressions to suffer from an increase in collinearity of signal variables simply because the marginal contribution of signal variables to overall fit diminishes with higher collinearity of signals.

5.1.2 Designs with non-zero correlations between signal and noise variables (design set II)

In the second set of experiments, 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:

$$\begin{aligned}
&\text{signal variables: } x_{it} = (\varepsilon_{it} + g_t)/\sqrt{2}, \text{ for } i = 1,2,3,4, \\
&\text{noise variables: (pseudo-signal) } x_{5t} = \varepsilon_{5t} + \kappa x_{1t},\ x_{6t} = \varepsilon_{6t} + \kappa x_{2t}, \text{ and} \\
&\qquad\qquad \text{(pure noise) } x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2}, \text{ for } i > 6,
\end{aligned}$$

where, as before, g_t, and ε_{it} are independent draws from $N(0,1)$ or $\left[\chi^2_t(2) - 2\right]/2$. 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: $\boldsymbol{x}_{nt} \sim IID\left(\mathbf{0}, \boldsymbol{\Sigma}_x\right)$ 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 $\left[\chi_t^2(2) - 2\right]/2$.

When pseudo-signal variables are present ($k^* > 0$), the OCMT procedure is expected to pick up the pseudo-signals in DGP-II(a) with high probability, but $\tilde{\boldsymbol{\beta}}_n$ remains consistent in the sense that $\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \to_p 0$, see Theorem 5. However, $\tilde{\boldsymbol{\beta}}_n$ will be asymptotically less efficient than the estimates of the true model due to presence of pseudo-signals. DGP-II(b) corresponds to the interesting case, where $\theta_i \neq 0$ for all $i = 1, 2, ..., n$.

5.1.3 Designs with zero net signal effects (design set III)

In the third set of experiments, 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 $\theta_4 = 0$ (the net effect of the fourth signal variable):

DGP-III One of the 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 $i \geq 4$.

We note that it cannot be the case that all four signal variables have zero net effects. The presence of zero net signal effects in DGP-III violates Assumption 2(b), and we use DGP-III to illustrate the effectiveness of OCMT procedure, where the fourth variable will be picked up with high probability in the second stage.

5.1.4 Designs with zero net signal effects and pseudo-signal variables (design set IV)

In the fourth set of experiments, we allow for signal variables with zero θ as well as the pseudo-signals 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 Designs with $k = n$ signal variables (design set V)

In the fifth set of experiments, we consider $k = n$ signal variables. This design is inspired by the literature on approximately sparse models (Belloni, Chernozhukov, and Hansen (2014b)).

DGP-V $\beta_i = 1/i^2$, $\boldsymbol{x}_{nt}$ are generated as in design DGP-II(b).

All autoregressive processes are generated with zero starting values and 100 burn-in periods. In all DGPs, we set ς in (40) so that $R^2 = 30\%$, 50% or 70%. We consider $n = 100, 200$ and 300, $T = 100, 300$ and 500, and carry out $R_{MC} = 2000$ replications for each experiment.

5.2 Description of individual methods

We consider the OCMT, Lasso, Hard thresholding, Sica and boosting methods described in detail below. With the exception of the OCMT procedure all other methods use the set of standardised regressors $\{\tilde{x}_{it}\}$, defined by $\tilde{x}_{it} = (x_{it} - \bar{x}_i)/s_{xi}$, for $i = 1, 2, ..., n$, $t = 1, 2, ..., T$, where $\bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}$ and $s_{xi}^2 = T^{-1}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)^2$. OCMT does not require any standardisation and we use the original (non-standardised) data, but include an intercept in the regressions. It is worth noting the OCMT procedure is unaffected by scaling of the regressors, but the same is not true of penalised regression techniques.

5.2.1 OCMT method

The OCMT method is implemented as outlined in Section 4.6. We use critical value function, $c_p(n)$, defined by (7) with $f(n) = n^{\delta}$ and consider two choices for $\delta = 1$ and 1.25, and three choices for $p = 0.1$, 0.05, and 0.01, which gives six critical values in total. The choice of p did not matter much and in what follows we only report the results for $p = 0.01$ but provide a full set of results for all combinations of p and δ in an online Monte Carlo Supplement.

5.2.2 Penalised regression methods

Penalised regressions are implemented solving the following optimization problem,[3]

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}),\ Q(\boldsymbol{\beta}) = (2T)^{-1}\sum_{t=1}^{T}\left(\tilde{y}_t - \sum_{i=1}^{n}\beta_i\tilde{x}_{it}\right)^2 + \|P_{\lambda}(\boldsymbol{\beta}_n)\|_1,$$

[3]We used the same Matlab codes for the Lasso, Hard thresholding and Sica penalised regression methods as in Zheng, Fan, and Lv (2014). We are grateful to these authors for providing us with their codes.

where $\tilde{y}_t = y_t - T^{-1}\sum_{t=1}^{T} y_t$ and $P_\lambda(\boldsymbol{\beta}_n) = [p_\lambda(|\beta_1|), p_\lambda(|\beta_2|), ..., p_\lambda(|\beta_n|)]'$. Depending on the choice of the penalty function, we have:

$$\text{Lasso: } p_\lambda(\beta) = \lambda\beta$$

$$\text{Sica: } p_\lambda(\beta, a) = \lambda(a+1)\beta/(a+\beta), \text{ with a small shape parameter } a = 10^{-4}$$

$$\text{Hard thresholding: } p_\lambda(\beta) = \frac{1}{2}\left\{\lambda^2 - (\lambda-\beta)_+^2\right\}, \beta \geq 0.$$

These penalty functions are popular in the literature, see, e.g., Tibshirani (1996), Lv and Fan (2009), and Zheng, Fan, and Lv (2014). We consider the same set of possible values for the penalization parameter λ as in Zheng, Fan, and Lv (2014), namely $\lambda \in \Lambda \equiv \{\lambda_{\min}, \lambda_{\min} + \lambda_\epsilon, \lambda_{\min} + 2\lambda_\epsilon, ..., \lambda_{\max}\}$, where

$$\lambda_{\max} = \max_{i=1,2,...,n} \left|T^{-1}\tilde{\boldsymbol{x}}_i'\tilde{\boldsymbol{y}}\right|, \lambda_{\min} = \epsilon\lambda_{\max}, \ \tilde{\boldsymbol{y}} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_T)'$$

$$\epsilon = \begin{cases} 0.001, & \text{for } n \leq T \\ 0.01, & \text{for } n > T \end{cases},$$

and $\lambda_\epsilon = (\lambda_{\max} - \lambda_{\min})/(K-1)$, with $K = 50$. Following the literature, we select λ using 10-fold cross-validation. That is, we divide the available sample into 10 sub-samples of equal length. One at a time, one sub-sample is used for validation and the remaining 9 for training. This gives us 10 different selected values of λ, which we then average, and this average is denoted as $\hat{\lambda}_a$. We then choose $\lambda = \arg\min_{\lambda\in\Lambda}\left|\lambda - \hat{\lambda}_a\right|$.

5.2.3 Boosting

We consider the boosting algorithm proposed by Buhlmann (2006). This algorithm can be described as follows

Algorithm 1 *1. (initialization). Let* $\tilde{\boldsymbol{x}}_{nt} = (\tilde{x}_{1t}, \tilde{x}_{2t}, ..., \tilde{x}_{nt})'$, $\tilde{\boldsymbol{X}}_n = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_n)$ *and* $\boldsymbol{e} = (e_1, e_2, ..., e_T)'$. *Define the least squares base procedure:*

$$\hat{g}_{\tilde{\boldsymbol{X}},\boldsymbol{e}}(\tilde{\boldsymbol{x}}_{nt}) = \hat{\delta}_{\hat{s}}\tilde{x}_{\hat{s}t}, \ \hat{s} = \arg\min_{1\leq i\leq n}\left(\boldsymbol{e} - \hat{\delta}_i\tilde{\boldsymbol{x}}_i\right)'\left(\boldsymbol{e} - \hat{\delta}_i\tilde{\boldsymbol{x}}_i\right), \ \ \hat{\delta}_i = \frac{\boldsymbol{e}'\tilde{\boldsymbol{x}}_i}{\tilde{\boldsymbol{x}}_i'\tilde{\boldsymbol{x}}_i},$$

2. Given data $\tilde{\boldsymbol{X}}_n$ *and* $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_T)'$,*apply the base procedure to obtain* $\hat{g}^{(1)}_{\tilde{\boldsymbol{X}},\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{x}}_{nt})$. *Set* $\hat{F}^{(1)}(\tilde{\boldsymbol{x}}_{nt}) = v\hat{g}^{(1)}_{\tilde{\boldsymbol{X}},\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{x}}_{nt})$, *for some* $v > 0$, *Set* $\hat{s}^{(1)} = \hat{s}$ *and* $m = 1$.

3. Compute the residual vector $\boldsymbol{e} = \tilde{\boldsymbol{y}} - \hat{F}^{(m)}(\tilde{\boldsymbol{X}}_n)$, *with* $\hat{F}^{(m)}(\tilde{\boldsymbol{X}}_n) = (\hat{F}^{(m)}(\tilde{\boldsymbol{x}}_{n1}), \hat{F}^{(m)}(\tilde{\boldsymbol{x}}_{n2}), ..., \hat{F}^{(m)}(\tilde{\boldsymbol{x}}_{nT}))'$, *and fit the base procedure to these residuals to obtain the fit values* $\hat{g}^{(m+1)}_{\tilde{\boldsymbol{X}},\boldsymbol{e}}(\tilde{\boldsymbol{x}}_{nt})$ *and* $\hat{s}^{(m)}$. *Update*

$$\hat{F}^{(m+1)}(\tilde{\boldsymbol{x}}_{nt}) = \hat{F}^{(m)}(\tilde{\boldsymbol{x}}_{nt}) + v\hat{g}^{(m+1)}_{\boldsymbol{X},\boldsymbol{e}}(\tilde{\boldsymbol{x}}_{nt}).$$

4. Increase the iteration index m by one and repeat step 3 until the stopping iteration M is achieved. The stopping iteration is given by

$$M = \arg \min_{1 \leq m \leq m_{\max}} AIC_C(m),$$

for some predetermined large $m_{\max}$, where

$$AIC_C(m) = \ln(\hat{\sigma}^2) + \frac{1 + tr\left(\mathcal{B}_m\right)/T}{1 - \left(tr\left(\mathcal{B}_m\right) + 2\right)/T}$$

$$\hat{\sigma}^2 = \frac{1}{T}\left(\boldsymbol{y} - \mathcal{B}_m \tilde{\boldsymbol{y}}\right)'\left(\boldsymbol{y} - \mathcal{B}_m \tilde{\boldsymbol{y}}\right)$$

$$\mathcal{B}_m = I - \left(I - v\mathcal{H}^{(\hat{s}_m)}\right)\left(I - v\mathcal{H}^{(\hat{s}_{m-1})}\right) ... \left(I - v\mathcal{H}^{(\hat{s}_1)}\right)$$

$$\mathcal{H}^{(j)} = \frac{\tilde{\boldsymbol{x}}_j \tilde{\boldsymbol{x}}_j'}{\tilde{\boldsymbol{x}}_j' \tilde{\boldsymbol{x}}_j}.$$

We set $m_{\max} = 500$ and consider two values for the tuning parameter: $v = 0.1$ and 1. The former is suggested in Buhlmann (2006).

5.3 Summary statistics for MC results

We evaluate the small sample performance of individual methods, using a number of criteria. In particular, we report the following summary statistics:

1. The true positive rate (TPR) defined by (10), and the false positive rate (FPR) define by (11).

2. The out-of-sample root mean square forecast error *relative* to that of the benchmark true model estimated by least squares using the only the signal variables, which we denote by rRMSFE.

3. The root mean square error of $\tilde{\boldsymbol{\beta}}$ relative to the true benchmark model, denoted by $\text{rRMSE}_{\tilde{\boldsymbol{\beta}}}$.

4. The probability (frequency) of selecting at least all the k signal variables, denoted by $\hat{\pi}_k$, and the the probability of at least selecting all the signal and pseudo-true signals (if any), denoted by $\hat{\pi}_{k+k^*}$.

5. The probability of selecting the true model, denoted by $\hat{\pi}$, and the probability of selecting pseudo-true model with all pseudo-signals, denoted by $\hat{\pi}^*$.

6. The following summary statistics are also reported on frequency distribution of the number of selected covariates, $\hat{\kappa}$:

(a) $\overline{\hat{\kappa}}$, the average number of selected covariates, denoted by $\overline{\hat{\kappa}}$;

(b) $\hat{\kappa}_5$, the 5^{th} quantile of the distribution of $\hat{\kappa}$;

(c) $\hat{\kappa}_{95}$, the 95^{th} quantile of the distribution of $\hat{\kappa}$;

(d) $\hat{\kappa}_{\max}$, the largest number of selected covariates.

rRMSFE is computed based on 100 out-of-sample forecasts. In the case of the OCMT method, we also report $r = J - 1$, the number of iterations of OCMT method before convergence.

Remark 18 *In the case of the approximately sparse DGP-V, TPR and FPR are computed assuming the first 11 covariates (that have coefficients* $1, 1/2, 1/2^2, ..., 1/11^2$*) are signal variables and the remaining covariates having coefficients* $\beta_i = 1/i^2$*, for* $i > 11$ *as noise variables.*[4]

5.4 MC findings

We present the MC findings in two parts. First, we consider the relative performance of the OCMT method compared to the penalised regression and boosting techniques, and also report some statistics on the relative computational times involved across the different methods. These comparisons are carried out in the case DGPs with Gaussian covariates and Gaussian errors. Next, we investigate the robustness of the OCMT procedure to non-Gaussian errors and serially correlated covariates. Penalised regressions are not computed for these experiments due to their high computational burden.

5.4.1 Comparison of OCMT method with penalised regression and boosting methods

Consider first the designs with zero correlations between signal and noise variables (design I). Table 1 reports the findings for $n \in \{100, 300\}$, averaged across $R^2 \in \{0.3, 0.5, 0.7\}$ and $T \in \{100, 300, 500\}$ to economize on space. The full set of results for different values of R^2 and T are available in an online Monte Carlo Supplement. Table 1 reports the results for the OCMT method with $\delta = 1$ and $p = 0.01$ and compares them with those obtained using penalised regressions and boosting techniques.[5] The findings for DGP-I(a),(b) and (c) in Table 1 are very similar and can be summarized as follows. OCMT has the best TPR/FPR trade-off, the lowest average relative root mean square forecast error (< 1.004) and the highest average probability of selecting the true model (0.89-0.92). The average probability of selecting

[4] In choosing the threshold $i = 11$, we were guided by the fact that $|\beta_i| / \sqrt{Var(\varsigma u_t)}$, which is a good measure of the strength of the signal, exceeds 0.01 only for $i \leq 11$ when $R^2 = 70\%$.

[5] Findings for other choices of δ and p are very similar and are reported in the online Monte Carlo supplement.

the true model is very low for other methods. In particular, the Lasso tends to select more regressors (about 8-12, on average), and the average probability of selecting the correct model is only 0.05-0.12. In contrast, Sica and Hard thresholding tend to under-select, but have higher probability of selecting the correct model (0.20-0.37) than Lasso, although these probabilities are still much lower than those achieved by OCMT for these experiments. In the case of boosting methods, we show findings only for $v = 0.1$, a choice recommended by Buhlmann (2006). The boosting tend to over-select even more heavily than Lasso and, as a result, its probability of selecting the true model is very small, often near zero. This seems to be a general feature of boosting. It holds across all of the experiments that we consider and is not much affected if we use a larger value of v. In the online Monte Carlo Supplement, we provide further results for boosting using $v = 1$.

Decreasing the collinearity among the covariates from $\omega = 0.5$ in the case of DGP-I(a) to $\omega = 0.2$ in the case of DGP-I(d) has opposite effects on the performance of the OCMT and penalised regressions. Decreasing ω reduces the magnitude of θ_i and therefore lowers the power of selecting the signals with the OCMT method. The average probability of selecting the correct model with OCMT drops to 0.79-0.82 in DGP-I(d) from 0.91-0.92 in DGP-I(a). For the penalised regressions, on the other hand, we see slight improvements with a fall in the collinearity of the signal variables. One possible explanation for this is that the marginal contribution of signals to the overall fit of the model has increased, which resulted in a better performance of the penalised regression methods. We observe an increase in $\hat{\pi}$ which ranges between 0.02 and 0.63, depending on the choice of the penalty function. The findings for design DGP-I(d) with a high ($\omega = 0.8$) pair-wise collinearity of signals (reported in the online Monte Carlo Supplement) show a substantial improvement in OCMT and a deterioration in the penalised regression methods, as to be expected.

