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

A.Chudik (a) G. Kapetanios (b) M. Hashem Pesaran (c)

(a) Federal Reserve Bank of Dallas
(b) King's College London, University of London
(c) Department of Economics & USC Dornsife INET, University of Southern California, USA, and Trinity College, Cambridge, UK

Introduction

- The problem of correctly specifying a model has been a recurring theme in econometrics.
- There are a number of competing approaches such as those based on specification testing or the use of information criteria
- Up until recently, the problem of variable or model selection has been analysed using standard statistical framework where the number of observations is considerably larger than the number of potential variables.
- In **high-dimensional regression** settings, where the number of variables (models) is larger than the number of available observations, model selection and estimation have been largely approached using a set of methods collectively known as **penalised (or regularised) regression.**

- Other approaches (often referred as machine learning techniques) such boosting, regression trees, and step-wise regressions are also used, but they lack a rigorous theory and the stopping criteria used in these algorithms are often ad hoc.
- In this paper we propose an alternative approach to the problem of variable selection in high-dimensional linear regression models.

Introduction: The Variable Selection Problem

- Suppose that an investigator is faced with the problem of explaining the target variable, denoted as y_t, in terms of n potential covariates, $\mathbf{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$. It is assumed that the model generating y_t is spanned by a sparse sub-set of $\mathbf{x}_{nt}$.
- Both n and T are large (and possibly $n > T$).
- Accordingly, the n potential covariates are classified in three categories:
  - k **signals**, that together generate y_t (k is fixed as $n \to \infty$)
  - k^* **pseudo-signals**, which are not included in the model generating y_t (the DGP for short), but are correlated with signals.
  - $n - k - k^*$ remaining **noise** covariates which are uncorrelated with the k signal variables.
- The identity of the signals is not known.

Penalised regressions

- Penalised regressions can be applied to linear as well as non-linear regression models. In the linear case

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

penalised regressions are used when n is large relative to T.

- The potential covariates, $\{x_{it},$ for $i = 1, 2, .., n\}$, are typically
    - standardised and in some cases also orthogonalised,
    - assumed to be strictly exogenous - in some papers endogeneity is allowed but it is assumed there exist suitable instruments.

Examples of penalty functions used in the literature:

$$\text{Ridge regression: } \sum_{i=1}^{n} \beta_i^2 < K < \infty,$$

$$\text{Lasso regression: } \sum_{i=1}^{n} |\beta_i| < K < \infty,$$

$$\text{Non-convex penalised regression: } \sum_{i=1}^{n} |\beta_i|^{\gamma} < K < \infty, \quad 0 < \gamma < 1$$

or

$$\text{Elastic net regression: } \sum_{i=1}^{n} \left[(1-\alpha) |\beta_i| + \alpha \beta_i^2 \right] < K < \infty.$$

- The penalised regressions are then computed by solving the optimisation problem $[\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)']$

$$\min_{\beta} \left\{ \sum_{t=1}^{T} \left(y_t - a - \boldsymbol{\beta}_n' \mathbf{x}_{nt} \right)^2 + \lambda \sum_{i=1}^{n} \left[(1-\alpha)|\beta_i|^{\gamma} + \alpha \beta_i^2 \right] \right\},$$

$\mathbf{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$ for given values of λ, α and γ.

- OLS corresponds to the no penalty case of $\lambda = 0$ and when $\lambda \neq 0$, $\alpha = 1$ yields the Ridge regressions and $\alpha = 0$, $\gamma = 1$ the Lasso regression, with the latter being better suited for a predictor selection as originally noted by Tibshirani (1996, JRSS).
- λ, α and γ are estimated using cross-validation techniques.

This Paper: A New Method

- In this paper we propose **O**ne **C**ovariate at a time **M**ultiple **T**esting (**OCMT**) procedure, where covariates are selected **one at a time**, based on a t-test.
- The general case requires an iterative stage to account for the statistical contribution of covariates that have not been previously selected (again one at a time) to the unexplained part of y_t.
- The selected covariates are then used, in a final multiple regression stage, to provide the final coefficient estimates.
- In carrying out the t-tests we adjust the critical values to take account of the multiple testing nature of the procedure.

- **OCMT** is conceptually and computationally simple, and stands at the other extreme to the penalised regression technique that considers all the covariates simultaneously.
- Also unlike penalised regression and boosting, **OCMT** has an important inferential element which helps in providing a bridge between large and small dimensional analysis and inference.
- Assumptions that underlie **OCMT** are in many ways weaker than those for penalised regression and can be relaxed in a more transparent manner due to **OCMT**'s roots in classical inference.

- **OCMT** can accommodate non-sparse $Cov(\mathbf{x}_{nt})$, so long as the correlations of signal and noise variables are sufficiently weak. No restrictions are imposed on the correlations amongst the noise variables.
- But like penalised and boosting techniques, **OCMT** only applies when the underlying DGP is sparse.
- Also, it is possible for **OCMT** to select some pseudo-signals (if they are sufficiently correlated with signals). In such cases one could apply standard model selection criteria (such as the Schwarz criterion) to the set of covariates selected by **OCMT** (which is likely to be a lot less than T, in practice).

Net Effect Coefficients

- Consider the data generating process (DGP) given by

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

- Following Pesaran and Smith (2014, Economics Letters) we introduce the following total or **'net' effect** of x_{it} on y_t:

$$\theta_{i,T} = \sum_{j=1}^{n} I(\beta_j \neq 0)\beta_j \sigma_{ij,T} = \sum_{j=1}^{k} \beta_j \sigma_{ij,T}, \quad (2)$$

where $\sigma_{ij,T} = E\left(T^{-1}\mathbf{x}_i' \mathbf{M}_\tau \mathbf{x}_j\right)$, $\mathbf{M}_\tau = \mathbf{I}_T - \boldsymbol{\tau}_T \boldsymbol{\tau}_T' / T$, and $\boldsymbol{\tau}_T = (1, 1, ..., 1)'$.

Net Effect Coefficients

- To simplify the notations we suppress the T subscript and use θ_i and σ_{ij} below.
- θ_i plays a crucial role in our proposed approach. We base inference on θ_i and then decide if such an inference can help in deciding whether or not $\beta_i = 0$.

Relationship between β_i and θ_i

- Ideally we would like to be able to base our selection decision directly on β_i and its estimate (jointly with the other covariates). But when n is large such a strategy is not feasible without some form of penalization.
- Instead of following the penalization route, 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.

The inverse mapping from θ_i to β_i

	$\theta_i \neq 0$	$\theta_i = 0$
$\beta_i \neq 0$	(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

- We consider each case in turn under appropriate restrictions.
- Cases I and IV, $\beta_i \neq 0$ iff $\theta_i \neq 0$ arise if signal and noise variables are uncorrelated.

- Case II, $\beta_i \neq 0$ and $\theta_i = 0$ can arise for some signals (not all), and will be covered by iterating on the OCMT procedure.
- Case III, $\beta_i = 0$ and $\theta_i \neq 0$ arises if the i^{th} covariate is correlated with one or more signal variables. To identify the signal variables we need these correlations to be reasonably weak, in the sense that $\sum_{j=k+1}^{n} |\theta_j| < K < \infty$ is satisfied. Finite k^* or (when $k^* \to \infty$) $\sum_{j=1}^{n} |\sigma_{ij}| < K < \infty$, for $i = 1, 2, ..., k$, is sufficient.

- To deal with Case II, we generalise θ_i to consider a conditional 'net' effect of x_{it} on y_t, where we condition on the effect of a subset of the signal (and pseudo signal) variables on y_t.
- We show this conditional 'net' effect will always be non-zero for some such subsets.
- We denote such conditional net effects of covariate i by $\theta_{i,T}(\boldsymbol{z}) = \sum_{j=1}^{k} \beta_j \sigma_{ij,T}(\boldsymbol{z})$ where $\sigma_{ij,T}(\boldsymbol{z}) = E\left(T^{-1}\boldsymbol{x}_i' \boldsymbol{M}_z \boldsymbol{x}_j\right)$, $\boldsymbol{M}_z = \boldsymbol{I}_T - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_T)'$, and $\boldsymbol{z}_t$ is a vector of variables that includes the constant and a subset of covariates. We suppress the T subscript and use $\theta_i(\boldsymbol{z})$ and $\sigma_{ij}(\boldsymbol{z})$ below.

The OCMT Approach

- We analyse the iterative scheme underlying OCMT. The need for an iterative scheme arises due to the possibility, discussed above, that $\theta_i = 0$ and $\beta_i \neq 0$.
- We call such signal variables hidden signals.
- Not all signal variables can be hidden.
- Once one conditions on the set of signal variables that are not hidden then, we show that there exists i such that $\theta_i(\boldsymbol{z}) \neq 0$, while $\theta_i = 0$ and $\beta_i \neq 0$, where $\boldsymbol{z}$ denotes the signal variables that are not hidden.
- Using this fact one can successively uncover all hidden signals (if any).

The OCMT Approach - Stage 1

- If we denote by P the set of iterations that need to be considered to uncover all hidden signals, then $1 \leq P_0 \leq k$, where P_0 is the true value of P.
- In the first stage, we run the n bivariate OLS regressions of y_t on x_{it},

$$y_t = c_{i,(1)} + \phi_{i,(1)} x_{it} + e_{it,(1)},\ t = 1, 2, ..., T, \text{ for } i = 1, 2, ..., n, \tag{3}$$

and consider the t-ratio of $\theta_{i,(1)}$, denoted by $t_{\hat{\theta}_{T,i,(1)}}$ in this simple regression.

The OCMT Approach -Stage 1

- The first stage multiple testing estimator of $I_{(1)}\left(\beta_i \neq 0\right)$ is given by

$$I_{(1)}\widehat{\left(\beta_i \neq 0\right)} = I\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n)\right], \text{ for } i = 1, 2, ..., n, \quad (4)$$

where

$$c_p(n, \delta) = \Phi^{-1}\left(1 - \frac{p}{2f\left(n, \delta\right)}\right) \quad (5)$$

is a **critical value function**, and $f\left(n, \delta\right) = n^{\delta}$, for $0 < \delta < \infty$. We use $\delta^{*} > \delta$ in higher stages.

- Covariates for which $I_{(1)}\widehat{\left(\beta_i \neq 0\right)} = 1$ are selected as signals or pseudo-signals.

- It is useful to note that $c_p^2(n,\delta) = O[\delta \ln(n)]$, and

$$\exp\left[-\frac{\varkappa c_p^2(n,\delta)}{2}\right] = \ominus\left(n^{-\delta\varkappa}\right). \quad (6)$$

- Assuming there exists $\kappa_1 > 0$, such that $T = \ominus(n^{\kappa_1})$, it follows that $c_p(n,\delta) = o\left(T^{C_0}\right)$, for all $C_0 > 0$.
- As we shall see, the OCMT procedure applies irrespective of whether n is small or large relative to T, so long as $T = \ominus(n^{\kappa_1})$, for any finite $\kappa_1 > 0$.

The OCMT Approach - Stage j>1

- Denote the number of covariates 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 $\boldsymbol{X}^s_{(1)}$.
- Further, 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 iteration 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 $i \in \mathcal{N}_{(j-1)}$.