We turn next to the experiments with non-zero correlations between signal and noise variables (design II). The concepts of true and pseudo-true models (selected by OCMT) do not coincide in these experiments, but the OCMT estimator of $\boldsymbol{\beta}_n$, namely $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_n)'$, with $\tilde{\beta}_i$ defined by (9), is still consistent (see Theorem 5 and Remark 16). Table 2 reports findings for DGP-II(a) featuring 2 pseudo-signals and DGP-II(b) featuring all noise variables correlated with signals. The OCMT procedure continues to perform well in these designs, and the true and false positive rate trade-off seems to be the best among the methods considered. Similarly to DGP-I, Lasso and boosting continue to over-select and the Hard and Sica methods under-select the true number of signals.

We now consider the findings for the experiments with zero net effects (design III). For these experiments, the signals with zero net effect will not be picked up in the first stage of OCMT method (even asymptotically). Nevertheless, such signals do get picked up with a high probability at the second or higher stages of the OCMT procedure. This feature is clearly

seen from the findings presented in Table 3, where the average probability of selecting the correct model using the OCMT method continues to be much higher than those obtained for the penalised methods. It is also worth noting that the average number of iterations needed for OCMT to converge ($r = J - 1$) only increases marginally to slightly above 1. OCMT continues to have the best performance on average in terms of RMSFE and rRMSE of $\tilde{\boldsymbol{\beta}}_n$.

We turn next to experiments with zero net effects as well as pseudo-signals (design IV) summarized in Table 4. As expected, the probability of selecting the correct model, $\hat{\pi}$, dropped to 0 in the case of OCMT method, due to the presence of pseudo-signals. Similarly to Table 2, the probability of selecting the pseudo-true model $\hat{\pi}^*$ remain high and the OCMT method continue to have the best forecasting performance and TPR/FPR trade-off.

Finally, we consider the experiments with an unbounded number of signals (design V). There are $k = n$ signals in these experiments, but only a few of the signals are strong. For these experiments we compute TPR and FPR statistic assuming that the first 11 covariates with coefficients $\beta_i = 1/i^2$, for $i = 1, 2, ..., 11$ are the 'true' signals. We also report the root mean square forecast error and RMSE of $\tilde{\boldsymbol{\beta}}_n$ relative to the benchmark model which feature the first 11 covariates only. Findings reported in Table 5 show that OCMT continues to achieve the best forecasting performance and the lowest RMSE.

Overall, the small sample evidence suggests that the OCMT method outperforms the penalised regressions that have become the de facto benchmark in the literature, at least in the case of the experiments considered in this paper. Another important advantage of the OCMT procedure is that it is easy to implement and very fast to compute. Table 6 shows relative computational times in the case of DGP-II(b), which features the type of covariance regressor matrix commonly employed in the literature.[6] The OCMT method is about 10^2 to 10^4 times faster than penalised regression methods, and about 50 times faster compared to boosting.

5.4.2 Robustness of OCMT method to non-Gaussianity and serial correlation

Findings presented so far correspond to experiments with Gaussian (G) innovations and, with the exception of DGP-I(b), serially uncorrelated (SU) covariates (we refer to these experiments as G-SU). We now consider additional experiments to investigate the robustness of OCMT method to non-Gaussianity and highly serially correlated covariates. In particular, we consider three additional sets of experiments: non-Gaussian innovations with serially uncorrelated covariates (NG-SU), Gaussian innovations with serially correlated covariates (G-SC), and non-Gaussian innovations with serially correlated covariates (NG-SC). Serially correlated

[6] Computational times are similar across the individual DGPs.

covariates in the case of G-SC and NG-SC experiments are generated using

$$\varepsilon_{it} = 0.9\varepsilon_{i,t-1} + \sqrt{1-0.9^2}e_{it}, \tag{43}$$

where e_{it} are generated as independent draws from $N(0,1)$ or $[\chi^2_t(2)-2]/2$. We set the autoregressive coefficient in (43) to a relatively high value of 0.9, since the moderately low value of 0.5 in DGP-I(b) did not have any substantial impact on the findings. As before, we report findings for $n \in \{100, 300\}$, and average individual summary statistics across $R^2 \in \{0.3, 0.5, 0.7\}$ and $T \in \{100, 300, 500\}$. To economize on space further, we only report findings for rRMSFE and rRMSE of $\tilde{\beta}_n$ in the body of the paper, see Tables 7 and 8, respectively. The full set of results is reported in the online MC Supplement.

The results for the forecasting performance are reported in Table 7. The ones with Gaussian innovations are reported under columns labeled "G-SU" and "G-SC", and those with non-Gaussian innovations under columns labelled "NG-SU" and "NG-SC". According to these results, comparing "G-SU" with "NG-SU" and "G-SC" with "NG-SC", the effects of allowing for non-Gaussian innovations seem to be rather marginal. The deterioration in the relative forecasting performance is very small for both reported sets of critical values, $p = 0.01$ and $\delta = 1$ or 1.25. In contrast, comparing "G-SU" with "G-SC" and "NG-SU" with NG-SC", the deterioration in performance due to serial correlation of covariates is much larger (up to 35%, depending on the design). This is because longer time series observations are needed to detect spurious correlation when the covariates are highly serially correlated (in the present set of experiments set to 0.90). Findings for rRMSE of $\tilde{\boldsymbol{\beta}}_n$ in Table 8 are qualitatively similar, but show much larger deterioration in relative performance in the case of the serially correlated covariates.

6 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. A distinctive characteristic of the method is the consideration of regressors individually rather than within a multiple regression setting. In this sense, there are other alternatives that could be considered such as versions of boosting. A formal extension of the method to serially correlated covariates along the lines considered in Section 4.7 would also be welcome. A further possibility is to extend the idea of considering regressors individually to other testing frameworks, such as tests of forecasting ability. Finally, it is also important that the performance of the OCMT approach is evaluated in empirical contexts. It is hoped that the theoretical results and the Monte Carlo evidence presented in this paper provide a sound basis for such further developments and applications.

Table 1: Monte Carlo findings for experiments with zero correlation between signal and noise variables (design set I)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\boldsymbol{\beta}}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\overline{\hat{\kappa}}$	$\hat{\kappa}_5$	$\hat{\kappa}_{95}$	$\hat{\kappa}_{\max}$	r
	DGP-I(a): Temporally uncorrelated and weakly collinear regressors											
OCMT	100	0.9769	0.0003	1.002	1.084	0.95	0.92	3.9	3.7	4.0	5.9	0.012
	300	0.9681	0.0001	1.003	1.129	0.93	0.91	3.9	3.7	4.0	5.6	0.012
Lasso	100	0.9723	0.0541	1.021	1.513	0.91	0.09	9.1	3.9	17.9	35.7	-
	300	0.9669	0.0282	1.029	1.715	0.89	0.06	12.2	3.9	27.4	59.4	-
Sica	100	0.6818	0.0016	1.050	5.692	0.40	0.36	2.9	1.8	4.3	10.7	-
	300	0.6440	0.0005	1.059	6.551	0.36	0.33	2.7	1.8	4.2	11.3	-
Hard	100	0.6805	0.0050	1.054	5.511	0.34	0.23	3.2	2.0	5.7	15.0	-
	300	0.6221	0.0011	1.065	6.695	0.27	0.21	2.8	1.8	4.9	12.0	-
Boosting	100	0.9850	0.3360	1.062	3.726	0.94	0.00	36.2	27.3	45.4	54.3	-
	300	0.9813	0.2750	1.115	6.691	0.93	0.00	85.3	77.6	93.1	101.6	-
	DGP-I(b): Temporally correlated and weakly collinear regressors											
OCMT	100	0.9768	0.0003	1.002	1.087	0.94	0.92	3.9	3.7	4.0	5.9	0.010
	300	0.9663	0.0001	1.004	1.140	0.93	0.89	3.9	3.6	4.1	6.0	0.013
Lasso	100	0.9710	0.0557	1.021	1.501	0.90	0.08	9.2	3.9	18.3	36.6	-
	300	0.9675	0.0296	1.028	1.705	0.89	0.05	12.6	4.1	27.6	60.2	-
Sica	100	0.6731	0.0017	1.055	6.019	0.39	0.35	2.9	1.8	4.3	11.0	-
	300	0.6363	0.0006	1.065	6.728	0.35	0.32	2.7	1.7	4.0	11.7	-
Hard	100	0.6727	0.0054	1.058	5.682	0.33	0.23	3.2	2.0	5.9	14.9	-
	300	0.6141	0.0012	1.070	6.846	0.26	0.20	2.8	1.8	4.9	12.0	-
Boosting	100	0.9835	0.3224	1.064	3.629	0.94	0.00	34.9	25.7	44.3	53.8	-
	300	0.9807	0.2581	1.118	6.419	0.93	0.00	80.3	72.4	88.3	97.8	-
	DGP-I(c): Strongly collinear and persistent noise variables											
OCMT	100	0.9761	0.0002	1.002	1.159	0.94	0.93	3.9	3.7	4.0	8.4	0.007
	300	0.9682	0.0001	1.003	1.297	0.93	0.91	3.9	3.7	4.0	18.6	0.009
Lasso	100	0.9737	0.0415	1.018	1.453	0.91	0.12	7.9	3.9	15.1	37.8	-
	300	0.9711	0.0211	1.024	1.598	0.90	0.08	10.1	3.9	21.9	51.1	-
Sica	100	0.6895	0.0016	1.049	5.843	0.41	0.37	2.9	1.8	4.2	11.4	-
	300	0.6546	0.0005	1.057	6.454	0.37	0.34	2.8	1.8	4.1	12.4	-
Hard	100	0.7103	0.0051	1.048	5.134	0.38	0.26	3.3	2.1	5.9	15.4	-
	300	0.6515	0.0012	1.060	6.078	0.30	0.24	3.0	1.9	5.3	12.2	-
Boosting	100	0.9869	0.3277	1.059	5.258	0.95	0.00	35.4	25.9	43.9	51.4	-
	300	0.9835	0.2125	1.091	6.949	0.94	0.00	66.8	58.1	75.4	86.9	-
	DGP-I(d): $\omega = 0.2$											
OCMT	100	0.9183	0.0003	1.015	1.711	0.84	0.82	3.7	3.3	4.0	5.8	0.020
	300	0.8984	0.0001	1.020	1.968	0.81	0.79	3.6	3.1	3.9	5.9	0.024
Lasso	100	0.9848	0.0791	1.029	2.576	0.95	0.03	11.5	4.9	21.5	40.6	-
	300	0.9799	0.0404	1.041	3.170	0.94	0.02	15.9	5.3	32.6	60.8	-
Sica	100	0.8770	0.0021	1.030	3.420	0.70	0.63	3.7	2.9	4.7	11.1	-
	300	0.8512	0.0008	1.038	3.912	0.65	0.60	3.6	2.8	5.0	11.1	-
Hard	100	0.8794	0.0033	1.032	3.459	0.70	0.60	3.8	2.9	5.3	11.9	-
	300	0.8399	0.0009	1.043	4.365	0.63	0.56	3.6	2.8	5.0	11.0	-
Boosting	100	0.9951	0.3399	1.065	5.391	0.98	0.00	36.6	28.0	45.7	55.6	-
	300	0.9914	0.2699	1.119	9.648	0.97	0.00	83.8	76.2	91.6	100.6	-

Notes: There are $k = 4$ signal variables ($i = 1, 2, 3, 4$) and $k^* = 0$ pseudo-signal variables. TPR (FPR) is the true (false) positive rate, rRMSFE is the root mean square forecast error relative to the true benchmark model, $\text{rRMSE}_{\tilde{\boldsymbol{\beta}}}$ is the root mean square error of $\tilde{\boldsymbol{\beta}}$ relative to the true benchmark model, $\hat{\pi}_k$ is the probability that signal variables $i = 1, 2, ..., k$ are among the selected variables, $\hat{\pi}$ is the probability of selecting the true model (featuring the first k covariates), $\overline{\hat{\kappa}}$ is the average number of selected covariates, $\hat{\kappa}_5$ and $\hat{\kappa}_{95}$, respectively, are the 5^{th} and the 95^{th} quantiles of the distribution of the number of selected covariates, and $\hat{\kappa}_{\max}$ is the largest number of selected covariates. This table reports OCMT for $p = 0.01$ and $\delta = 1$ and Boosting for $v = 0.1$. See Section 5 for details.

Table 2: Monte Carlo findings for experiments with non-zero correlations between signal and pseudo-signal variables (design set II)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	rRMSE$_{\tilde{\boldsymbol{\beta}}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\hat{\pi}_{k+k^*}$	$\hat{\pi}^*$	$\overline{\hat{\kappa}}$	$\hat{\kappa}_5$	$\hat{\kappa}_{95}$	$\hat{\kappa}_{\max}$	r
	DGP-II(a): Two pseudo-signal variables													
OCMT	100	0.9768	0.0194	1.006	1.862	0.95	0.01	0.87	0.85	5.8	5.2	6.0	7.7	0.010
	300	0.9667	0.0061	1.007	1.842	0.93	0.02	0.85	0.83	5.7	5.1	6.0	7.6	0.014
Lasso	100	0.9650	0.0577	1.022	1.807	0.88	0.06	0.05	0.00	9.4	3.9	18.6	37.8	-
	300	0.9604	0.0293	1.029	1.947	0.87	0.05	0.04	0.00	12.5	4.1	27.6	58.6	-
Sica	100	0.6685	0.0020	1.052	6.129	0.38	0.35	0.00	0.00	2.9	1.8	4.3	11.9	-
	300	0.6303	0.0006	1.061	6.979	0.34	0.32	0.00	0.00	2.7	1.8	4.0	10.0	-
Hard	100	0.6650	0.0057	1.055	6.320	0.31	0.22	0.00	0.00	3.2	2.0	5.8	14.3	-
	300	0.6077	0.0012	1.067	7.421	0.25	0.20	0.00	0.00	2.8	1.8	4.8	10.9	-
Boosting	100	0.9788	0.3377	1.062	3.984	0.92	0.00	0.14	0.00	36.3	27.5	45.7	56.0	-
	300	0.9743	0.2760	1.116	6.860	0.91	0.00	0.12	0.00	85.6	77.7	93.7	101.8	-
	DGP-II(b): All noise variables collinear with signals													
OCMT	100	0.9514	0.0059	1.007	1.349	0.88	0.39	-	-	4.4	3.6	5.2	7.1	0.013
	300	0.9376	0.0017	1.009	1.417	0.86	0.41	-	-	4.3	3.6	5.2	6.9	0.016
Lasso	100	0.9737	0.0644	1.025	1.843	0.91	0.05	-	-	10.1	4.0	19.7	40.3	-
	300	0.9679	0.0334	1.034	2.148	0.90	0.03	-	-	13.8	4.6	29.8	61.6	-
Sica	100	0.7402	0.0016	1.041	5.408	0.47	0.43	-	-	3.1	2.2	4.4	11.1	-
	300	0.7054	0.0006	1.049	6.249	0.42	0.39	-	-	3.0	2.0	4.3	12.0	-
Hard	100	0.7207	0.0038	1.047	5.849	0.39	0.30	-	-	3.2	2.1	5.3	13.6	-
	300	0.6656	0.0009	1.059	7.175	0.32	0.27	-	-	2.9	2.0	4.8	10.6	-
Boosting	100	0.9884	0.3695	1.068	4.618	0.96	0.00	-	-	39.4	29.4	49.2	57.4	-
	300	0.9833	0.2715	1.114	7.153	0.94	0.00	-	-	84.3	76.7	92.1	101.8	-

Notes: There are $k = 4$ signal variables ($i = 1, 2, 3, 4$), and $k^* = 2$ pseudo-signal variables ($i = 5, 6$) in the case of DGPII(a), whereas all noise variables are collinear with signals in the case of DGPII(b). See notes to Table 1 for a brief summary of the reported statistics. This table reports OCMT for $p = 0.01$ and $\delta = 1$ and Boosting for $v = 0.1$. See Section 5 for details.

Table 3: Monte Carlo findings for experiments with zero net signal effects (design set III)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	rRMSE$_{\tilde{\boldsymbol{\beta}}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\overline{\hat{\kappa}}$	$\hat{\kappa}_5$	$\hat{\kappa}_{95}$	$\hat{\kappa}_{\max}$	r
OCMT	100	0.9184	0.0003	1.017	2.020	0.86	0.84	3.7	3.2	4.0	5.8	0.920
	300	0.9015	0.0001	1.022	2.290	0.84	0.81	3.6	3.2	4.1	5.9	0.902
Lasso	100	0.9600	0.1367	1.056	5.663	0.89	0.00	17.0	7.6	29.3	46.4	-
	300	0.9394	0.0679	1.080	7.857	0.84	0.00	23.9	9.3	43.9	80.9	-
Sica	100	0.9069	0.0024	1.026	3.010	0.81	0.73	3.9	3.1	4.9	12.6	-
	300	0.8737	0.0010	1.039	3.824	0.77	0.70	3.8	2.9	5.2	12.7	-
Hard	100	0.8587	0.0045	1.045	5.140	0.71	0.57	3.9	2.7	5.7	15.6	-
	300	0.7975	0.0012	1.065	7.185	0.62	0.54	3.5	2.4	5.2	10.7	-
Boosting	100	0.9938	0.3606	1.078	5.164	0.98	0.00	38.6	30.2	47.1	55.0	-
	300	0.9821	0.2559	1.135	8.621	0.94	0.00	79.7	72.3	87.3	95.8	-

Notes: There are 4 signal variables ($i = 1, 2, 3, 4$) of which the last one has zero net effect ($\theta_4 = 0$). See notes to Table 1 for a brief summary of the reported statistics. This table reports OCMT for $p = 0.01$ and $\delta = 1$ and Boosting for $v = 0.1$. See Section 5 for details.