The OCMT Approach - Stage j>1

- We then compute the following t-ratios

$$t_{\hat{\phi}_{T,i,(j)}} = \frac{\hat{\phi}_{T,i,(j)}}{s.e.\left(\hat{\phi}_{T,i,(j)}\right)} = \frac{\mathbf{x}_i' \boldsymbol{M}_{(j-1)} \mathbf{y}}{\hat{\sigma}_{i,(j)} \sqrt{\mathbf{x}_i' \boldsymbol{M}_{(j-1)} \boldsymbol{x}_i}}, \text{ for } i \in \mathcal{N}_{(j-1)},\ j \geq 2 \quad (7)$$

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

The OCMT Approach -Stage j>1

- Regressors for which $I_{(j)}\widehat{(\beta_i \neq 0)} = I\left[\left|t_{\hat{\phi}_{T,i,(j)}}\right| > c_p(n,\delta^*)\right] = 1$, are then added to the set of already selected signal variables from the previous stages.
- Denote the number of variables selected in stage j by $\hat{k}^s_{(j)}$, 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 $\mathbf{X}_{(j)} = (\mathbf{X}_{(j-1)}, \mathbf{X}^s_{(j)}) = (\boldsymbol{x}_{(j),1}, ..., \boldsymbol{x}_{(j),T})'$, $\hat{k}_{(j)} = \hat{k}_{(j-1)} + \hat{k}^s_{(j)}$, $\mathcal{S}_{(j)} = \mathcal{S}_{(j-1)} \cup \mathcal{S}^s_{(j)}$, and $\mathcal{N}_{(j)} = \{1, 2, ..., n\} \setminus \mathcal{S}_{(j)}$, and then proceed to stage $j+1$. Note that $\hat{\phi}_{T,i,(j)} \to_p \theta_{i,(j)}/\sigma_{ii} \equiv \theta_i\left(\boldsymbol{x}_{(j-1)}\right)/\sigma_{ii}$ and $\theta_i(\boldsymbol{z})$ denotes the conditional net effect introduced previously. Note also that $\theta_{i,(1)}$ is θ_i.

The OCMT Approach - Final Stage

- The procedure stops when no regressors are selected at a given stage j. Then stage $j-1$ will be denoted by $\hat{P}_{n,T}$, the estimator of P_0. So,

$$\hat{P}_{n,T} = \min\left\{ j : \sum_i I_{(j)}\left(\widehat{\beta_i \neq 0}\right) = 0 \right\} - 1, \tag{8}$$

  and

$$I\left(\widehat{\beta_i \neq 0}\right) = \sum_{j=1}^{\hat{P}_{n,T}} I_{(j)}\left(\widehat{\beta_i \neq 0}\right).$$

In the final step a multivariate regression of y_t on all the selected regressors is considered for inference and forecasting.

The OCMT approach: Assumptions

We consider the following assumptions:

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

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

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

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

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

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

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

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 $E\left(\boldsymbol{q}_{\cdot t} x_t\right)$ and $\boldsymbol{\Sigma}_{qq} = E\left(\boldsymbol{q}_{\cdot t}\boldsymbol{q}'_{\cdot t}\right)$ exist and $\boldsymbol{\Sigma}_{qq}$ is invertible. Define $\gamma_{qx,T} = \boldsymbol{\Sigma}_{qq}^{-1}\left[T^{-1}\sum_{t=1}^{T} E\left(\boldsymbol{q}_{\cdot t} x_t\right)\right]$ and $u_{x,t,T} =: u_{x,t} = x_t - \gamma'_{qx,T}\boldsymbol{q}_{\cdot t}$. All elements of the vector of projection coefficients, $\gamma_{qx,T}$, are uniformly bounded and only a finite number of the elements of $\gamma_{qx,T}$ are different from zero.

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

Discussion of the assumptions

- We allow for stochastic covariates, but require them to be martingale differences. This is less restrictive than the **IID** assumption often used in the literature.
- But the martingale difference assumption need not be imposed on the (pure) noise variables. Mixing can be used instead.
- The pure noise variables can have any arbitrary degree of correlation with the other noise variables.
- Exponential probability tail assumptions are ubiquitous in the literature.

Proposition

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

Example

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

Theorem 1 - Statement of conditions

Consider the DGP (1) with k signal variables. In addition suppose that there are k^* pseudo-signal variables and $n - k - k^*$ noise variables, and that Assumptions 1-4 and 6 hold. Assumption 5 holds for all pairs $(x_{it}, \boldsymbol{X}_{(j-1)})$, $i \in \mathcal{N}_{(j-1)}$, $j = 1, 2, \ldots$, where j denotes the stage of the OCMT procedure, and $\boldsymbol{X}_{(j-1)}$, and $\mathcal{N}_{(j-1)}$ are defined above.
$c_p(n, \delta)$ is given by (5) with $0 < p < 1$ and let $f(n, \delta) = cn^{\delta}$, for the first stage of OCMT and $f(n, \delta^*) = cn^{\delta^*}$, for subsequent stages, for some $c > 0$, $\delta^* > \delta > 0$. $n, T \to \infty$, such that $T = \ominus(n^{\kappa_1})$, for some $\kappa_1 > 0$, and $k^* = \ominus(n^{\epsilon})$ for some positive $\epsilon < \min\{1, \kappa_1/3\}$.

Theorem 1 - Probability of $\hat{P}_{n,T} > k$

For any $0 < \varkappa < 1$, for some constant $C_0 > 0$, and under the conditions above, the probability that the number of stages in the OCMT procedure, $\hat{P}_{n,T}$, defined by (8), exceeds k is given by

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

Theorem 1 - Probability of choosing the pseudo-true model

It is instructive to define formally the concept of the pseudo-true model, which we consider this to be a set of models. Each model in the set contains x_{it}, $i = 1, ..., k$. No model can contain any of the variables x_{it}, $i = k + k^* + 1, ..., n$. The models in the set may contain some or all of x_{it}, $i = k + 1, ..., k + k^*$. Formally, let

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

For any $0 < \varkappa < 1$, for some constant $C_0 > 0$, and under the conditions above, we have

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

Theorem 1 - True positive rate (TPR)

Consider the true positive rate ($TPR_{n,T}$),

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

For any $0 < \varkappa < 1$, for some constant $C_0 > 0$, and under the conditions above, we have

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

and if $\delta > 1 - \kappa_1/3$, then $TPR_{n,T} \to_p 1$;

Theorem 1 - False positive rate (FPR)

Consider the false positive rate ($FPR_{n,T}$),

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

For any $0 < \varkappa < 1$, for some constant $C_0 > 0$, and under the conditions above, we have

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

and if $\delta > \min\{0, 1-\kappa_1/3\}$, and $\delta^* > 1$, then $FPR_{n,T} \rightarrow_p 0$.