Table 4: Monte Carlo findings for experiments with zero net signal effects and pseudo-signals (design set IV)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	rRMSE$_{\tilde{\boldsymbol{\beta}}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\hat{\pi}_{k+k^*}$	$\hat{\pi}^*$	$\overline{\hat{\kappa}}$	$\hat{\kappa}_5$	$\hat{\kappa}_{95}$	$\hat{\kappa}_{\max}$	r
	DGP-IV(a): Two pseudo-signal variables													
OCMT	100	0.9198	0.0174	1.020	2.649	0.86	0.03	0.74	0.72	5.3	4.6	5.9	7.9	0.919
	300	0.9034	0.0055	1.024	2.904	0.84	0.03	0.71	0.69	5.2	4.3	5.9	7.3	0.903
Lasso	100	0.9544	0.1393	1.056	6.106	0.88	0.00	0.09	0.00	17.2	7.8	29.1	47.7	-
	300	0.9324	0.0689	1.081	8.301	0.83	0.00	0.06	0.00	24.1	9.6	44.2	75.3	-
Sica	100	0.8925	0.0029	1.028	3.807	0.77	0.70	0.00	0.00	3.9	3.1	4.9	10.7	-
	300	0.8600	0.0011	1.041	4.636	0.73	0.67	0.00	0.00	3.8	3.0	5.0	11.9	-
Hard	100	0.8282	0.0060	1.050	9.708	0.62	0.50	0.00	0.00	3.9	2.8	5.8	14.3	-
	300	0.7682	0.0016	1.071	11.540	0.54	0.46	0.00	0.00	3.6	2.4	5.4	11.7	-
Boosting	100	0.9894	0.3623	1.078	5.706	0.96	0.00	0.17	0.00	38.7	30.1	47.2	54.9	-
	300	0.9772	0.2571	1.135	9.164	0.93	0.00	0.13	0.00	80.0	72.4	87.8	96.0	-
	DGP-IV(b): All noise variables collinear with signals													
OCMT	100	0.8921	0.0076	1.016	2.325	0.72	0.16	-	-	4.3	3.4	5.2	6.8	0.730
	300	0.8729	0.0022	1.020	2.558	0.68	0.17	-	-	4.2	3.3	5.1	6.6	0.697
Lasso	100	0.9287	0.0982	1.042	4.237	0.79	0.01	-	-	13.1	5.0	24.7	44.1	-
	300	0.9046	0.0481	1.057	5.451	0.71	0.00	-	-	17.9	5.3	36.4	72.2	-
Sica	100	0.7829	0.0019	1.037	6.120	0.63	0.57	-	-	3.3	2.1	4.7	12.0	-
	300	0.7424	0.0007	1.047	6.762	0.57	0.53	-	-	3.2	1.9	4.7	12.4	-
Hard	100	0.7309	0.0038	1.051	7.299	0.54	0.41	-	-	3.3	1.9	5.4	13.1	-
	300	0.6646	0.0009	1.064	9.051	0.45	0.37	-	-	2.9	1.8	5.0	10.4	-
Boosting	100	0.9857	0.3826	1.075	5.335	0.95	0.00	-	-	40.7	31.0	49.9	57.6	-
	300	0.9697	0.2637	1.123	8.150	0.90	0.00	-	-	81.9	74.3	89.8	98.6	-

Notes: There are $k = 4$ signal variables ($i = 1, 2, 3, 4$), and $k^* = 2$ pseudo-signal variables ($i = 5, 6$) in the case of DGPIV(a), whereas all noise variables are collinear with signals in the case of DGPIV(b). See notes to Table 1 for a brief summary of the reported statistics. This table reports OCMT for $p = 0.01$ and $\delta = 1$ and Boosting for $v = 0.1$. See Section 5 for details.

Table 5: Monte Carlo findings for experiments with k=n signal variables (design set V)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_{11}$	$\overline{\hat{\kappa}}$	$\hat{\kappa}_5$	$\hat{\kappa}_{95}$	$\hat{\kappa}_{\max}$	r
OCMT	100	0.2820	0.0003	0.986	0.433	0.00	3.1	2.2	4.1	6.2	0.018
	300	0.2691	0.0001	0.986	0.443	0.00	3.0	2.1	4.1	5.8	0.019
Lasso	100	0.3450	0.0522	1.001	0.570	0.00	8.4	2.7	18.2	38.3	-
	300	0.3121	0.0265	1.008	0.647	0.00	11.1	2.7	26.4	58.8	-
Sica	100	0.1294	0.0011	1.010	1.264	0.00	1.5	1.0	3.0	11.1	-
	300	0.1216	0.0004	1.014	1.351	0.00	1.5	1.0	2.8	10.6	-
Hard	100	0.1231	0.0012	1.012	1.374	0.00	1.5	1.0	2.8	10.6	-
	300	0.1117	0.0003	1.015	1.433	0.00	1.3	1.0	2.4	9.4	-
Boosting	100	0.5751	0.3696	1.045	1.683	0.00	39.2	29.1	49.1	58.3	-
	300	0.5119	0.2731	1.089	2.620	0.00	84.6	76.8	92.4	101.1	-

Notes: Slope coefficients in DGPV are set to $\beta_i = 1/i^2$, for $i = 1, 2, ..., n$. TPR is computed assuming that covariates $i = 1, 2, ..., 11$ are the signal variables , FPR is computed assuming covariates $i > 11$ are the noise variables, rRMSFE is an out-of-sample root mean square forecast error relative to the benchmark model featuring the first 11 covariates, $\text{rRMSE}_{\tilde{\beta}}$ is the root mean square error of $\tilde{\beta}$ relative to the benchmark model featuring the first 11 covariates, and $\hat{\pi}_{11}$ is the probability that covariates $i = 1, 2, ..., 11$ are among the selected covariates. $\overline{\hat{\kappa}}$, $\hat{\kappa}_5$, $\hat{\kappa}_{95}$ and $\hat{\kappa}_{\max}$ are, respectively, the average, 5^{th} quantile, 95^{th} quantile and the maximum of the number of selected covariates. This table reports OCMT for $p = 0.01$ and $\delta = 1$ and Boosting for $v = 0.1$. See Section 5 for details.

Table 6: Computational times relative to OCMT method

Experiments:	DGPII(b), $T = 100$, $R^2 = 50\%$		
	$N = 100$	$N = 200$	$N = 300$
OCMT (benchmark)	1	1	1
Lasso	292	280	226
Hard	713	522	400
Sica	10349	8540	6047
Boosting ($v = 0.1$)	55	66	54

Notes: This table reports computational times relative to OCMT for $p = 0.01$ and $\delta = 1$. Boosting is implemented using $v = 0.1$. See Section 5 for details.

Table 7: Robustness of OCMT to Non-Gaussianity and Serial Correlation (rRMSFE findings)

Summary statistics are averaged across T and R^2

G-SU: Gaussian innovations with serially uncorrelated covariates
NG-SU: non-Gaussian innovations with serially uncorrelated covariates
G-SC: Gaussian innovations with serially correlated covariates
NG-SC: non-Gaussian innovations with serially correlated covariates

		MC findings for rRMSFE							
		$p = 0.01, \delta = 1$				$p = 0.01, \delta = 1.25$			
	n	G-SU	NG-SU	G-SC	NG-SC	G-SU	NG-SU	G-SC	NG-SC
DGP-I(a)	100	1.002	1.004	1.005	1.007	1.002	1.005	1.004	1.006
	300	1.003	1.006	1.01	1.011	1.004	1.007	1.007	1.008
DGP-I(b)	100	1.002	1.004	1.005	1.007	1.003	1.004	1.004	1.006
	300	1.004	1.006	1.009	1.011	1.004	1.006	1.007	1.009
DGP-I(c)	100	1.002	1.003	1.008	1.011	1.002	1.004	1.005	1.008
	300	1.003	1.005	1.025	1.038	1.004	1.006	1.017	1.026
DGP-I(d, $\omega = 0.2$)	100	1.015	1.017	1.038	1.041	1.019	1.021	1.037	1.040
	300	1.020	1.022	1.081	1.082	1.026	1.028	1.065	1.066
DGP-I(d, $\omega = 0.8$)	100	1.001	1.002	1.002	1.002	1.001	1.002	1.001	1.001
	300	1.001	1.004	1.003	1.003	1.001	1.003	1.001	1.002
DGP-II(a)	100	1.006	1.009	1.015	1.016	1.007	1.009	1.013	1.014
	300	1.007	1.010	1.018	1.021	1.008	1.010	1.015	1.018
DGP-II(b)	100	1.007	1.008	1.093	1.095	1.008	1.009	1.078	1.080
	300	1.009	1.012	1.334	1.348	1.011	1.013	1.243	1.232
DGP-III	100	1.017	1.018	1.091	1.093	1.021	1.023	1.084	1.086
	300	1.022	1.025	1.197	1.205	1.028	1.031	1.152	1.155
DGP-IV(a)	100	1.020	1.023	1.097	1.098	1.024	1.027	1.092	1.091
	300	1.024	1.029	1.212	1.213	1.030	1.034	1.168	1.165
DGP-IV(b)	100	1.016	1.018	1.12	1.118	1.019	1.021	1.105	1.104
	300	1.020	1.024	1.386	1.377	1.024	1.027	1.257	1.255
DGP-V	100	0.986	0.986	1.023	1.022	0.986	0.986	1.008	1.008
	300	0.986	0.989	1.239	1.245	0.987	0.988	1.136	1.141

Notes: In the case of DGP-I(b) in G-SU or NG-SU set of experiments, covariates are serially correlated but the extent of serial correlation is low ($\rho_i = 0.5$). The correlation coefficients ρ_i are set equal to 0.9 in G-SC and NG-SC sets of experiments. See notes to Tables 1-5.

Table 8: Robustness of OCMT to Non-Gaussianity and Serial Correlation (rRMSE$_{\hat{\beta}}$ findings)

Summary statistics are averaged across T and R^2

G-SU: Gaussian innovations with serially uncorrelated covariates
NG-SU: non-Gaussian innovations with serially uncorrelated covariates
G-SC: Gaussian innovations with serially correlated covariates
NG-SC: non-Gaussian innovations with serially correlated covariates

		MC findings for rRMSE$_{\tilde{\beta}}$							
		$p = 0.01, \delta = 1$				$p = 0.01, \delta = 1.25$			
	n	G-SU	NG-SU	G-SC	NG-SC	G-SU	NG-SU	G-SC	NG-SC
DGP-I(a)	100	1.084	1.159	1.212	1.224	1.071	1.141	1.122	1.128
	300	1.129	1.252	1.425	1.446	1.115	1.210	1.216	1.234
DGP-I(b)	100	1.087	1.125	1.206	1.218	1.070	1.104	1.119	1.130
	300	1.140	1.215	1.383	1.421	1.111	1.171	1.189	1.225
DGP-I(c)	100	1.159	1.315	5.883	10.888	1.090	1.221	3.389	5.111
	300	1.297	1.543	26.26	34.221	1.141	1.332	42.05	24.531
DGP-I(d, $\omega = 0.2$)	100	1.711	1.813	2.883	2.892	1.863	1.945	2.533	2.574
	300	1.968	2.096	6.487	7.251	2.186	2.259	4.586	4.774
DGP-I(d, $\omega = 0.8$)	100	1.012	1.023	1.033	1.019	0.999	1.005	1.004	1.002
	300	1.014	1.040	1.035	1.035	0.993	1.007	1.007	0.996
DGP-II(a)	100	1.862	1.904	2.124	2.099	1.838	1.859	1.997	1.989
	300	1.842	1.956	2.257	2.404	1.811	1.893	2.027	2.096
DGP-II(b)	100	1.349	1.428	4.57	4.477	1.365	1.424	3.779	3.643
	300	1.417	1.544	25.23	25.999	1.449	1.525	21.12	13.537
DGP-III	100	2.020	2.141	4.951	4.776	2.318	2.435	4.532	4.370
	300	2.290	2.492	12.56	15.919	2.793	2.890	8.425	8.776
DGP-IV(a)	100	2.649	2.782	5.866	5.735	2.888	3.025	5.337	5.253
	300	2.904	3.031	14.66	13.765	3.362	3.363	10.2	10.039
DGP-IV(b)	100	2.325	2.351	6.224	6.238	2.556	2.546	5.551	5.535
	300	2.558	2.637	35.5	35.771	2.876	2.847	16.96	17.192
DGP-V	100	0.433	0.478	1.241	1.213	0.420	0.454	1.008	0.989
	300	0.443	0.504	7.048	7.637	0.428	0.468	3.698	3.804

Notes: See notes to Table 7.

A Appendix

A.1 Proof of theorems

A.1.1 Proof of Theorem 5

We note that $\cup_{i=0,...,k+k^*-1;j=0,...,n-k-k^*}\mathcal{A}_{i,j} = \mathcal{C}$, where

$$\mathcal{C} = \left\{ \left[\sum_{i=1}^{k+k^*} I\left(\widehat{\beta_i \neq 0}\right) < k + k^* \right] \right\}.$$

Then,

$$\begin{aligned} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\|\right) &= E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{C}\right) \Pr\left(\mathcal{C}\right) + \\ &\quad \sum_{j=0}^{n-k-k^*} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{A}_{k+k^*,j}\right) \Pr\left(\mathcal{A}_{k+k^*,j}\right) \\ &= A_{n,T} + B_{n,T}. \end{aligned}$$

Consider first $A_{n,T}$, and note that by (A.102) of Lemma 19 to the regression of y_t on the $\hat{k}$ selected regressors, for some finite positive constant C_0, we have

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{C}\right) \leq C_0 \left(\frac{l_{\max}^4}{T} + l_{\max}\right),$$

where $l_{\max}$ denotes an upper bound to $\hat{k} = \dim\left(\tilde{\boldsymbol{\beta}}_n\right)$. Also, by (A.87), for some finite positive constants C_1 and C_2,

$$\Pr\left(\mathcal{C}\right) \leq \exp\left(-C_1 T^{C_2}\right).$$

Therefore,

$$A_{n,T} \leq C_0 \left(\frac{l_{\max}^4}{T} + l_{\max}\right) \exp\left(-C_1 T^{C_2}\right).$$

Consider now $B_{n,T}$, and note that under $\hat{k} < l_{\max}$, it can be written as

$$\begin{aligned} B_{n,T} &= \sum_{j=0}^{n-k-k^*} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{A}_{k+k^*,j}\right) \Pr\left(\mathcal{A}_{k+k^*,j}\right) \\ &= \sum_{j=0}^{l_{\max}-k-k^*} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{A}_{k+k^*,j}\right) \Pr\left(\mathcal{A}_{k+k^*,j}\right) \\ &\quad + E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \cup_{j=l_{\max}-k-k^*+1}^{n-k-k^*} \mathcal{A}_{k+k^*,j}\right) \Pr\left(\cup_{j=l_{\max}-k-k^*+1}^{n-k-k^*} \mathcal{A}_{k+k^*,j}\right). \end{aligned}$$

Using (A.92) of Lemma 17, it follows that for some $C_0 > 0$.

$$\Pr\left(\cup_{j=l_{\max}-k-k^*+1}^{n-k-k^*} \mathcal{A}_{k+k^*,j}\right) \leq \Pr\left(\hat{k} > l_{\max} - k - k^* + 1\right) \leq C_0 \frac{p\ n}{f\left(n\right)\left(l_{\max} - k - k^* + 1\right)}.$$

and, noting that $\Pr\left(\hat{k}-k-k^*=j\right) \leq \Pr\left(\hat{k}-k-k^*>j-1\right)$, it also follows that

$$\Pr\left(\mathcal{A}_{k+k^*,j}\right) \leq \frac{C_0 np}{j\ f\left(n\right)}. \tag{A.1}$$

Further, by (A.101) of Lemma 19,

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{A}_{k+k^*,j}\right) = C_0\left(\frac{j^4}{T}\right),$$

and

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \cup_{j=l_{\max}-k-k^*+1}^{n-k-k^*} \mathcal{A}_{k+k^*,j}\right) = C_0\left(\frac{l_{\max}^4}{T}\right).$$

Combining the above results gives

$$B_{n,T} = O\left[\left(\frac{l_{\max}^4}{T}\right) n\frac{p}{f\left(n\right)}\right].$$

Hence

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\|\right) = O\left[\left(\frac{l_{\max}^4}{T} + l_{\max}\right)\exp\left(-C_1 T^{C_2}\right)\right] + O\left[\left(\frac{l_{\max}^4}{T}\right)\frac{pn}{f\left(n\right)}\right],$$

which completes the proof.

A.1.2 Proof of Theorem 6

Note that regardless of the number of selected regressors, denoted as $\hat{k}$, $0 \leq \hat{k} \leq n$, 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

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

for any $\hat{k}$. We also note that if for two random variables $x, y > 0$ defined on a probability space, Ω,

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

then $E(x) \leq E(y)$. The above results imply that $E\left\|\tilde{\boldsymbol{u}}\right\|^2 \leq E\left\|\boldsymbol{y}\right\|^2$. Also, by Assumptions 1 and 4, $E\left(y_t^2\right)$ is bounded, and so we have $E\left\|\boldsymbol{y}\right\|^2 = O\left(T\right)$, and therefore $E\left\|\tilde{\boldsymbol{u}}\right\|^2 = O\left(T\right)$.

Now let $\mathcal{A}_0$ be the set of pseudo-true models as defined in Section 4.3 and let $\mathcal{A}_0^c$ be its complement. Then

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

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

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

where C_0 is a finite constant that does not depend on n and/or T. Now using (32), we note that

$$P(\mathcal{A}_0) \geq 1 - n\frac{p}{f(n)} - \exp\left(-C_1 T^{C_2}\right),$$

for some finite positive constants C_1 and C_2, and (assuming $k + k^*$ does not increase with n)

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

in (A.2), we obtain

$$E\left(\frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2\right) \to \sigma^2 \text{ so long as } 1 - n\frac{p}{f(n)} - \exp\left(-C_1 T^{C_2}\right) \to 1. \tag{A.3}$$

Finally, if $n/f(n) = o(1/T)$, it immediately follows that

$$E\left(\frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2\right) - \sigma^2 = O\left(\frac{1}{T}\right), \tag{A.4}$$

which establishes the desired result.

A.2 Lemmas

Lemma 1 *Let $0 < \varkappa \leq 1$, $\delta > 0$, $0 < p < 1$, and consider the critical value function*

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

where $\Phi^{-1}(.)$ is the inverse function of the cumulative standard normal distribution and $f(n) = n^{\delta}$. Then:

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

(*ii*) $\exp\left[-\varkappa c_p^2(n)/2\right] = \ominus\left(n^{-\delta\varkappa}\right)$, and

(*iii*) if $\delta > 1/\varkappa$, then $n\exp\left[-\varkappa c_p^2(n)/2\right] \to 0$ as $n \to \infty$.

Proof. Using Lemma 3 of Bailey, Pesaran, and Smith (2015),

$$c_p(n) \leq \sqrt{2\left[\ln f(n) - \ln p\right]},$$

and therefore for $f(n) = n^{\delta}$, with $\delta > 0$, we have

$$c_p^2(n) \leq 2\left[\delta \ln(n) - \ln(p)\right] = O\left[\ln(n)\right],$$

which establishes result (i). Further, by Proposition 24 of Dominici (2003) we have that

$$\lim_{n\to\infty} c_p(n)/LW\left\{\frac{1}{2\pi\left[\left(1-\frac{p}{2f(n)}\right)-1\right]^2}\right\}^{1/2} = 1,$$

where LW denotes the Lambert W function which satisfies $\lim_{n\to\infty} LW(n)/\left\{\ln(n) - \ln\left[\ln(n)\right]\right\} = 1$ as $n \to \infty$. We note that $\lim_{n\to\infty} \ln(n)/\left\{\ln(n) - \ln\left[\ln(n)\right]\right\} = 1$ as $n \to \infty$. So

$$\lim_{n\to\infty} \frac{LW\left\{\frac{1}{2\pi\left[\left(1-\frac{p}{2f(n)}\right)-1\right]^2}\right\}^{1/2}}{\left\{2\ln\left(\frac{\sqrt{2}f(n)}{\sqrt{\pi}p}\right)\right\}^{1/2}} = 1.$$

Hence, for any $0 < \varkappa \leq 1$,

$$\lim_{n\to\infty} \frac{\exp\left[-\varkappa c_p^2(n)/2\right]}{\exp\left[-\frac{\varkappa\left\{\left[2\ln\left(\frac{\sqrt{2}f(n)}{\sqrt{\pi}p}\right)\right]^{1/2}\right\}^2}{2}\right]} = \lim_{n\to\infty} \frac{\exp\left[-\varkappa c_p^2(n)/2\right]}{\left[f(n)\right]^{-\varkappa}\pi^{\varkappa}p^{2\varkappa}2^{-\varkappa}} = 1 \text{ as } n \to \infty,$$

and substituting n^{δ} for $f(n)$ yields ,

$$\lim_{n\to\infty} \frac{\exp\left[-\varkappa c_p^2(n)/2\right]}{n^{-\delta\varkappa}} \to \frac{2^{\varkappa}}{\pi^{\varkappa}p^{2\varkappa}}. \tag{A.5}$$

It follows from (A.5) that $\exp\left[-\varkappa c_p^2(n)/2\right] = \ominus\left(n^{-\delta\varkappa}\right)$, as required. This completes the proof of result (ii). Finally, it readily follows from (ii) that $n\exp\left[-\varkappa c_p^2(n)/2\right] = \ominus\left(n^{1-\delta\varkappa}\right)$ and therefore $n\exp\left[-\varkappa c_p^2(n)/2\right] \to 0$ when $\delta > 1/\varkappa$, as desired. This completes the proof of the last result. ■

Lemma 2 *Let X_{iT}, for $i = 1, 2, ..., l_T$, Y_T and Z_T be random variables. Then, for some finite positive constants C_0, C_1 and C_2, and any constants π_i, for $i = 1, 2, ..., l_T$, satisfying $0 < \pi_i < 1$ and $\sum_{i=1}^{l_T} \pi_i = 1$, we have*

$$\Pr\left(\sum_{i=1}^{l_T} |X_{iT}| > C_0\right) \leq \sum_{i=1}^{l_T} \Pr\left(|X_{iT}| > \pi_i C_0\right), \tag{A.6}$$

$$\Pr\left(|X_T| \times |Y_T| > C_0\right) \leq \Pr\left(|X_T| > C_0/C_1\right) + \Pr\left(|Y_T| > C_1\right), \tag{A.7}$$

and

$$\begin{aligned}\Pr\left(|X_T| \times |Y_T| \times |Z_T| > C_0\right) \leq \Pr\left(|X_T| > C_0/\left(C_1C_2\right)\right) + \Pr\left(|Y_T| > C_1\right) + \\ \Pr\left(|Z_T| > C_2\right).\end{aligned} \tag{A.8}$$

Proof. Without loss of generality we consider the case $l_T = 2$. Consider the two random variables X_{1T} and X_{2T}. Then, for some finite positive constants C_0 and C_1, we have

$$\begin{aligned}\Pr\left(|X_{1T}| + |X_{2T}| > C_0\right) &\leq \Pr\left(\{|X_{1T}| > (1-\pi)C_0\} \cup \{|X_{2T}| > \pi C_0\}\right) \\ &\leq \Pr\left(|X_{1T}| > (1-\pi)C_0\right) + \Pr\left(|X_{2T}| > \pi C_0\right),\end{aligned}$$

proving the first result of the lemma. Also

$$\begin{aligned}\Pr\left(|X_T| \times |Y_T| > C_0\right) = \Pr\left(|X_T| \times |Y_T| > C_0 \ \ |\{|Y_T| > C_1\}\right) \Pr\left(|Y_T| > C_1\right) + \\ \Pr\left(|X_T| \times |Y_T| > C_0 \ \ |\{|Y_T| \leq C_1\}\right) \Pr\left(|Y_T| \leq C_1\right),\end{aligned}$$

and since

$$\Pr\left(|X_T| \times |Y_T| > C_0 \ \ |\{|Y_T| > C_1\}\right) \leq \Pr\left(|X_T| > C_0/C_1\right),$$

and

$$0 \leq \Pr\left(|X_T| \times |Y_T| > C_0 \ \ |\{|Y_T| \leq C_1\}\right) \leq 1,$$

then

$$\Pr\left(|X_T| \times |Y_T| > C_0\right) \leq \Pr\left(|X_T| > C_0/C_1\right) + \Pr\left(|Y_T| > C_1\right),$$

proving the second result of the lemma. The third result follows by a repeated application of the second result. ■

Lemma 3 *Consider the scalar random variable X_T, and the constants B and C. Then, if $|B| \geq C > 0$,*

$$\Pr\left(|X + B| \leq C\right) \leq \Pr\left(|X| > |B| - C\right). \tag{A.9}$$

Proof. We note that the event we are concerned with is of the form $\mathcal{A} = \{|X + B| \leq C\}$. Consider two cases: (i) $B > 0$. Then, $\mathcal{A}$ can occur only if $X < 0$ and $|X| > B - C = |B| - C$. (ii) $B < 0$. Then, $\mathcal{A}$ can occur only if $X > 0$ and $X = |X| > |B| - C$. It therefore follows that the event $\{|X| > |B| - C\}$ implies $\mathcal{A}$ proving (A.9). ■

Lemma 4 *Consider the scalar random variable, ω_T, and the deterministic sequence, $\alpha_T > 0$, such that $\alpha_T \to 0$ as $T \to \infty$. Then there exists $T_0 > 0$ such that for all $T > T_0$ we have*

$$\Pr\left(\left|\frac{1}{\sqrt{\omega_T}} - 1\right| > \alpha_T\right) \leq \Pr\left(|\omega_T - 1| > \alpha_T\right). \tag{A.10}$$

Proof. We first note that for $\alpha_T < 1/2$

$$\left|\frac{1}{\sqrt{\omega_T}} - 1\right| < |\omega_T - 1| \text{ for any } \omega_T \in [1-\alpha_T, 1+\alpha_T].$$

Also, since $a_T \to 0$ then there must exist a $T_0 > 0$ such that $a_T < 1/2$, for all $T > T_0$, and hence if event $A : |\omega_T - 1| \leq a_T$ is satisfied, then it must be the case that event $B : \left|\frac{1}{\sqrt{\omega_T}} - 1\right| \leq a_T$ is also satisfied for all $T > T_0$. Further, since $A \Rightarrow B$, then $B^c \Rightarrow A^c$, where A^c denotes the complement of A. Therefore, $\left|\frac{1}{\sqrt{\omega_T}} - 1\right| > a_T$ implies $|\omega_T - 1| > a_T$, for all $T > T_0$, and we have $\Pr\left(\left|\frac{1}{\sqrt{\omega_T}} - 1\right| > \alpha_T\right) \leq \Pr\left(|\omega_T - 1| > \alpha_T\right)$, as required. ■

Lemma 5 *Let $\mathbf{A}_T = (a_{ij,T})$ be a symmetric $l_T \times l_T$ matrix with eigenvalues $\mu_1 \leq \mu_2 \leq ... \leq \mu_{l_T}$. Let $\mu_i = \ominus(l_T)$, $i = l_T - M + 1, ..., l_T$, for some finite M, and $\sup_{1 \leq i \leq l_T - M} \mu_i < C_0 < \infty$, for some finite positive C_0. Then,*

$$\|\mathbf{A}_T\|_F = \ominus(l_T). \tag{A.11}$$

If, in addition, $\inf_{1 \leq i < l_T} \mu_i > C_1 > 0$, for some finite positive C_1, then

$$\left\|\mathbf{A}_T^{-1}\right\|_F = \ominus\left(\sqrt{l_T}\right). \tag{A.12}$$

Proof. We have

$$\|\mathbf{A}_T\|_F^2 = Tr\left(\mathbf{A}_T \mathbf{A}_T'\right) = Tr\left(\mathbf{A}_T^2\right) = \sum_{i=1}^{l_T} \mu_i^2,$$

where μ_i, for $i = 1, 2, ..., l_T$, are the eigenvalues of $\mathbf{A}_T$. But by assumption $\mu_i = \ominus(l_T)$, for $i = l_T - M + 1, ..., l_T$, and $\sup_{1 \leq i \leq l_T - M} \mu_i < C_0 < \infty$, then $\sum_{i=1}^{l_T} \mu_i^2 = M \ominus(l_T^2) + O(l_T - M) = \ominus(l_T^2)$, and since M is fixed then (A.11) follows. Note that $\mathbf{A}_T^{-1}$ is also symmetric, and using similar arguments as above, we have

$$\left\|\mathbf{A}_T^{-1}\right\|_F^2 = Tr\left(\mathbf{A}_T^{-2}\right) = \sum_{i=1}^{l_T} \mu_i^{-2},$$

but all eigenvalues of $\mathbf{A}_T$ are bounded away from zero under the assumptions of the lemma, which implies $\mu_i^{-2} = \ominus(1)$ and therefore $\left\|\mathbf{A}_T^{-1}\right\|_F = \ominus\left(\sqrt{l_T}\right)$, which establishes (A.12). ■

Lemma 6 *Let z be a random variable and suppose there exists finite positive constants C_0, C_1 and $s > 0$ such that*

$$\Pr\left(|z| > \alpha\right) \leq C_0 \exp\left(-C_1 \alpha^s\right), \text{ for all } \alpha > 0. \tag{A.13}$$

Then for any finite $p > 0$ and p/s finite, there exists $C_2 > 0$ such that

$$E|z|^p \leq C_2. \tag{A.14}$$

Proof. We have that

$$E\left|z\right|^{p}=\int_{0}^{\infty}\alpha^{p}d\Pr\left(\left|z\right|\leq\alpha\right).$$

Using integration by parts, we get

$$\int_{0}^{\infty}\alpha^{p}d\Pr\left(\left|z\right|\leq\alpha\right)=p\int_{0}^{\infty}\alpha^{p-1}\Pr\left(\left|z\right|>\alpha\right)d\alpha.$$

But, using (A.13), and a change of variables, implies

$$E\left|z\right|^{p}\leq pC_{0}\int_{0}^{\infty}\alpha^{p-1}\exp\left(-C_{1}\alpha^{s}\right)d\alpha=\frac{pC_{0}}{s}\int_{0}^{\infty}u^{\frac{p-s}{s}}\exp\left(-C_{1}u\right)du=C_{0}C_{1}^{-p/s}\left(\frac{p}{s}\right)\Gamma\left(\frac{p}{s}\right),$$

where $\Gamma\left(\cdot\right)$ is a gamma function. But for a finite positive p/s, $\Gamma\left(p/s\right)$ is bounded and (A.14) follows. ■

Lemma 7 *Let $\boldsymbol{A}_{T}=\left(a_{ij,T}\right)$ be an $l_{T}\times l_{T}$ matrix and $\hat{\boldsymbol{A}}_{T}=\left(\hat{a}_{ij,T}\right)$ be an estimator of $\boldsymbol{A}_{T}$. Suppose that $\boldsymbol{A}_{T}$ is invertible and there exists a finite positive C_{0}, such that*

$$\sup_{i,j}\Pr\left(\left|\hat{a}_{ij,T}-a_{ij,T}\right|>b_{T}\right)\leq\exp\left(-C_{0}Tb_{T}^{2}\right),\tag{A.15}$$

for all $b_{T}>0$. Then

$$\Pr\left(\left\|\hat{\boldsymbol{A}}_{T}-\boldsymbol{A}_{T}\right\|_{F}>b_{T}\right)\leq l_{T}^{2}\exp\left(-C_{0}\frac{Tb_{T}^{2}}{l_{T}^{2}}\right),\tag{A.16}$$

and

$$\begin{aligned}\Pr\left(\left\|\hat{\boldsymbol{A}}_{T}^{-1}-\boldsymbol{A}_{T}^{-1}\right\|_{F}>b_{T}\right)\leq{}& l_{T}^{2}\exp\left(\frac{-C_{0}Tb_{T}^{2}}{l_{T}^{2}\left\|\boldsymbol{A}_{T}^{-1}\right\|_{F}^{2}\left(\left\|\boldsymbol{A}_{T}^{-1}\right\|_{F}+b_{T}\right)^{2}}\right)\\&+l_{T}^{2}\exp\left(-C_{0}\frac{T}{\left\|\boldsymbol{A}_{T}^{-1}\right\|_{F}^{2}l_{T}^{2}}\right).\end{aligned}\tag{A.17}$$