Theorem 1 - False discovery rate (FDR)

Consider the false discovery rate (if $\sum_{i=1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0$) defined by

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

For any $0 < \varkappa < 1$, for some constant $C_0 > 0$, and under the conditions above, we have

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

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

Theorem 1 - The residual norm of the selected model

Consider the norm: $F_{\tilde{u}} = \frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2$, where $\tilde{u}_t$ is the fitted value based on the estimates of the selected regression model.

Under the conditions above, we have

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

The norm of the estimated coefficients

- The OCMT estimator of β_i, denoted by $\tilde{\beta}_i$, is given by

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

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 $I\left(\widehat{\beta_i \neq 0}\right) = 1$, and a constant term.

- Consider the following norm:

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

We assume the following additional regularity condition.

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

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

Theorem 2: The norm of the estimated coefficients

Suppose conditions of Theorem 1 hold, and consider the coefficient norm of the selected model, $F_{\tilde{\beta}}$. In addition, suppose that Assumption 7 hold. Denote the maximum number of selected regressors that is allowed to enter the final stage regression by $l_{\max}$ and suppose that $l_{\max} = \ominus\left(n^{\kappa_2}\right)$, for some $\kappa_2 > 0$. Let $T = \ominus\left(n^{\kappa_1}\right)$, for some $\kappa_1 > 0$, $k^* = \ominus\left(n^{\epsilon}\right)$ for some positive $\epsilon < \min\left\{\kappa_2, \kappa_1/3\right\}$, $\delta^* > \delta > 1$ and $\delta^* > 2$. Then, for any $0 < \varkappa < 1$, and some constant $C_0 > 0$, we have

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

Main intuition behind the theoretical results

- Lemma 16 (stated also in the background slides) is a key lemma in the paper.
- Assuming $k^* = 0$ (for simplicity), and $T = \ominus(n^{\kappa_1})$, for some $\kappa_1 > 0$ (κ_1 could be much smaller than 1), Lemma 16 implies that for any $0 < \varkappa < 1$ there exist finite positive constants C_0 and C_1 such that

$$\sum_{i=k+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n)|\theta_i = 0\right] \leq$$

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

- But in view of (6), $(n-k)\exp\left[-\varkappa c_p^2(n,\delta)/2\right] = \ominus\left(n^{1-\delta\varkappa}\right) \to 0$ for $\delta > 1$ and $\varkappa < 1$. Therefore, the probability of selecting at least one noise variable in stage 1 declines to 0 with n at the rate $n^{1-\delta\varkappa}$.
- Higher stages are necessary only to uncover the *hidden signals*, and $\delta^* > 2$ is sufficient for the probability of selecting at least one noise variable in higher stages of OCMT to tend to 0 in n.

A Monte Carlo study

- We compare the small sample performance of the **OCMT** method with Lasso (and also SICA, Hard thresholding and boosting methods.)
- We consider four sets of designs depending on the choice of signal and noise θ's:

	Noise θ's	
Signal θ's	All are zero	Some are nonzero
All are nonzero	**Design set I**	**Design set II**
Some are zero	**Design set III**	**Design set IV**

- Design sets I-IV consider bounded number of signal variables. In addition to these four sets of designs, we also consider experiments with $k = n$ geometrically declining signal variables (**design set V**).

Design Set I: All signals have nonzero net effects

- DGP is given by the following model with $k = 4$ signals:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \varkappa u_t,\ u_t \sim IIDN\left(0, 1\right),$$

for $t = 1, 2, ..., T$. We set $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ and consider the following ways of generating $\mathbf{x}_t = \left(x_{1t}, x_{2t}, ..., x_{nt}\right)'$:

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

$$\text{signals: } x_{it} = \left(\varepsilon_{it} + g_t\right)/\sqrt{2}, \text{for } i = 1, 2, 3, 4, \quad (10)$$

$$\text{noise: } x_{5t} = \varepsilon_{5t},\ x_{it} = \left(\varepsilon_{i-1,t} + \varepsilon_{it}\right)/\sqrt{2}, \text{ for } i > 5, \quad (11)$$

where $g_t \sim IIDN\left(0, 1\right)$ and $\varepsilon_{it} \sim IIDN\left(0, 1\right)$. In this design $k^* = 0$, and the signal and noise variables are uncorrelated, but signals and noise variables are correlated with each other.

DGP-I(b) *Temporally correlated and weakly collinear covariates*: Variables are generated according to (10)-(11) with $\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + \sqrt{1-\rho_i^2} e_{it}$, $e_{it} \sim IIDN(0,1)$. We set $\rho_i = 0.5$ for all i.

DGP-I(c) *Strongly collinear noise variables due to a persistent unobserved common factor*. Signal variables are generated according to (10) and noise variables are generated as

$$x_{5t} = (\varepsilon_{5t} + b_i f_t) / \sqrt{3},\ x_{it} = \left[(\varepsilon_{i-1,t} + \varepsilon_{it}) / \sqrt{2} + b_i f_t \right] / \sqrt{3},$$

for $i > 5$, $b_i \sim IIDN(1,1)$, and $f_t = 0.95 f_{t-1} + \sqrt{1-0.95^2} v_t$, $v_t \sim IIDN(0,1)$, and $\varepsilon_{it} \sim IIDN(0,1)$.

DGP-I(d) *Equal (low or high) pair-wise correlation of signal variables*:

$$\text{signal variables: } x_{it} = (\varepsilon_{it} + \nu g_t) / \sqrt{1+\nu^2}, \text{ for } i = 1, 2, 3, 4,$$

and noise variables are generated according to (11), where $\varepsilon_{it} \sim IIDN(0,1)$, $g_t \sim IIDN(0,1)$ and we set $\nu = \sqrt{\omega/(1-\omega)}$, for $\omega = 0.2$ (low) and 0.8 (high). This ensures the correlation among the signal variables is ω. There is no correlation among noise variables.

- In all DGPs (in all design sets) $n = 100, 200, 300$, $T = 100, 300, 500$, and $R_{MC} = 2000$. In addition, we set $\varkappa$ so that $R^2 = 30\%$, 50% or 70%.
- To save on space, we average reported statistics across $R^2 = 30, 50, 70\%$, and report these averages only for n, $T = 100, 300$. Complete set of findings is provided in the MC Supplement.
- We report:
  - TPR, FPR and FDR,
  - the out-of-sample root mean square forecast error (RMSFE),
  - the root mean square error of $\tilde{\beta}$ ($\text{RMSE}_{\tilde{\beta}}$).
- Other statistics (such as the probability of selecting the correct model) are reported in the MC Supplement.

Table 1: MC findings for the first set of experiments

	DGP-I (averaged across $R^2 = 30\%, 50\%, 70\%$)							
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT
	$n = T = 100$				$n = 100,\ T = 300$			
TPR	1.000	0.909	0.734	0.906	1.000	0.987	0.916	0.999
FPR	0.000	0.054	0.010	0.000	0.000	0.054	0.005	0.000
FDR	0.000	0.478	0.149	0.005	0.000	0.461	0.065	0.003
RMSFE	3.419	3.551	3.570	3.460	3.362	3.408	3.407	3.363
$RMSE_{\tilde{\beta}}$	1.471	1.542	2.879	1.786	0.473	0.605	1.195	0.489
	$n = 300,\ T = 100$				$n = T = 300$			
TPR	1.000	0.898	0.745	0.877	1.000	0.986	0.919	0.999
FPR	0.000	0.034	0.010	0.000	0.000	0.026	0.004	0.000
FDR	0.000	0.615	0.289	0.006	0.000	0.539	0.129	0.003
RMSFE	3.418	3.602	3.635	3.480	3.362	3.421	3.414	3.363
$RMSE_{\tilde{\beta}}$	1.466	1.831	3.548	1.924	0.464	0.654	1.256	0.483

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$.

Design Set II: Featuring pseudo-signals

- y_t is generated in the same way as in the first set of designs, but we consider the following ways of generating $\mathbf{x}_t$:

DGP-II(a) *Two pseudo-signal variables:*

signal variables: $x_{it} = (\varepsilon_{it} + g_t) / \sqrt{2}$, for $i = 1, 2, 3, 4$,

noise variables: (pseudo-signal) $x_{5t} = \varepsilon_{5t} + \kappa x_{1t}$, $x_{6t} = \varepsilon_{6t} + \kappa x_{2t}$,

(pure noise) $x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it}) / \sqrt{2}$, for $i > 6$,

where $g_t \sim IIDN(0, 1)$, and $\varepsilon_{it} \sim IIDN(0, 1)$. 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*: $\mathbf{x}_t \sim IIDN(\mathbf{0}, \mathbf{\Sigma}_x)$ with the elements of $\mathbf{\Sigma}_x$ given by $0.5^{|i-j|}$, $1 \leq i, j \leq n$.

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

Table 2: MC findings for the second set of experiments

	DGP-II (averaged across $R^2 = 30\%, 50\%, 70\%$)							
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT
	$n = T = 100$				$n = 100,\ T = 300$			
TPR	1.000	0.913	0.754	0.890	1.000	0.993	0.944	1.000
FPR	0.000	0.060	0.011	0.008	0.000	0.062	0.005	0.014
FDR	0.000	0.514	0.170	0.145	0.000	0.508	0.077	0.225
RMSFE	3.283	3.425	3.445	3.335	3.226	3.274	3.269	3.234
$\text{RMSE}_{\tilde{\beta}}$	0.989	1.499	2.455	1.619	0.308	0.563	0.877	0.475
	$n = 300,\ T = 100$				$n = T = 300$			
TPR	1.000	0.904	0.765	0.859	1.000	0.991	0.944	0.999
FPR	0.000	0.037	0.011	0.003	0.000	0.029	0.005	0.004
FDR	0.000	0.643	0.314	0.135	0.000	0.578	0.142	0.217
RMSFE	3.284	3.478	3.513	3.355	3.226	3.290	3.279	3.234
$\text{RMSE}_{\tilde{\beta}}$	0.979	1.773	3.162	1.726	0.320	0.626	0.994	0.487

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$.

Design set III: Featuring signals with zero net effects

- Signal and noise variables $\{x_{it}\}$ are as in DGP-I(a) (see (10)-(11)), but β's are no longer equal to one in order to allow for zero net effects. We assume the fourth signal has zero net effect:

DGP-III. 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 $i \geq 4$.

Table 3: MC findings for the third set of experiments

	DGP-III (averaged across $R^2 = 30\%, 50\%, 70\%$)							
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT
	$n = T = 100$				$n = 100,\ T = 300$			
TPR	1.000	0.883	0.768	0.726	1.000	0.998	0.984	0.997
FPR	0.000	0.111	0.017	0.000	0.000	0.146	0.010	0.000
FDR	0.000	0.679	0.238	0.007	0.000	0.740	0.127	0.002
RMSFE	2.105	2.341	2.301	2.265	2.072	2.141	2.103	2.074
$\text{RMSE}_{\tilde{\beta}}$	0.404	2.324	2.340	2.145	0.128	0.717	0.568	0.222
	$n = 300,\ T = 100$				$n = T = 300$			
TPR	1.000	0.829	0.724	0.670	1.000	0.992	0.972	0.991
FPR	0.000	0.060	0.016	0.000	0.000	0.071	0.011	0.000
FDR	0.000	0.778	0.435	0.008	0.000	0.799	0.263	0.003
RMSFE	2.113	2.431	2.393	2.309	2.070	2.174	2.121	2.077
$\text{RMSE}_{\tilde{\beta}}$	0.410	2.878	2.854	2.448	0.128	1.041	0.721	0.330

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$.