Proof. First note that since $b_{T}>0$, then

$$\begin{aligned}\Pr\left(\left\|\hat{\boldsymbol{A}}_{T}-\boldsymbol{A}_{T}\right\|_{F}>b_{T}\right)&=\Pr\left(\left\|\hat{\boldsymbol{A}}_{T}-\boldsymbol{A}_{T}\right\|_{F}^{2}>b_{T}^{2}\right)\\&=\Pr\left(\left[\sum_{j=1}^{l_{T}}\sum_{i=1}^{l_{T}}\left(\hat{a}_{ij,T}-a_{ij,T}\right)^{2}>b_{T}^{2}\right]\right),\end{aligned}$$

and using the probability bound result, (A.6), and setting $\pi_{i}=1/l_{T}$, we have

$$\begin{aligned}\Pr\left(\left\|\hat{\boldsymbol{A}}_{T}-\boldsymbol{A}_{T}\right\|_{F}>b_{T}\right)&\leq\sum_{j=1}^{l_{T}}\sum_{i=1}^{l_{T}}\Pr\left(\left|\hat{a}_{ij,T}-a_{ij,T}\right|^{2}>l_{T}^{-2}b_{T}^{2}\right)\\&=\sum_{j=1}^{l_{T}}\sum_{i=1}^{l_{T}}\Pr\left(\left|\hat{a}_{ij,T}-a_{ij,T}\right|>l_{T}^{-1}b_{T}\right)\\&\leq l_{T}^{2}\sup_{ij=1,2,...,l_{T}}\left[\Pr\left(\left|\hat{a}_{ij,T}-a_{ij,T}\right|>l_{T}^{-1}b_{T}\right)\right].\end{aligned}$$

Hence by (A.15) we obtain (A.16). To establish (A.17) define the sets

$$\mathcal{A}_T = \left\{ \left\| \boldsymbol{A}_T^{-1} \right\|_F \left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F \leq 1 \right\} \text{ and } \mathcal{B}_T = \left\{ \left\| \hat{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}_T^{-1} \right\|_F > b_T \right\}$$

and note that by (2.15) of Berk (1974) if $\mathcal{A}_T$ holds we have

$$\left\| \hat{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}_T^{-1} \right\|_F \leq \frac{\left\| \boldsymbol{A}_T^{-1} \right\|_F^2 \left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F}{1 - \left\| \boldsymbol{A}_T^{-1} \right\|_F \left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F}. \tag{A.18}$$

Hence

$$\begin{aligned} \Pr\left(\mathcal{B}_T \,|\mathcal{A}_T\right) &\leq \Pr\left(\frac{\left\| \boldsymbol{A}_T^{-1} \right\|_F^2 \left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F}{1 - \left\| \boldsymbol{A}_T^{-1} \right\|_F \left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F} > b_T \right) \\ &= \Pr\left(\left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F > \frac{b_T}{\left\| \boldsymbol{A}_T^{-1} \right\|_F \left(\left\| \boldsymbol{A}_T^{-1} \right\|_F + b_T \right)} \right). \end{aligned} \tag{A.19}$$

Note also that

$$\Pr\left(\mathcal{B}_T\right) = \Pr\left(\left\{ \mathcal{B}_T \cap \mathcal{A}_T \right\} \cup \left\{ \mathcal{B}_T \cap \mathcal{A}_T^C \right\} \right) = \Pr\left(\mathcal{B}_T | \mathcal{A}_T\right) \Pr\left(\mathcal{A}_T\right) + \Pr\left(\mathcal{B}_T | \mathcal{A}_T^C\right) \Pr\left(\mathcal{A}_T^C\right). \tag{A.20}$$

Furthermore

$$\begin{aligned} \Pr\left(\mathcal{A}_T^C\right) &= \Pr\left(\left\| \boldsymbol{A}_T^{-1} \right\|_F \left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F > 1 \right) \\ &= \Pr\left(\left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F > \left\| \boldsymbol{A}_T^{-1} \right\|_F^{-1} \right), \end{aligned}$$

and by (A.16) we have

$$\Pr\left(\mathcal{A}_T^C\right) \leq l_T^2 \exp\left(-C_0 \frac{T}{\left\| \boldsymbol{A}_T^{-1} \right\|_F^2 l_T^2} \right).$$

Using the above result and (A.19) in (A.20), we now have

$$\begin{aligned} \Pr\left(\mathcal{B}_T\right) \leq \Pr&\left(\left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F > \frac{b_T}{\left\| \boldsymbol{A}_T^{-1} \right\|_F \left(\left\| \boldsymbol{A}_T^{-1} \right\|_F + b_T \right)} \right) \Pr\left(\mathcal{A}_T\right) \\ &+ \Pr\left(\mathcal{B}_T | \mathcal{A}_T^C\right) l_T^2 \exp\left(-C_0 \frac{T}{\left\| \boldsymbol{A}_T^{-1} \right\|_F^2 l_T^2} \right). \end{aligned}$$

Furthermore, since $\Pr\left(\mathcal{A}_T\right) \leq 1$ and $\Pr\left(\mathcal{B}_T | \mathcal{A}_T^C\right) \leq 1$ then

$$\begin{aligned} \Pr\left(\mathcal{B}_T\right) = \Pr\left(\left\| \hat{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}_T^{-1} \right\|_F > b_T \right) \leq \Pr&\left(\left\| \hat{\boldsymbol{A}}_T - \boldsymbol{A}_T \right\|_F > \frac{b_T}{\left\| \boldsymbol{A}_T^{-1} \right\|_F \left(\left\| \boldsymbol{A}_T^{-1} \right\|_F + b_T \right)} \right) \\ &+ l_T^2 \exp\left(-C_0 \frac{T}{\left\| \boldsymbol{A}_T^{-1} \right\|_F^2 l_T^2} \right). \end{aligned}$$

Result (A.17) now follows if we apply (A.16) to the first term on the RHS of the above. ■

Lemma 8 *Let $\boldsymbol{A}_T = (a_{ij,T})$ be a $l_T \times l_T$ matrix and $\hat{\boldsymbol{A}}_T = (\hat{a}_{ij,T})$ be an estimator of $\boldsymbol{A}_T$. Let $\left\|\boldsymbol{A}_T^{-1}\right\|_F > 0$ and suppose that for some $s > 0$, any $b_T > 0$ and some finite positive constant C_0,*

$$\sup_{i,j} \Pr\left(|\hat{a}_{ij,T} - a_{ij,T}| > b_T\right) \leq \exp\left[-C_0 \left(T b_T\right)^{s/(s+2)}\right].$$

Then

$$\Pr\left(\left\|\hat{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}_T^{-1}\right\|_F > b_T\right) \leq l_T^2 \exp\left(\frac{-C_0 \left(T b_T\right)^{s/(s+2)}}{l_T^{s/(s+2)} \left\|\boldsymbol{A}_T^{-1}\right\|_F^{s/(s+2)} \left(\left\|\boldsymbol{A}_T^{-1}\right\|_F + b_T\right)^{s/(s+2)}}\right) + l_T^2 \exp\left(-C_0 \frac{T^{s/(s+2)}}{\left\|\boldsymbol{A}_T^{-1}\right\|_F^{s/(s+2)} l_T^{s/(s+2)}}\right). \tag{A.21}$$

Proof:

First note that since $b_T > 0$, then

$$\begin{aligned}\Pr\left(\left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F > b_T\right) &= \Pr\left(\left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F^2 > b_T^2\right) \\ &= \Pr\left[\sum_{j=1}^{l_T}\sum_{i=1}^{l_T} \left(\hat{a}_{ij,T} - a_{ij,T}\right)^2 > b_T^2\right],\end{aligned}$$

and using the probability bound result, (A.6), and setting $\pi_i = 1/l_T^2$, we have

$$\begin{aligned}\Pr\left(\left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F > b_T\right) &\leq \sum_{j=1}^{l_T}\sum_{i=1}^{l_T} \Pr\left(|\hat{a}_{ij,T} - a_{ij,T}|^2 > l_T^{-2} b_T^2\right) \\ &= \sum_{j=1}^{l_T}\sum_{i=1}^{l_T} \Pr\left(|\hat{a}_{ij,T} - a_{ij,T}| > l_T^{-1} b_T\right) \\ &\leq l_T^2 \sup_{ij}\left[\Pr\left(|\hat{a}_{ij,T} - a_{ij,T}| > l_T^{-1} b_T\right)\right] = l_T^2 \exp\left(-C_0 T^{s/(s+1)} \frac{b_T^{s/(s+2)}}{l_t^{s/(s+2)}}\right).\end{aligned} \tag{A.22}$$

To establish (A.21) define the sets

$$\mathcal{A}_T = \left\{\left\|\boldsymbol{A}_T^{-1}\right\|_F \left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F \leq 1\right\} \text{ and } \mathcal{B}_T = \left\{\left\|\hat{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}_T^{-1}\right\| > b_T\right\}$$

and note that by (2.15) of Berk (1974) if $\mathcal{A}_T$ holds we have

$$\left\|\hat{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}_T^{-1}\right\| \leq \frac{\left\|\boldsymbol{A}_T^{-1}\right\|_F^2 \left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F}{1 - \left\|\boldsymbol{A}_T^{-1}\right\|_F \left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F}.$$

Hence

$$\begin{aligned}\Pr\left(\mathcal{B}_T \,|\mathcal{A}_T\right) &\leq \Pr\left(\frac{\left\|\boldsymbol{A}_T^{-1}\right\|_F^2 \left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F}{1 - \left\|\boldsymbol{A}_T^{-1}\right\|_F \left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F} > b_T\right) \\ &= \Pr\left[\left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F > \frac{b_T}{\left\|\boldsymbol{A}_T^{-1}\right\|_F \left(\left\|\boldsymbol{A}_T^{-1}\right\|_F + b_T\right)}\right].\end{aligned}$$

Note also that

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

Furthermore

$$\begin{aligned}\Pr\left(\mathcal{A}_T^C\right) &= \Pr\left(\left\|\boldsymbol{A}_T^{-1}\right\|_F \left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F > 1\right) \\ &= \Pr\left(\left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F > \left\|\boldsymbol{A}_T^{-1}\right\|_F^{-1}\right),\end{aligned}$$

and by (A.22) we have

$$\Pr\left(\mathcal{A}_T^C\right) \leq l_T^2 \exp\left(-C_0 T^{s/(s+1)} \frac{b_T^{s/(s+2)}}{l_t^{s/(s+2)}}\right) = \exp\left(-C_0 \frac{T^{s/(s+2)}}{\left\|\boldsymbol{A}_T^{-1}\right\|_F^{s/(s+2)} l_T^{s/(s+2)}}\right).$$

Using the above result, we now have

$$\begin{aligned}\Pr\left(\mathcal{B}_T\right) \leq \Pr&\left(\left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F > \frac{b_T}{\left\|\boldsymbol{A}_T^{-1}\right\|_F\left(\left\|\boldsymbol{A}_T^{-1}\right\|_F + b_T\right)}\right)\Pr\left(\mathcal{A}_T\right) \\ &+ \Pr\left(\mathcal{B}_T|\mathcal{A}_T^C\right)\exp\left(-C_0 \frac{T^{s/(s+2)}}{\left\|\boldsymbol{A}_T^{-1}\right\|_F^{s/(s+2)} l_T^{s/(s+2)}}\right).\end{aligned}$$

Furthermore, since $\Pr\left(\mathcal{A}_T\right) \leq 1$ and $\Pr\left(\mathcal{B}_T|\mathcal{A}_T^C\right) \leq 1$ then

$$\begin{aligned}\Pr\left(\mathcal{B}_T\right) = \Pr\left(\left\|\hat{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}_T^{-1}\right\| > b_T\right) \leq \Pr&\left(\left\|\hat{\boldsymbol{A}}_T - \boldsymbol{A}_T\right\|_F > \frac{b_T}{\left\|\boldsymbol{A}_T^{-1}\right\|_F\left(\left\|\boldsymbol{A}_T^{-1}\right\|_F + b_T\right)}\right) \\ &+ \exp\left(-C_0 \frac{T^{s/(s+2)}}{\left\|\boldsymbol{A}_T^{-1}\right\|_F^{s/(s+2)} l_T^{s/(s+2)}}\right).\end{aligned}$$

Result (A.21) now follows if we apply (A.22) to the first term on the RHS of the above.

Lemma 9 *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*

$$\Pr\left(\left|\sum_{t=1}^T z_t\right| > \zeta_T\right) \leq \exp\left[\frac{-\left(1-\pi\right)^2 \zeta_T^2}{2T\sigma_z^2}\right]. \tag{A.23}$$

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

$$\Pr\left(\left|\sum_{t=1}^T z_t\right| > \zeta_T\right) \leq \exp\left[-C_3 \zeta_T^{s/(s+1)}\right]. \tag{A.24}$$

Proof. We proceed to prove (A.23) first and then prove (A.24). 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

$$\Pr\left(\left|\sum_{t=1}^{T}[z_t - E(z_t)]\right| > \zeta_T\right) \leq \Pr\left(\left|\sum_{t=1}^{T}[w_t - E(w_t)]\right| > (1-\pi)\,\zeta_T\right) + \Pr\left(\left|\sum_{t=1}^{T}[v_t - E(v_t)]\right| > \pi\zeta_T\right), \tag{A.25}$$

for any $0 < \pi < 1$.[7] 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.25) 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.25) and first note that

$$\Pr\left(\left|\sum_{t=1}^{T}[v_t - E(v_t)]\right| > \pi\zeta_T\right) \leq \Pr\left[\sum_{t=1}^{T}|v_t - E(v_t)| > \pi\zeta_T\right], \tag{A.26}$$

and by Markov's inequality,

$$\begin{aligned}\Pr\left(\sum_{t=1}^{T}|[v_t - E(v_t)]| > \pi\zeta_T\right) &\leq \left(\frac{1}{\pi\zeta_T}\right)\sum_{t=1}^{T}E\,|v_t - E(v_t)| \\ &\leq \left(\frac{2}{\pi\zeta_T}\right)\sum_{t=1}^{T}E\,|v_t|\,. \end{aligned} \tag{A.27}$$

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

$$\begin{aligned} E\,|v_t| &= E\left(|z_t I\left[|z_t| > D_T\right]|\right) \\ &\leq \left(E\,|z_t|^p\right)^{1/p}\left\{E\left[|I\left(|z_t| > D_T\right)|^q\right]\right\}^{1/q} \\ &= \left(E\,|z_t|^p\right)^{1/p}\left\{E\left[I\left(|z_t| > D_T\right)\right]\right\}^{1/q} \\ &= \left(E\,|z_t|^p\right)^{1/p}\left[\Pr\left(|z_t| > D_T\right)\right]^{1/q}. \end{aligned} \tag{A.28}$$

Also, for any finite $p \geq 1$ there exists a finite positive constant C_2 such that $E\,|z_t|^p \leq C_2 < \infty$, by Lemma 6. Further, by assumption

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

[7] 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, $\Pr\left(|A_T| > \zeta_T\right) \leq \Pr\left(|B_{1,T}| + |B_{2,T}| > \zeta_T\right)$. Equation (A.25) now readily follows using the same steps as in the proof of (A.6).

Using this upper bound in (A.28) together with the upper bound on $E|z_t|^p$, we have

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

Therefore, using (A.26)-(A.27),

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

We need to determine D_T such that

$$(2/\pi)\, C_2^{1/p} C_0^{1/q} \zeta_T^{-1} T \left[\exp\left(-C_1 D_T^s\right)\right]^{1/q} \leq \exp\left[\frac{-(1-\pi)^2 \zeta_T^2}{2\left[T\sigma_z^2 + (1-\pi) D_T \zeta_T\right]}\right]. \tag{A.29}$$

Taking logs, we have

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

or

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

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

$$\begin{aligned}
&\left(2\sigma_z^2 C_1 q^{-1}\right) T D_T^s + \left(2 C_1 q^{-1}\right)(1-\pi) D_T^{s+1} \zeta_T - 2(1-\pi) D_T \zeta_T \ln\left(\zeta_T^{-1} T\right) - \\
&2(1-\pi) D_T \zeta_T \ln\left[(2/\pi)\, C_2^{1/p} C_0^{1/q}\right] \\
&\geq 2\sigma_z^2 T \ln\left[(2/\pi)\, C_2^{1/p} C_0^{1/q}\right] + 2\sigma_z^2 T \ln\left(\zeta_T^{-1} T\right) + (1-\pi)^2 \zeta_T^2.
\end{aligned} \tag{A.30}$$

The above expression can now be simplified for values of $T \to \infty$, by dropping the constants and terms that are asymptotically dominated by other terms on the same side of the inequality.[8] Since $\zeta_T = \ominus\left(T^{\lambda}\right)$, for some $0 < \lambda \leq (s+1)/(s+2)$, and considering values of D_T such that $D_T = \ominus\left(T^{\psi}\right)$, for some $\psi > 0$, implies that the third and fourth term on the LHS of (A.30), 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.30) which is of order $\ominus\left(T^{\lambda+\psi+s\psi}\right)$. Further the first term on the RHS of (A.30) 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.30) is $\ominus(T)$. Result (A.29) follows if we show that there exists D_T such that

$$\left(C_1 q^{-1}\right)\left[2\sigma_z^2 T D_T^s + 2(1-\pi) D_T^{s+1} \zeta_T\right] \geq 2\sigma_z^2 T \ln\left(\zeta_T^{-1} T\right) + (1-\pi)^2 \zeta_T^2. \tag{A.31}$$

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

Set

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

and note that (A.31) can be written as

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

Hence, the required condition is met if

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

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

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

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.29) is dominated by T which takes us back to the Exponential inequality with bounded random variables and proves (A.23). 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.29). So the bound takes the form $\exp\left[\frac{-(1-\pi)\zeta_T^2}{C_4 D_T \zeta_T}\right]$, 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.24). ■

Remark 19 *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 10 *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 $\sup_t \Pr\left(|x_t| > \alpha\right) \leq C_0 \exp\left(-C_1\alpha^s\right)$ and $\sup_t \Pr\left(|u_t| > \alpha\right) \leq C_0 \exp\left(-C_1\alpha^s\right)$, 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

$$\Pr\left(\left|\sum_{t=1}^{T}\left(x_t u_t - \mu_t\right)\right| > \zeta_T\right) \leq \exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2T\sigma_{(T)}^2}\right], \tag{A.32}$$

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

$$\Pr\left(\left|\sum_{t=1}^{T}\left(x_t u_t - \mu_t\right)\right| > \zeta_T\right) \leq \exp\left[-C_2\zeta_T^{s/(s+2)}\right]. \tag{A.33}$$

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 (A.7) with C_0 set equal to α and C_1 set equal to $\sqrt{\alpha}$)

$$\Pr\left[|x_t u_t| > \alpha\right] \leq \Pr\left[|x_t| > \alpha^{1/2}\right] + \Pr\left[|u_t|^2 > \alpha^{1/2}\right]. \tag{A.34}$$

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

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

Therefore, the process $x_t u_t$ satisfies the conditions of Lemma 9 and the results of the Lemma apply. ■

Lemma 11 *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 $\sup_t \Pr\left(|x_t| > \alpha\right) \leq C_0 \exp\left(-C_1\alpha^s\right)$ and $\sup_{i,t} \Pr\left(|q_{i,t}| > \alpha\right) \leq C_0 \exp\left(-C_1\alpha^s\right)$, for all $a > 0$. Consider the linear projection*

$$x_t = \sum_{j=1}^{l_T}\beta_j q_{jt} + u_{x,t}, \tag{A.35}$$

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

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

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 $0 < \pi < 1$,

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

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

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

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 12 *Let x_{it}, $i = 1, 2, ..., n$, $t = 1, ..., T$, and η_t be martingale difference processes that satisfy exponential tail probability bounds of the form (13) and (14), 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 $\boldsymbol{\Sigma}_{qq} = T^{-1}\sum_{t=1}^{T} E\left(\boldsymbol{q}_{\cdot t}\boldsymbol{q}_{\cdot t}'\right)$ 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 (15) 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 $\hat{\boldsymbol{u}}_{x_i} = (\hat{u}_{x_i,1}, \hat{u}_{x_i,2}, ..., \hat{u}_{x_i,T})' = \mathbf{M}_q\boldsymbol{x}_i$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})'$, $\hat{\boldsymbol{u}}_\eta = (\hat{u}_{\eta,1}, \hat{u}_{\eta,2}, ..., \hat{u}_{\eta,T})' = \mathbf{M}_q\boldsymbol{\eta}$, $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_T)'$, $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}\left(\mathbf{Q}'\mathbf{Q}\right)^{-1}\mathbf{Q}$, $\mathcal{F}_t = \mathcal{F}_t^\eta \cup \mathcal{F}_t^x$, $\mu_{x_i\eta,t} = E\left(u_{x_i,t}u_{\eta,t}\,|\mathcal{F}_{t-1}\right)$, $\omega^2_{x_i\eta,1,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[\left(x_{it}\eta_t - E\left(x_{it}\eta_t\,|\mathcal{F}_{t-1}\right)\right)^2\right]$, and $\omega^2_{x_i\eta,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[\left(u_{x_i,t}u_{\eta,t} - \mu_{x_i\eta,t}\right)^2\right]$. Let $\zeta_T = \ominus(T^\lambda)$. Then, for any π in the range $0 < \pi < 1$, we have,*

$$\Pr\left(\left|\sum_{t=1}^{T} x_{it}\eta_t - E\left(x_{it}\eta_t\,|\mathcal{F}_{t-1}\right)\right| > \zeta_T\right) \leq \exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2T\omega^2_{x_i\eta,1,T}}\right], \tag{A.36}$$

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

$$\Pr\left(\left|\sum_{t=1}^{T} x_{it}\eta_t - E\left(x_{it}\eta_t\,|\mathcal{F}_{t-1}\right)\right| > \zeta_T\right) \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right], \tag{A.37}$$

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

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}\hat{u}_{\eta,t} - \mu_{x_i\eta,t}\right)\right| > \zeta_T\right) \leq C_0 \exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2T\omega^2_{x_i\eta,T}}\right] + \exp\left[-C_1T^{C_2}\right]. \tag{A.38}$$

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

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}\hat{u}_{\eta,t}-\mu_{x_i\eta,t}\right)\right|>\zeta_T\right)\leq C_0\exp\left[-C_3\zeta_T^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right], \tag{A.39}$$

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.36) and (A.37) follow immediately given our assumptions and Lemma 10. 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

$$\begin{aligned}\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}\hat{u}_{\eta,t}-\mu_{x_i\eta,t}\right)&=\hat{\boldsymbol{u}}_{x_i}'\hat{\boldsymbol{u}}_\eta-\sum_{t=1}^{T}\mu_{x_i\eta,t}=\boldsymbol{u}_{x_i}'\mathbf{M}_q\boldsymbol{u}_\eta-\sum_{t=1}^{T}\mu_{x_i\eta,t}\\&=\sum_{t=1}^{T}\left(u_{x_i,t}u_{\eta,t}-\mu_{x_i\eta,t}\right)-\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\hat{\boldsymbol{\Sigma}}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_\eta\right),\end{aligned} \tag{A.40}$$

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

$$\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\hat{\boldsymbol{\Sigma}}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_\eta\right)=\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\left(\hat{\boldsymbol{\Sigma}}_{qq}^{-1}-\boldsymbol{\Sigma}_{qq}^{-1}\right)\left(\mathbf{Q}'\boldsymbol{u}_\eta\right)+\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\boldsymbol{\Sigma}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_\eta\right). \tag{A.41}$$

By (A.6) 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 (A.7) to the last two terms of the above we obtain

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

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

$$\begin{aligned}
&\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}_{x,t}\hat{u}_{\eta,t} - \mu_{x\eta,t}\right)\right| > \zeta_T\right) \\
&\leq \Pr\left(\left|\sum_{t=1}^{T}\left(u_{x,t}u_{\eta,t} - \mu_{x\eta,t}\right)\right| > \pi_1\zeta_T\right) + \Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_{qq}^{-1} - \boldsymbol{\Sigma}_{qq}^{-1}\right\|_F > \frac{\zeta_T}{\delta_T}\right) \\
&\quad + \Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F > \left(\pi_2\delta_T T\right)^{1/2}\right) + \Pr\left(\left\|\boldsymbol{u}_x'\mathbf{Q}\right\|_F > \left(\pi_2\delta_T T\right)^{1/2}\right), \\
&\quad + \Pr\left(\left\|\boldsymbol{u}_x'\mathbf{Q}\right\|_F > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) + \Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right). \qquad \text{(A.42)}
\end{aligned}$$

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 10, we have, for any π in the range $0 < \pi < 1$,

$$\Pr\left(\left|\sum_{t=1}^{T}\left(u_{x_i,t}u_{\eta,t} - \mu_{x_i\eta,t}\right)\right| > \pi_1\zeta_T\right) \leq \exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2T\omega_{x\eta,T}^2}\right], \qquad \text{(A.43)}$$

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

$$\Pr\left(\left|\sum_{t=1}^{T}\left(u_{x_i,t}u_{\eta,t} - \mu_{x_i\eta,t}\right)\right| > \pi_1\zeta_T\right) \leq \exp\left[-C_0\zeta_T^{s/(s+1)}\right], \qquad \text{(A.44)}$$

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.42) are of order $\exp\left[-C_1T^{C_2}\right]$, for some finite positive constants C_1 and C_2. We will make use of Lemma 10 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.32) or (A.33) of Lemma 10, apply. We start with (A.32). Then, for some finite positive constant C_0, we have[9]

$$\sup_i \Pr\left(\left\|\mathbf{q}_i'\boldsymbol{u}_{\eta}\right\| > \left(\pi_2\delta_T T\right)^{1/2}\right) \leq \exp\left(-C_0\delta_T\right). \qquad \text{(A.45)}$$

[9]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 11.

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 (A.6),

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

which upon using (A.45) yields (for some finite positive constant C_0)

$$\Pr\left(\|\mathbf{Q}'\boldsymbol{u}_\eta\|_F > (\pi_2\delta_T T)^{1/2}\right) \leq l_T\exp\left(-\frac{C_0\delta_T}{l_T}\right),\ \ \Pr\left(\|\mathbf{Q}'\boldsymbol{u}_x\| > (\pi_2\delta_T T)^{1/2}\right) \leq l_T\exp\left(-\frac{C_0\delta_T}{l_T}\right). \tag{A.46}$$

Similarly,

$$\begin{aligned}
\Pr\left(\|\mathbf{Q}'\boldsymbol{u}_\eta\|_F > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) &\leq l_T\exp\left(\frac{-C_0\zeta_T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right), \qquad (A.47)\\
\Pr\left(\|\mathbf{Q}'\boldsymbol{u}_x\| > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) &\leq l_T\exp\left(\frac{-C_0\zeta_T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right).
\end{aligned}$$

Turning to the second term of (A.42), 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

$$\sup_{ij}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left[q_{it}q_{jt} - E(q_{it}q_{jt})\right]\right| > \frac{\pi_2\zeta_T}{\delta_T}\right) \leq \exp(\frac{-C_0T\zeta_T^2}{\delta_T^2}). \tag{A.48}$$

Therefore, by Lemma 7, for some finite positive constant C_0, we have

$$\begin{aligned}
\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_{qq}^{-1} - \boldsymbol{\Sigma}_{qq}^{-1}\right\| > \frac{\zeta_T}{\delta_T}\right) \leq l_T^2\exp\left[\frac{-C_0T\zeta_T^2}{\delta_T^2 l_T^2\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^2\left(\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F + \delta_T^{-1}\zeta_T\right)^2}\right] + \qquad (A.49)\\
l_T^2\exp\left(\frac{-C_0T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^2 l_T^2}\right).
\end{aligned}$$

Further by Lemma 5, $\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F = \ominus\left(l_T^{1/2}\right)$, and

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

Consider now the different terms in the above expression and let

$$P_{11} = \frac{\delta_T}{l_T},\ P_{12} = \frac{\zeta_T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T},$$
$$P_{13} = \frac{T}{l_T^2 \left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^2 \left[\delta_T \zeta_T^{-1} \left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F + 1\right]^2}, \text{ and } P_{14} = \frac{T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^2 l_T^2}.$$

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

$$P_{11} = \frac{\delta_T}{l_T} = \ominus\left(T^{\alpha-d}\right), \tag{A.50}$$

$$P_{12} = \frac{\zeta_T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T} = \ominus\left(T^{\lambda-3d/2}\right), \tag{A.51}$$

$$\begin{aligned} P_{13} = \frac{T}{l_T^2 \left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^2 \left[\delta_T \zeta_T^{-1} \left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F + 1\right]^2} &= \ominus\left(T^{\max\{1-3d-(2\alpha-2\lambda+d),1-3d-(\alpha-\lambda+d/2),1-3d\}}\right) \\ &= \ominus\left(T^{\max\{1+2\lambda-4d-2\alpha,1+\lambda-7d/2-\alpha,1-3d\}}\right), \end{aligned} \tag{A.52}$$

and

$$P_{14} = \frac{T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^2 l_T^2} = \ominus\left(T^{1-3d}\right). \tag{A.53}$$

Suppose that $d < 1/3$, and by (A.51) 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 $\exp\left[-C_1 T^{C_2}\right]$, for some finite positive constants C_1 and C_2, and (A.38) follows. The same analysis can be repeated under (A.33). In this case, (A.46), (A.47), (A.48) and (A.49) are replaced by

$$\begin{aligned} \Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F > \left(\pi_2 \delta_T T\right)^{1/2}\right) &\leq l_T \exp\left(-\frac{C_0 \delta_T^{s/2(s+2)} T^{s/2(s+2)}}{l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\delta_T T}{l_T}\right)^{s/2(s+2)}\right], \\ \Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{x}\right\| > \left(\pi_2 \delta_T T\right)^{1/2}\right) &\leq l_T \exp\left(-\frac{C_0 \delta_T^{s/2(s+2)} T^{s/2(s+2)}}{l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\delta_T T}{l_T}\right)^{s/2(s+2)}\right], \end{aligned} \tag{A.54}$$

$$\begin{aligned} \Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F > \frac{\pi_3^{1/2} \zeta_T^{1/2} T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) &\leq l_T \exp\left(\frac{-C_0 \zeta_T^{s/2(s+2)} T^{s/2(s+2)}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{s/2(s+2)} l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\zeta_T T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right)^{s/2(s+2)}\right], \\ \Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{x}\right\| > \frac{\pi_3^{1/2} \zeta_T^{1/2} T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) &\leq l_T \exp\left(\frac{-C_0 \zeta_T^{s/2(s+2)} T^{s/2(s+2)}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{s/2(s+2)} l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\zeta_T T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right)^{s/2(s+2)}\right], \end{aligned} \tag{A.55}$$

$$\sup_{ij} \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left[q_{it}q_{jt} - E(q_{it}q_{jt})\right]\right| > \frac{\pi_2\zeta_T}{\delta_T}\right) \leq \exp\left[\frac{-C_0 T^{s/(s+2)}\zeta_T^{s/(s+2)}}{\delta_T^{s/(s+2)}}\right], \tag{A.56}$$

and, using Lemma 8,

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

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

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

$$l_T^2 \exp\left[-C_0\left(\frac{T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F l_T}\right)^{s/(s+2)}\right]. \tag{A.57}$$

respectively. Once again, we need to derive conditions that imply that $P_{21} = \frac{\delta_T T}{l_T}$, $P_{22} = \frac{\zeta_T T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F l_T}$, $P_{23} = \frac{T\zeta_T}{\delta_T l_T \left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F \left(\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F + \delta_T^{-1}\zeta_T\right)}$ and $P_{24} = \frac{T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F l_T}$ 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.39) 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

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

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

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.32), 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. ■

Remark 20 *It is important to highlight one particular feature of the above proof. In (A.46), $q_{it}u_{x,t}$ 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. Of course, we go on to relax this assumption in Section 4.7, where we allow noise variables to be mixing processes.

Lemma 13 *Let x_{it}, $i = 1, 2, ..., n$, be martingale difference processes that satisfy exponential tail probability bounds of the form (13), 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 (15) as $u_{x_i t} = x_{it} - \boldsymbol{\gamma}'_{qx_i,T}\boldsymbol{q}_{\cdot t}$. 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 $i = 1, 2, ..., l_T$. Let $\hat{\boldsymbol{u}}_{x_i} = (\hat{u}_{x_i,1}, \hat{u}_{x_i,2}, ..., \hat{u}_{x_i,T})' = \mathbf{M}_q\mathbf{x}_{i\cdot}$, where $\mathbf{x}_{i\cdot} = (x_{i1}, x_{i2}, ..., x_{iT})'$ and $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}$. Moreover, suppose that $E\left(u^2_{x_i,t} - \sigma^2_{x_i t}|\mathcal{F}_{t-1}\right) = 0$, where $\mathcal{F}_t = \mathcal{F}^x_t$ and $\sigma^2_{x_i t} = E(u^2_{x_i,t})$. 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,*

$$\Pr\left[\left|\sum_{t=1}^{T}\left(x^2_{it} - \sigma^2_{x_i t}\right)\right| > \zeta_T\right] \leq C_0 \exp\left[\frac{-(1-\pi)^2\zeta^2_T}{2T\omega^2_{i,1,T}}\right]. \tag{A.58}$$

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

$$\Pr\left[\left|\sum_{t=1}^{T}\left(x^2_{it} - \sigma^2_{x_i t}\right)\right| > \zeta_T\right] \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right]. \tag{A.59}$$

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

$$\Pr\left[\left|\sum_{t=1}^{T}\left(\hat{u}^2_{x_i,t} - \sigma^2_{x_i t}\right)\right| > \zeta_T\right] \leq C_0 \exp\left[\frac{-(1-\pi)^2\zeta^2_T}{2T\omega^2_{i,T}}\right] + \exp\left[-C_1 T^{C_2}\right], \tag{A.60}$$

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

$$\Pr\left[\left|\sum_{t=1}^{T}\left(\hat{u}^2_{x_i,t} - \sigma^2_{x_i t}\right)\right| > \zeta_T\right] \leq C_0 \exp\left[-C_3\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1 T^{C_2}\right], \tag{A.61}$$

for some finite positive constants C_0, C_1, C_2 and C_3, where $\omega^2_{i,1,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[\left(x^2_{it} - \sigma^2_{x_i t}\right)^2\right]$ and $\omega^2_{i,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[\left(u^2_{x_i,t} - \sigma^2_{x_i t}\right)^2\right]$.