Design set IV: Featuring signals with zero net effects and pseudo-signals

- We allow for both, zero net effects as well as pseudo-signals:

DGP-IV(a) We generate $\boldsymbol{x}_t$ 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 $\boldsymbol{x}_t$ 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 $i > 1$.

Table 4: MC findings for the fourth set of experiments

	DGP-IV (averaged across $R^2 = 30\%, 50\%, 70\%$)							
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT
	$n = T = 100$				$n = 100,\ T = 300$			
TPR	1.000	0.843	0.705	0.710	1.000	0.983	0.942	0.957
FPR	0.000	0.095	0.015	0.006	0.000	0.126	0.010	0.014
FDR	0.000	0.641	0.222	0.126	0.000	0.695	0.121	0.244
RMSFE	2.240	2.445	2.426	2.385	2.198	2.266	2.241	2.219
$\text{RMSE}_{\tilde{\beta}}$	0.458	2.027	2.235	1.889	0.143	0.743	0.727	0.440
	$n = 300,\ T = 100$				$n = T = 300$			
TPR	1.000	0.794	0.672	0.662	1.000	0.968	0.926	0.945
FPR	0.000	0.051	0.013	0.002	0.000	0.061	0.010	0.004
FDR	0.000	0.740	0.380	0.111	0.000	0.759	0.238	0.235
RMSFE	2.238	2.514	2.495	2.413	2.199	2.298	2.260	2.224
$\text{RMSE}_{\tilde{\beta}}$	0.444	2.437	2.665	2.100	0.145	0.985	0.855	0.510

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$.

Design set V: All variables are signals

- In the fifth set of experiments, we consider $k = n$ signal variables. MC Findings are reported assuming the first 11 variables are signals (as a benchmark).

DGP-V $\beta_i = 1/i^2$ and $\mathbf{x}_t \sim IIDN\left(\mathbf{0}, \mathbf{\Sigma}_x\right)$ with the elements of $\mathbf{\Sigma}_x$ given by $0.5^{|i-j|}$, $1 \leq i, j \leq n$.

Table 5: MC findings for the fifth set of experiments

	DGP-V (averaged across $R^2 = 30\%, 50\%, 70\%$)							
	Oracle*	Lasso	A-Lasso	OCMT	Oracle*	Lasso	A-Lasso	OCMT
	$n = T = 100$				$n = 100,\ T = 300$			
TPR*	1.000	0.270	0.122	0.197	1.000	0.359	0.147	0.300
FPR*	0.000	0.049	0.002	0.000	0.000	0.052	0.001	0.000
FDR*	0.000	0.474	0.058	0.006	0.000	0.435	0.016	0.003
RMSFE	1.371	1.362	1.358	1.326	1.311	1.314	1.328	1.299
$\text{RMSE}_{\tilde{\beta}}$	0.420	0.183	0.168	0.140	0.128	0.068	0.086	0.050
	$n = 300,\ T = 100$				$n = T = 300$			
TPR*	1.000	0.244	0.131	0.185	1.000	0.321	0.156	0.285
FPR*	0.000	0.031	0.004	0.000	0.000	0.024	0.002	0.000
FDR*	0.000	0.635	0.159	0.008	0.000	0.535	0.053	0.004
RMSFE	1.372	1.380	1.366	1.328	1.313	1.321	1.328	1.300
$\text{RMSE}_{\tilde{\beta}}$	0.424	0.227	0.233	0.149	0.127	0.076	0.089	0.049

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$. $(^*)$ Findings are reported assuming the first 11 variables are signals.

MC results: Lasso versus OCMT

- Although the **OCMT** *on average* beats **Lasso** in our set of experiments, it is not the case that **Lasso** is always outperformed by **OCMT**.
- There are instances, where **Lasso** beats **OCMT**, and none of the methods uniformly outperform others.
- **A-Lasso** improves on **Lasso**'s FDR/FPR, but it worsens TPR.
- The trade-offs between the **Lasso** and **OCMT** depend, in large part, on the correlation of the signal variables.
  - An increased correlation between the signals tends to increase the size of the net effect coefficients, and therefore the power of **OCMT** improves.
  - **Lasso** tends to deteriorate with an increase in the correlation of signal variables, since the marginal contribution of the signals to the overall fit diminishes.

MC evidence on the performance of other popular methods

- We also consider:
  - **Hard thresholding** (Zheng, Fan and Lv, 2014 JRSS)
  - **SICA** (Lv and Fan, 2009, Ann. Statist.). SICA stands for '*smooth integration of counting and absolute deviation*' - which refers to the type of penalty function used.
  - **Boosting methods**: We implement the boosting method proposed by Buhlmann (2006, Ann. Statist.) and consider two step sizes, $v = 0.1$ (recommended by Buhlmann), and $v = 1$.

Penalties

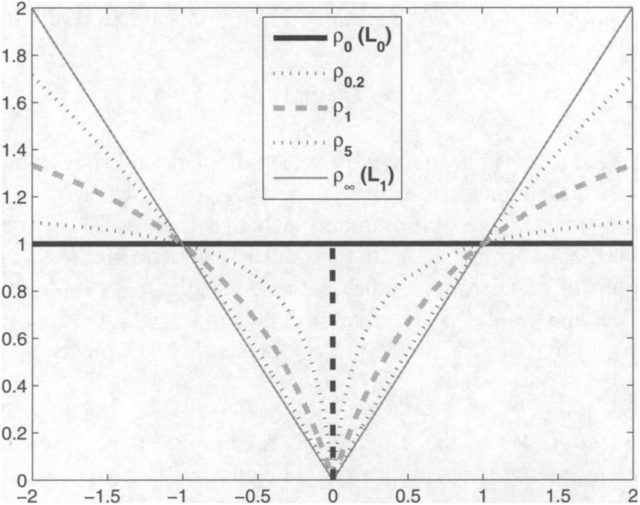
- Consider the penalised least squares

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}),\ Q(\boldsymbol{\beta}) = (2T)^{-1} \left\| \mathbf{y} - \sum_{i=1}^{n} \beta_i \boldsymbol{x}_i \right\|_2^2 + \| P_\lambda(\boldsymbol{\beta}) \|_1 ,$$

  where we use the compact notation
  $P_\lambda(\boldsymbol{\beta}) = P_\lambda(|\boldsymbol{\beta}|) = [p_\lambda(|\beta_1|), p_\lambda(|\beta_2|), ..., p_\lambda(|\beta_n|)]'$.
- Depending on the choice of the penalty function, we obtain:

$$\begin{aligned} \text{Lasso: } & p_\lambda(b) = \lambda b \\ \text{SICA: } & p_\lambda(b) = \lambda(a+1)b/(a+b) \ \text{ with } a = 10^{-4} \\ \text{Hard thresholding: } & p_\lambda(b) = \frac{1}{2}\left\{\lambda^2 - (\lambda - b)_+^2\right\},\ b \geq 0. \end{aligned}$$

- This figure (taken from Lv and Fan, 2009, Ann. Statist.) illustrates the role of SICA (for different values of a) and Lasso ($a = \infty$) penalties .



Computational demands: OCMT vs. other methods

- Computational demands of data-rich methods can be a problem in certain applications.
- **OCMT** is simple and fast. It takes less than 0.01 seconds to apply the **OCMT** in Matlab to a sample of $n = 300$ variables and $T = 100$ observations using a laptop.
- In contrast, penalised regressions are much more demanding. They take us about 200 to 10,000 times longer than the **OCMT** using the same hardware.
- The boosting method (with 500 iterations) is less demanding than the penalised regression methods - it takes 'only' about 50 times longer than **OCMT**.

MC results on the performance of OCMT, Lasso, Hard, Sica and Boosting methods

- We show the following statistics:
  - TPR, FPR and FDR (as before)
  - RMSFE *relative* to the benchmark Oracle method (rRMSFE)
  - $\text{RMSE}_{\tilde{\beta}}$ *relative* to the benchmark Oracle method ($\text{rRMSE}_{\tilde{\beta}}$)
  - the probability that regressors $1, 2, ..., k$ are among the selected ($\hat{\pi}_k$), and the probability of selecting the correct model ($\hat{\pi}$).
  - In addition to $\hat{\pi}$, we also report the probability of selecting pseudo-true model with all pseudo-signals, denoted by $\hat{\pi}^*$ in DGP-II(a) and DGP-IV(a).
- The benchmark model in DGP-V consists of the first 11 variables.