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.58) and (A.59) follow similarly to (A.36) and (A.37). For (A.60) and (A.61), we first note that

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

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

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

by Lemma 11, then the conditions of Lemma 9 are met and we have

$$\Pr\left[\left|\sum_{t=1}^{T}\left(u_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq \exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2T\omega_{i,T}^2}\right]. \tag{A.62}$$

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

$$\Pr\left[\left|\sum_{t=1}^{T}\left(u_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right], \tag{A.63}$$

if $\lambda > (s/2+1)/(s/2+2)$. Then, using the same line of reasoning as in the proof of Lemma 12 we establish the desired result. ■

Lemma 14 *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 1-4, 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 (15) 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 $3d/2 < \lambda \leq (s/2+1)/(s/2+2)$,*

$$\Pr\left(\left|\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma_{x,(T)}^2} - 1\right| > a_T\right) \leq \exp\left[\frac{-\sigma_{x,(T)}^4\left(1-\pi\right)^2 Ta_T^2}{2\omega_{x,(T)}^2}\right] + \exp\left[-C_0T^{C_1}\right], \tag{A.64}$$

and

$$\Pr\left[\left|\left(\frac{\sigma_{x,(T)}^2}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2} - 1\right| > a_T\right] \leq \exp\left[\frac{-\sigma_{x,(T)}^4\left(1-\pi\right)^2 Ta_T^2}{2\omega_{x,(T)}^2}\right] + \exp\left[-C_0T^{C_1}\right], \tag{A.65}$$

where

$$\sigma_{x,(T)}^2 = \frac{1}{T}\sum_{t=1}^{T} E\left(u_{x,t}^2\right),\ \omega_{x,(T)}^2 = \frac{1}{T}\sum_{t=1}^{T} E\left[\left(u_{x,t}^2 - \sigma_{xt}^2\right)^2\right]. \tag{A.66}$$

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

$$\Pr\left(\left|\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma_{x,(T)}^2} - 1\right| > a_T\right) \leq \exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right] + \exp\left[-C_1T^{C_2}\right], \tag{A.67}$$

and

$$\Pr\left[\left|\left(\frac{\sigma^2_{x,(T)}}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right]. \tag{A.68}$$

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

$$\Pr\left(\left|\frac{T^{-1}\boldsymbol{e}'\boldsymbol{e}}{\sigma^2_{u,(T)}}-1\right|>a_T\right)\leq\exp\left[\frac{-\sigma^4_{u,(T)}\left(1-\pi\right)^2Ta_T^2}{2\omega^2_{u,(T)}}\right]+\exp\left[-C_0T^{C_1}\right], \tag{A.69}$$

and

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{\mathbf{e}'\mathbf{e}/T}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[\frac{-\sigma^4_{u,(T)}\left(1-\pi\right)^2Ta_T^2}{2\omega^2_{u,T}}\right]+\exp\left[-C_0T^{C_1}\right], \tag{A.70}$$

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

$$\sigma^2_{u,(T)}=\frac{1}{T}\sum_{t=1}^{T}\sigma_t^2,\ \textit{and}\ \omega^2_{u,T}=\frac{1}{T}\sum_{t=1}^{T}E\left[\left(u_t^2-\sigma_t^2\right)^2\right]. \tag{A.71}$$

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

$$\Pr\left(\left|\frac{T^{-1}\boldsymbol{e}'\boldsymbol{e}}{\sigma^2_{u,(T)}}-1\right|>a_T\right)\leq\exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right], \tag{A.72}$$

and

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{\mathbf{e}'\mathbf{e}/T}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right], \tag{A.73}$$

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 13 to $\sum_{t=1}^{T}\left(\hat{u}^2_{x,t}-\sigma^2_{xt}\right)$ with $\zeta_T=Ta_T$ we have

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}^2_{x,t}-\sigma^2_{xt}\right)\right|>\zeta_T\right)\leq\exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2\omega^2_{x,(T)}T}\right]+\exp\left[-C_0T^{C_1}\right],$$

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

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}^2_{x,t}-\sigma^2_{xt}\right)\right|>\zeta_T\right)\leq\exp\left[-C_0\zeta_T^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$, where $\omega^2_{x,(T)}$ is defined by (A.66). Also

$$\Pr\left[\left|\frac{T^{-1}\sum_{t=1}^{T}\left(\hat{u}^2_{x,t}-\sigma^2_{xt}\right)}{\sigma^2_{x,(T)}}\right|>\frac{\zeta_T}{T\sigma^2_{x,(T)}}\right]\leq\exp\left[\frac{-\left(1-\pi\right)^2\zeta^2_T}{2\omega^2_{x,(T)}T}\right]+\exp\left[-C_0T^{C_1}\right],$$

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

$$\Pr\left[\left|\frac{T^{-1}\sum_{t=1}^{T}\left(\hat{u}^2_{x,t}-\sigma^2_{xt}\right)}{\sigma^2_{x,(T)}}\right|>\frac{\zeta_T}{T\sigma^2_{x,(T)}}\right]\leq\exp\left[-C_0\zeta_T^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right],$$

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

$$\Pr\left(\left|\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}-1\right|>a_T\right)\leq\exp\left[\frac{-\sigma^4_{x,(T)}\left(1-\pi\right)^2Ta^2_T}{2\omega^2_{x,(T)}}\right]+\exp\left[-C_0T^{C_1}\right], \tag{A.74}$$

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

$$\Pr\left(\left|\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}-1\right|>a_T\right)\leq\exp\left[-C_0\zeta_T^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right], \tag{A.75}$$

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 4, we have

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

and hence

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[\frac{-\sigma^4_{x,(T)}\left(1-\pi\right)^2Ta^2_T}{\omega^2_{x,(T)}}\right]+\exp\left[-C_0T^{C_1}\right], \tag{A.76}$$

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

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[-C_0\zeta_T^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right], \tag{A.77}$$

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 2 for some finite positive constant C, we have

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

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.74),

$$\Pr\left[\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2}-1\right|>a_T\right]\leq \exp\left[\frac{-\sigma^4_{x,(T)}\left(1-\pi\right)^2Ta_T^2}{2\omega^2_{x,(T)}}\right]+\exp\left[-C_0T^{C_1}\right], \quad \text{(A.78)}$$

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

$$\Pr\left[\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2}-1\right|>a_T\right]\leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right], \quad \text{(A.79)}$$

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 13 to $\sum_{t=1}^T\left(e_t^2-\sigma_t^2\right)$, and following similar lines of reasoning we have

$$\Pr\left[\left|\sum_{t=1}^T\left(e_t^2-\sigma_t^2\right)\right|>\zeta_T\right]\leq\exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2\omega^2_{u,(T)}T}\right]+\exp\left[-C_0T^{C_1}\right],$$

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

$$\Pr\left[\left|\sum_{t=1}^T\left(e_t^2-\sigma_t^2\right)\right|>\zeta_T\right]\leq\exp\left[-C_0\zeta_T^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right],$$

if $\lambda>(s/2+1)/(s/2+2)$, which yield (A.69) and (A.72). Result (A.70) also follows along similar lines as used above to prove (A.65). ■

Lemma 15 *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 1-4. Let $\boldsymbol{q}_{.t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ 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 $\boldsymbol{\Sigma}_{qq} = \frac{1}{T}\sum_{t=1}^{T} E(\boldsymbol{q}_{.t}\boldsymbol{q}_{.t}')$ and $\hat{\boldsymbol{\Sigma}}_{qq} = \mathbf{Q}'\mathbf{Q}/T$ are both invertible, where $\mathbf{Q} = (\boldsymbol{q}_{1.}, \boldsymbol{q}_{2.}, ..., \boldsymbol{q}_{l_T.})$ and $\boldsymbol{q}_{i.} = (q_{i1}, q_{i2}, .., q_{iT})'$, for $i = 1, 2, ..., l_T$. Suppose that Assumption 5 holds for the pair y_t and $(\boldsymbol{q}_{.t}', x_t)'$, where x_t is a generic element of $\{x_{1t}, x_{2t}, ..., x_{nt}\}$ that does not belong to $\boldsymbol{q}_{.t}$, and denote the corresponding projection residuals defined by (15) as $e_t = y_t - \boldsymbol{\gamma}_{yqx,T}'(\boldsymbol{q}_{.t}', x_t)'$. Define $\boldsymbol{x} = (x_1, x_2, ..., x_T)'$, $\boldsymbol{e} = (e_1, e_2, ..., e_T)'$, 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*

$$\Pr\left[\left|\frac{a_T}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}\right| > c_p(n)\right] \leq \Pr\left(\left|\frac{a_T}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{c_p(n)}{1+d_T}\right) + \exp\left[-C_0 T^{C_1}\right] \tag{A.80}$$

for any random variable a_T, some finite positive constants C_0 and C_1, and some bounded sequence $d_T > 0$, where $c_p(n)$ is defined in (7). Similarly,

$$\Pr\left[\left|\frac{a_T}{\sqrt{(\mathbf{e}'\mathbf{e}/T)}}\right| > c_p(n)\right] \leq \Pr\left(\left|\frac{a_T}{\sigma_{e,(T)}}\right| > \frac{c_p(n)}{1+d_T}\right) + \exp\left[-C_0 T^{C_1}\right]. \tag{A.81}$$

Proof. We prove (A.80). (A.81) 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 2, note that for any $d_T > 0$ bounded in T,

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

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

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

Using (A.65), (A.68), (A.70) and (A.73),

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

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,

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

which establishes the desired the result. ■

Lemma 16 *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 1-4, 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 (15) 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 $\theta = E\left(T^{-1}\boldsymbol{x}'\mathbf{M}_q\mathbf{X}_b\right)\boldsymbol{\beta}_b$, where $\mathbf{X}_b$ is $T \times k_b$ matrix of observations on $\boldsymbol{x}_{b,t}$. Finally, $c_p(n)$ is given by (7) and (8), for any positive finite δ and $0 < p < 1$, and there exists $\kappa > 0$ such that $n = O\left(T^{\kappa}\right)$. Then, for any π in the range $0 < \pi < 1$, $d_T > 0$ and bounded in T, and for some finite positive constants C_0 and C_1,*

$$\Pr\left[|t_x| > c_p(n)\,|\theta = 0\right] \leq \exp\left[\frac{-(1-\pi)^2\sigma^2_{e,(T)}\sigma^2_{x,(T)}c_p^2(n)}{2(1+d_T)^2\omega^2_{xe,T}}\right] + \exp\left[-C_0 T^{C_1}\right], \tag{A.82}$$

where

$$t_x = \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}, \tag{A.83}$$

$$\sigma^2_{e,(T)} = E\left(T^{-1}\mathbf{e}'\mathbf{e}\right), \ \sigma^2_{x,(T)} = E\left(T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}\right), \tag{A.84}$$

and

$$\omega_{xe,T}^2 = \frac{1}{T}\sum_{t=1}^{T} E\left[(u_{x,t}\eta_t)^2\right]. \tag{A.85}$$

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

$$\begin{aligned}\Pr\left[|t_x| > c_p(n)\,|\theta = 0\right] \leq \exp\left[\frac{-(1-\pi)^2 c_p^2(n)}{2\left(1+d_T\right)^2}\right] \\ + \exp\left(-C_0 T^{C_1}\right).\end{aligned} \tag{A.86}$$

In the case where $\theta \neq 0$, *for* $d_T > 0$ *and bounded in* T, *and for some* $C_2, C_3 > 0$, *we have*

$$\Pr\left[|t_x| > c_p(n)\,|\theta \neq 0\right] > 1 - \exp\left(-C_2 T^{C_3}\right). \tag{A.87}$$

Proof. The DGP, given by (17), 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,

$$t_x = \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}} = \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{X}_b\boldsymbol{\beta}_b}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}} + \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{u}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}. \tag{A.88}$$

Let $\theta = E\left(T^{-1}\boldsymbol{x}'\mathbf{M}_q\mathbf{X}_b\right)\boldsymbol{\beta}_b$, $\boldsymbol{\eta} = \mathbf{X}_b\boldsymbol{\beta}_b + \mathbf{u}$, $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_T)'$, and write (A.88) as

$$t_x = \frac{\sqrt{T}\theta}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}} + \frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta\right)}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}. \tag{A.89}$$

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

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

Now by Lemma 15, we have

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

where $\sigma^2_{e,(T)}$ and $\sigma^2_{x,(T)}$ are defined by (A.84). Hence, noting that, by Remark 2, $c_p(n) = o(T^{C_0})$, for all $C_0 > 0$, under Assumption 4, and by Lemma 12, we have

$$\Pr\left[|t_x| > c_p(n)\,|\theta = 0\right] \leq \exp\left[\frac{-(1-\pi)^2\,\sigma^2_{e,(T)}\sigma^2_{x,(T)}c_p^2(n)}{2\left(1+d_T\right)^2\omega^2_{xe,T}}\right] + \exp\left(-C_0T^{C_1}\right),$$

where

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

and $u_{x,t}$, being the error in the regression of x_t on $\mathbf{Q}$, is defined by (15). 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

$$\omega^2_{xe,T} = \frac{1}{T}\sum_{t=1}^{T} E\left(u^2_{x,t}\right)\boldsymbol{\beta}'_b E\left(\boldsymbol{x}_{bq,t}\boldsymbol{x}'_{bq,t}\right)\boldsymbol{\beta}_b + \frac{1}{T}\sum_{t=1}^{T} E\left(u^2_{xt}\right)E\left(u^2_t\right). \tag{A.90}$$

Similarly

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

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

$$\begin{aligned}\sigma^2_{e,(T)} &= \boldsymbol{\beta}'_b E\left(T^{-1}\mathbf{X}'_b\mathbf{M}_q\mathbf{X}_b\right)\boldsymbol{\beta}_b + \frac{1}{T}\sum_{t=1}^{T} E\left(u^2_t\right) \\ &= \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\beta}'_b E\left(\boldsymbol{x}_{bq,t}\boldsymbol{x}'_{bq,t}\right)\boldsymbol{\beta}_b + \frac{1}{T}\sum_{t=1}^{T} E\left(u^2_t\right).\end{aligned} \tag{A.91}$$

Using (A.90) and (A.91), it is now easily seen that if either $E\left(u^2_{x,t}\right) = \sigma^2_{ux}$ or $E\left(u^2_t\right) = \sigma^2$, for all t, then we have $\omega^2_{xe,T} = \sigma^2_{e,(T)}\sigma^2_{x,(T)}$, and hence

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

giving a rate that does not depend on error variances. Next, we consider $\theta \neq 0$. By (A.80) of Lemma 15, for $d_T > 0$ and bounded in T,

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

We then have

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

Then

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

Note that since $c_p(n)$ is given by (7) and (8), then by Remark 2, $\frac{T^{1/2}|\theta|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p(n)}{1+d_T} > 0$. Then by Lemma 3,

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

But, setting $\zeta_T = T^{1/2}\left[\frac{T^{1/2}|\theta|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p(n)}{1+d_T}\right]$ and noting this choice of ζ_T satisfies $\zeta_T = \ominus\left(T^{\lambda}\right)$ with $\lambda = 1$, (A.39) of Lemma 12 applies regardless of $s > 0$, which gives us

$$\begin{aligned}&\Pr\left[\left|\frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta\right)}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{T^{1/2}|\theta|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p(n)}{1+d_T}\right] \\ &\leq C_4\exp\left\{-C_5\left[T^{1/2}\left(\frac{T^{1/2}|\theta|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{\theta c_p(n)}{1+d_T}\right)\right]^{s/(s+2)}\right\} \\ &+ \exp\left(-C_6T^{C_7}\right),\end{aligned}$$

for some C_4, C_5, C_6 and $C_7 > 0$. Hence, there must exist positive finite constants C_2 and C_3, such that

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

for any $s > 0$. So overall

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

■

Lemma 17 *Consider the data generating process (1) with k signal variables $\{x_{1t}, x_{2t}, ..., x_{kt}\}$, and assume that there are k^* pseudo-signal variables $\{x_{k+1,t}x_{k+2,t}, ..., x_{k+k^*,t}\}$ and $n-k-k^*$ noise variables $\{x_{k+k^*+1,t}, x_{k+k^*+2,t}, ..., x_{n,t}\}$. Moreover, suppose that conditions of Lemma 16 hold. Then for $0 < \varkappa < 1$ and the constants C_0, $C_1, C_2, C_3 > 0$, we have*

$$\Pr\left(\hat{k}-k-k^* > j\right) \leq \frac{k+k^*}{j}O\left[\exp\left(-C_2T^{C_3}\right)\right] + \frac{n-k-k^*}{j}\left\{\exp\left[-\frac{\varkappa c_p^2(n)}{2}\right] + O\left[\exp(-C_0T^{C_1})\right]\right\}, \tag{A.92}$$

for $j = 1, ..., n-k-k^*$, where $\hat{k}$ is the number of variables selected by the OCMT procedure, defined by

$$\hat{k} = \sum_{i=1}^{n} I\widehat{(\beta_i \neq 0)},$$

and $I\widehat{(\beta_i \neq 0)}$ is defined by (6).

Proof. We first note that by Markov's inequality

$$\Pr\left(\hat{k}-k-k^* > j\right) \leq \frac{E\left(\hat{k}-k-k^*\right)}{j}. \tag{A.93}$$

But

$$\begin{aligned}
E\left(\hat{k}\right) &= \sum_{i=1}^{n} E\left[I\widehat{(\beta_i \neq 0)}\right] \\
&= \sum_{i=1}^{k+k^*} E\left[I\widehat{(\beta_i \neq 0)}\left|\beta_i \neq 0\right.\right] + \sum_{i=k+k^*+1}^{n} E\left[I\widehat{(\beta_i \neq 0)}\left|\beta_i = 0\right.\right] \\
&= \sum_{i=1}^{k+k^*} \Pr\left(\left|t_{\hat{\phi}_i}\right| > c_p(n)|\theta_i \neq 0\right) + \sum_{i=k+k^*+1}^{n} \Pr\left(\left|t_{\hat{\phi}_i}\right| > c_p(n)|\theta_i = 0\right).
\end{aligned}$$

Now using (A.86) of Lemma 16, we have (for some $0 < \varkappa < 1$ and $C_0, C_1 > 0$)

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

Also, we have, using (A.87) of Lemma 16,

$$1 - \Pr\left(\left|t_{\hat{\phi}_i}\right| > c_p(n)|\theta_i \neq 0\right) < \exp(-C_2 T^{C_3}),$$

and $i = 1, 2, ..., k + k^*$. Hence,

$$E\left(\hat{k}\right) - k - k^* = (k + k^*)\, O\left[\exp\left(-C_2 T^{C_3}\right)\right] + (n - k - k^*)\left\{\exp\left[-\frac{\varkappa c_p^2(n)}{2}\right] + \exp(-C_0 T^{C_1})\right\}.$$

Using this result in (A.93) now establishes (A.92). ■

Lemma 18 *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 $\hat{\mathbf{\Sigma}}_{ab} = T^{-1}\mathbf{S}_a'\mathbf{S}_b$ and denote its expectations by $\mathbf{\Sigma}_{ab} = T^{-1}E\left(\mathbf{S}_a'\mathbf{S}_b\right)$. Let*

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

and suppose that

$$\sup_{i,j}\left[\sum_{t=1}^{T}\sum_{t'=1}^{T} E(z_{ij,t}z_{ij,t'})\right] = O\left(T\right). \tag{A.94}$$

Then,

$$E\left\|\hat{\mathbf{\Sigma}}_{ab} - \mathbf{\Sigma}_{ab}\right\|_F^2 = O\left(\frac{l_{a,T}l_{b,T}}{T}\right). \tag{A.95}$$

If, in addition,

$$\sup_{i,j,i',j'}\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} E(z_{ij,t}z_{ij,t'}z_{i'j',s}z_{i'j',s'})\right] = O\left(T^2\right), \tag{A.96}$$

then

$$E\left\|\hat{\mathbf{\Sigma}}_{ab} - \mathbf{\Sigma}_{ab}\right\|_F^4 = O\left(\frac{l_{a,T}^2 l_{b,T}^2}{T^2}\right). \tag{A.97}$$

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

$$a_{ij,T} = T^{-1}\sum_{t=1}^{T} z_{ij,t}, \tag{A.98}$$

and hence

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

and (A.95) follows from (A.94). Similarly,

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

But using (A.98) we have

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

and

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

Result (A.97) now follows from (A.96). ■

Remark 21 *It is clear that conditions (A.94) and (A.96) are met under Assumption 4 that requires z_{it} to be a martingale difference process. But it is easily seen that condition (A.94) 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 19 *Suppose that the data generating process (DGP) is given by*

$$\underset{T\times1}{\mathbf{y}}=\underset{T\times k+1}{\mathbf{X}}\cdot\underset{k+1\times1}{\boldsymbol{\beta}}+\underset{T\times1}{\mathbf{u}},\tag{A.99}$$

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

$$\underset{T\times 1}{\mathbf{y}} = \underset{T\times l_T}{\mathbf{S}} \cdot \underset{l_T\times 1}{\boldsymbol{\delta}} + \underset{T\times 1}{\varepsilon}. \tag{A.100}$$

where $\mathbf{u} = (u_1, u_2, ..., u_T)'$ *is independently distributed of* $\mathbf{X}$ *and* $\mathbf{S}$*,* $E(\mathbf{u}) = \mathbf{0}$*,* $E(\mathbf{uu}') = \sigma^2\mathbf{I}$*,* $0 < \sigma^2 < \infty$*,* $\mathbf{I}$ *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* $\mu_1 \leq \mu_2 \leq ... \leq \mu_{l_T}$*. Let* $\mu_i = O(l_T)$*,* $i = l_T - M + 1, ..., l_T$*, 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* $C_1 > 0$*.*

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* $\hat{\boldsymbol{\Sigma}}_{ss} = \mathbf{S}'\mathbf{S}/T$*.*

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.94) and (A.96) of Lemma 18. Moreover,* $z^*_{ij,t} = s_{it}x_{jt} - E(s_{it}x_{jt})$ *satisfies condition (A.94) of Lemma 18, and* $\|\boldsymbol{\Sigma}_{sx}\|_F = \|E(\mathbf{S}'\mathbf{X}/T)\|_F = O(1)$*.*

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

$$E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\beta}_0\right\| = O\left(\frac{l_T^4}{T}\right), \tag{A.101}$$

where $\hat{\boldsymbol{\delta}}$ *is the least square estimator of* $\boldsymbol{\delta}$ *in the regression model (A.100) 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*

$$E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\beta}_0\right\| = O(l_T) + O\left(\frac{l_T^4}{T}\right). \tag{A.102}$$

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

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

Hence

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

Using (2.15) of Berk (1974),

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

and using Cauchy-Schwarz inequality,

$$E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \leq \left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F^2 \left[E\left(\left\|\hat{\mathbf{\Sigma}}_{ss} - \mathbf{\Sigma}_{ss}\right\|_F^2\right)\right]^{1/2} \cdot \left\{E\left[\frac{1}{\left(1 - \left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F \left\|\hat{\mathbf{\Sigma}}_{ss} - \mathbf{\Sigma}_{ss}\right\|_F\right)^2}\right]\right\}^{1/2}. \tag{A.103}$$

We focus on the individual terms on the right side of (A.103) to establish an upper bound for $E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F$. The assumptions on eigenvalues of $\mathbf{\Sigma}_{ss}$ in this lemma are the same as in Lemma 5 with the only exception that $O(.)$ terms are used intstead of $\ominus(.)$. Using the same arguments as in the proof of Lemma 5, it readily follows that

$$\left\|\mathbf{\Sigma}_{ss}\right\|_F = O\left(l_T\right),$$

and

$$\left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F = O\left(\sqrt{l_T}\right). \tag{A.104}$$

Moreover, note that (i,j)-th element of $\left(\hat{\mathbf{\Sigma}}_{ss} - \mathbf{\Sigma}_{ss}\right)$, $z_{ijt} = s_{it}s_{jt} - E\left(s_{it}s_{jt}\right)$, satisfies the conditions of Lemma 18, which establishes

$$E\left(\left\|\hat{\mathbf{\Sigma}}_{ss} - \mathbf{\Sigma}_{ss}\right\|_F^2\right) = O\left(\frac{l_T^2}{T}\right). \tag{A.105}$$

Noting that $E\left(a^2\right) \leq \sqrt{E\left(a^4\right)}$, Assumption (ii) of this lemma implies that the last term on the right side of (A.103) is bounded, namely

$$E\left[\frac{1}{\left(1 - \left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F \left\|\hat{\mathbf{\Sigma}}_{ss} - \mathbf{\Sigma}_{ss}\right\|_F\right)^2}\right] = O\left(1\right), \tag{A.106}$$

Using (A.104), (A.105), and (A.106) in (A.103),

$$E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F = O\left(l_T\right)\sqrt{O\left(\frac{l_T^2}{T}\right)}O\left(1\right) = O\left(\frac{l_T^2}{\sqrt{T}}\right). \tag{A.107}$$

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

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

and using Cauchy-Schwarz inequality yields

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

where $\left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F^4 = O\left(l_T^2\right)$ by (A.104), $E\left(\left\|\hat{\mathbf{\Sigma}}_{ss} - \mathbf{\Sigma}_{ss}\right\|_F^4\right) = O\left(l_T^4/T^2\right)$ by (A.97) of Lemma 18, and $E\left[\left(1 - \left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F \left\|\hat{\mathbf{\Sigma}}_{ss} - \mathbf{\Sigma}_{ss}\right\|_F\right)^{-4}\right] = O\left(1\right)$ by Assumption ii of this lemma. Hence,

$$E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F^2 = O\left(l_T^2\right)\sqrt{O\left(\frac{l_T^4}{T^2}\right)}O\left(1\right) = O\left(\frac{l_T^4}{T}\right). \tag{A.108}$$

Using Lemma 18 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.95),

$$E\left(\left\|\hat{\mathbf{\Sigma}}_{sx} - \mathbf{\Sigma}_{sx}\right\|_F^2\right) = O\left(\frac{l_T}{T}\right). \tag{A.109}$$

We use the above results to derive an upper bound for

$$\begin{aligned} E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_*\right\| \leq\ & E\left[\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F\right]\|\boldsymbol{\beta}\| \\ &+ E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \left\|\mathbf{\Sigma}_{sx}\right\|_F \|\boldsymbol{\beta}\| \\ &+ \left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F \|\boldsymbol{\beta}\| \\ &+ E\left\|\hat{\mathbf{\Sigma}}_{ss}^{-1}\left(\frac{\mathbf{S}'\mathbf{u}}{T}\right)\right\|_F. \end{aligned} \tag{A.110}$$

First, note that $\|\boldsymbol{\beta}\| = O\left(1\right)$, 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.108), and $E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F^2 = O\left(l_T/T\right)$ by (A.109), and therefore

$$E\left[\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F\right]\|\boldsymbol{\beta}\| = \left[O\left(\frac{l_T^4}{T}\right)\right]^{1/2}\left[O\left(\frac{l_T}{T}\right)\right]^{1/2} = O\left(\frac{l_T^{5/2}}{T}\right). \tag{A.111}$$

Next, note that $E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F = O\left(l_T^2/\sqrt{T}\right)$ by (A.108), $\|\mathbf{\Sigma}_{sx}\|_F = O(1)$ by Assumption *iii* of this lemma (and $\|\boldsymbol{\beta}\| = O(1)$), and we obtain

$$E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F\|\mathbf{\Sigma}_{sx}\|_F\|\boldsymbol{\beta}\| = O\left(\frac{l_T^2}{\sqrt{T}}\right). \tag{A.112}$$

Moreover, using (A.104), and noting that $E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F = O\left(\sqrt{l_T/T}\right)$by (A.109),[10]

$$\left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F = O\left(\sqrt{l_T}\right)O\left(\frac{\sqrt{l_T}}{\sqrt{T}}\right) = O\left(\frac{l_T}{\sqrt{T}}\right),$$

and hence

$$\left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F\|\boldsymbol{\beta}\| = O\left(\frac{l_T}{\sqrt{T}}\right). \tag{A.113}$$

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 $E\left(\mathbf{u}\mathbf{u}'/T\right) = \sigma^2\mathbf{I}$, 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.107), we have

$$E\left|Tr\left(\hat{\mathbf{\Sigma}}_{ss}^{-1} - \mathbf{\Sigma}_{ss}^{-1}\right)\right| \leq l_T E\left\|\hat{\mathbf{\Sigma}}_{ss}^{-1} - \mathbf{\Sigma}_{ss}^{-1}\right\|_F = l_T E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F = O\left(\frac{l_T^3}{\sqrt{T}}\right).$$

[10] $E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F \leq \left[E\left(\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F^2\right)\right]^{1/2} = \sqrt{O\left(K_T/T\right)} = O\left(\sqrt{K_T/T}\right)$.

It follows,

$$E\left\|\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{u}\right\|_F^2 = O\left(\frac{l_T}{T}\right) + O\left(\frac{l_T^2}{\sqrt{T}}\right)\frac{1}{T}$$
$$= O\left(\frac{l_T}{T}\right) + O\left(\frac{l_T^3}{T^{3/2}}\right). \tag{A.114}$$

Overall, using (A.111), (A.112), (A.113), and (A.114) in (A.110),

$$E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_*\right\| = O\left(\frac{l_T^{5/2}}{T}\right) + O\left(\frac{l_T^2}{\sqrt{T}}\right) + O\left(\frac{l_T}{\sqrt{T}}\right)$$
$$+ O\left(\frac{l_T}{T}\right) + O\left(\frac{l_T^3}{T^{3/2}}\right).$$

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.101) 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

$$\left\|\boldsymbol{\delta}_*\right\| = \left\|\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{sx}\right\|_F$$
$$\leq \left\|\boldsymbol{\Sigma}_{ss}^{-1}\right\|_F \left\|\boldsymbol{\Sigma}_{sx}\right\|_F.$$

$\left\|\boldsymbol{\Sigma}_{ss}^{-1}\right\|_F = O\left(\sqrt{l_T}\right)$ by (A.104), and $\left\|\boldsymbol{\Sigma}_{sx}\right\|_F = O(1)$ by Assumption *iii* of this lemma. Hence, when at least one of the columns of $\mathbf{X}$ does not belong to $\mathbf{S}$,

$$\left\|\boldsymbol{\delta}_* - \boldsymbol{\beta}_0\right\| = O\left(l_T\right),$$

which completes the proof of (A.102). ■

Lemma 20 *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 $\mathbf{X}_k = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$. 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 $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$ exists. It is also assumed 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 $\boldsymbol{\beta}_b$. Define*

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

where $\boldsymbol{\Omega}_{b,T} = E(T^{-1}\mathbf{X}_b'\mathbf{M}_q\mathbf{X}_b)$. If $\boldsymbol{\Omega}_{b,T}$ is nonsingular, and $\boldsymbol{\beta}_k = (\beta_1, \beta_2, ..., \beta_k)' \neq \mathbf{0}$, 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 $\boldsymbol{\beta}_b \neq 0$, it follows that $\boldsymbol{\theta}_{b,T} \neq \mathbf{0}$, as desired. ■

References

Antoniadis, A., and J. Fan (2001): "Regularization of Wavelets Approximations," *Journal of the American Statistical Association*, 96, 939–967.

Bailey, N., M. H. Pesaran, and L. V. Smith (2015): "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.

Chudik, A., and M. H. Pesaran (2013): "Econometric Analysis of High Dimensional VARs Featuring a Dominant Unit," *Econometric Reviews*, 32, 592–649.

Dendramis, Y., L. Giraitis, and G. Kapetanios (2015): "Estimation of random coefficient time varying covariance matrices for large datasets," *Mimeo*.

Dominici, D. E. (2003): "The Inverse of the Cumulative Standard Normal Probability Function," *Integral Transforms and Special Functions*, 14(3), 281–292.

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.

Doornik, J. A., D. F. Hendry, and F. Pretis (2013): "Step Indicator Saturation," *Working paper, Oxford Martin School, Oxford University.*

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004): "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

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.

Frank, I. E., and J. H. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–148.

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.

Hendry, D. F., S. Johansen, and C. Santos (2008): "Automatic selection of indicators in a fully saturated regression," *Computational Statistics*, 33, 317–335.

Hendry, D. F., and H. M. Krolzig (2005): "The Properties of Automatic Gets Modelling," *Economic Journal*, 115, C32–C61.

Hocking, R. R. (1976): "The Analysis And Selection Of Variables In Linear Regression," *Biometrics*, 32(1).

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.

Said, E., and D. A. Dickey (1984): "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order," *Biometrika*, 71, 599–607.

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.