Table 7: Additional MC Findings for DGP-I(a)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	rRMSE$_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$
OCMT	100	0.9770	0.0001	0.0029	1.001	1.046	0.946	0.933
	300	0.9665	0.0001	0.0036	1.003	1.084	0.930	0.915
Lasso	100	0.9726	0.0546	0.4690	1.021	1.512	0.908	0.091
	300	0.9675	0.0285	0.5619	1.029	1.718	0.895	0.061
A-Lasso	100	0.8927	0.0060	0.0879	1.022	2.427	0.698	0.542
	300	0.8972	0.0061	0.1768	1.030	2.867	0.712	0.456
Sica	100	0.6796	0.0051	0.0960	1.054	5.545	0.339	0.231
	300	0.6199	0.0011	0.0723	1.066	6.785	0.267	0.211
Hard	100	0.6808	0.0016	0.0361	1.051	5.798	0.405	0.361
	300	0.6447	0.0005	0.0372	1.060	6.638	0.362	0.334
Boosting	100	0.9853	0.3356	0.8884	1.062	3.695	0.945	0.000
	300	0.9810	0.2748	0.9536	1.116	6.708	0.933	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 8: Additional MC Findings for DGP-I(b)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	rRMSE$_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$
OCMT	100	0.9770	0.0002	0.0047	1.002	1.054	0.945	0.924
	300	0.9683	0.0001	0.0051	1.003	1.079	0.931	0.908
Lasso	100	0.9719	0.0550	0.4745	1.021	1.504	0.905	0.086
	300	0.9681	0.0299	0.5821	1.028	1.716	0.896	0.052
A-Lasso	100	0.8910	0.0061	0.0892	1.023	2.427	0.697	0.544
	300	0.8961	0.0057	0.1808	1.029	2.788	0.707	0.447
Sica	100	0.6747	0.0051	0.0992	1.057	5.615	0.332	0.232
	300	0.6135	0.0012	0.0796	1.071	6.908	0.259	0.206
Hard	100	0.6735	0.0016	0.0373	1.054	6.012	0.391	0.352
	300	0.6352	0.0006	0.0410	1.064	6.598	0.348	0.322
Boosting	100	0.9836	0.3217	0.8839	1.064	3.657	0.940	0.000
	300	0.9804	0.2583	0.9507	1.118	6.436	0.930	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 9: Additional MC Findings for DGP-I(c)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	rRMSE$_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$
OCMT	100	0.9774	0.0001	0.0023	1.001	1.080	0.945	0.936
	300	0.9687	0.0001	0.0027	1.003	1.110	0.933	0.923
Lasso	100	0.9742	0.0412	0.4104	1.018	1.442	0.913	0.118
	300	0.9712	0.0214	0.5012	1.024	1.615	0.905	0.079
A-Lasso	100	0.8920	0.0042	0.0649	1.022	2.441	0.697	0.577
	300	0.8965	0.0037	0.1221	1.026	2.763	0.706	0.518
Sica	100	0.7091	0.0050	0.0934	1.048	5.065	0.378	0.260
	300	0.6520	0.0012	0.0746	1.059	6.148	0.305	0.237
Hard	100	0.6901	0.0017	0.0339	1.049	5.819	0.409	0.372
	300	0.6549	0.0005	0.0349	1.057	6.527	0.365	0.340
Boosting	100	0.9871	0.3272	0.8852	1.059	5.203	0.951	0.000
	300	0.9835	0.2124	0.9406	1.090	6.991	0.940	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 10: Additional MC Findings for DGP-I(d), $\omega = 0.2$ (low pair-wise correlation of signals)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	rRMSE$_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$
OCMT	100	0.9160	0.0001	0.0032	1.015	1.720	0.830	0.820
	300	0.8964	0.0000	0.0039	1.020	1.990	0.807	0.797
Lasso	100	0.9845	0.0784	0.5786	1.029	2.582	0.948	0.028
	300	0.9797	0.0402	0.6695	1.041	3.193	0.936	0.015
A-Lasso	100	0.9477	0.0074	0.1014	1.022	2.473	0.855	0.621
	300	0.9481	0.0073	0.2079	1.031	3.345	0.859	0.477
Sica	100	0.8777	0.0032	0.0588	1.032	3.459	0.696	0.597
	300	0.8391	0.0010	0.0605	1.044	4.419	0.626	0.555
Hard	100	0.8770	0.0020	0.0395	1.029	3.299	0.695	0.630
	300	0.8490	0.0008	0.0500	1.040	4.070	0.649	0.594
Boosting	100	0.9950	0.3392	0.8885	1.066	5.378	0.981	0.000
	300	0.9918	0.2698	0.9523	1.119	9.722	0.969	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 11: Additional MC Findings for DGP-II(a)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\hat{\pi}^*$
OCMT	100	0.9762	0.0192	0.3110	1.006	1.819	0.945	0.015	0.862
	300	0.9677	0.0061	0.3060	1.007	1.824	0.930	0.018	0.836
Lasso	100	0.9639	0.0568	0.4937	1.021	1.786	0.883	0.061	0.004
	300	0.9617	0.0292	0.5780	1.029	1.950	0.877	0.043	0.002
A-Lasso	100	0.8799	0.0065	0.1023	1.023	2.622	0.670	0.520	0.000
	300	0.8851	0.0063	0.1872	1.030	3.087	0.685	0.438	0.000
Sica	100	0.6661	0.0058	0.1183	1.055	6.320	0.316	0.219	0.000
	300	0.6081	0.0012	0.0876	1.067	7.431	0.251	0.201	0.000
Hard	100	0.6684	0.0020	0.0502	1.052	6.217	0.383	0.344	0.000
	300	0.6325	0.0006	0.0477	1.060	6.819	0.345	0.319	0.000
Boosting	100	0.9773	0.3370	0.8896	1.062	3.977	0.919	0.000	0.000
	300	0.9750	0.2762	0.9541	1.115	6.859	0.912	0.000	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 12: Additional MC Findings for DGP-II(b)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$
OCMT	100	0.9500	0.0059	0.1088	1.006	1.301	0.881	0.384
	300	0.9374	0.0016	0.0958	1.008	1.353	0.862	0.420
Lasso	100	0.9730	0.0645	0.5226	1.025	1.828	0.913	0.053
	300	0.9678	0.0335	0.6165	1.034	2.115	0.899	0.033
A-Lasso	100	0.9068	0.0068	0.0973	1.023	2.589	0.749	0.565
	300	0.9097	0.0068	0.1932	1.031	3.064	0.757	0.455
Sica	100	0.7197	0.0039	0.0758	1.048	5.809	0.394	0.303
	300	0.6663	0.0009	0.0597	1.058	7.074	0.320	0.273
Hard	100	0.7404	0.0016	0.0352	1.041	5.451	0.470	0.433
	300	0.7056	0.0006	0.0405	1.050	6.257	0.424	0.396
Boosting	100	0.9877	0.3703	0.8975	1.068	4.592	0.954	0.000
	300	0.9832	0.2713	0.9529	1.114	7.043	0.941	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 13: Additional MC Findings for DGP-III

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$
OCMT	100	0.9077	0.0001	0.0036	1.021	2.331	0.839	0.829
	300	0.8872	0.0001	0.0046	1.027	2.792	0.812	0.801
Lasso	100	0.9604	0.1365	0.7246	1.056	5.612	0.893	0.003
	300	0.9402	0.0685	0.7970	1.080	7.839	0.847	0.000
A-Lasso	100	0.9171	0.0120	0.1537	1.035	4.018	0.829	0.502
	300	0.8984	0.0128	0.3105	1.054	5.181	0.797	0.316
Sica	100	0.8592	0.0044	0.0830	1.045	5.022	0.712	0.580
	300	0.8008	0.0012	0.0818	1.064	7.036	0.628	0.541
Hard	100	0.9072	0.0024	0.0469	1.026	3.056	0.808	0.733
	300	0.8748	0.0010	0.0612	1.039	3.811	0.766	0.702
Boosting	100	0.9935	0.3606	0.8951	1.077	5.112	0.977	0.000
	300	0.9823	0.2559	0.9504	1.136	8.611	0.941	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 14: Additional MC Findings for DGP-IV(a)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	rRMSE$_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\hat{\pi}^*$
OCMT	100	0.9088	0.0172	0.2883	1.024	2.969	0.839	0.020	0.727
	300	0.8886	0.0054	0.2790	1.030	3.376	0.811	0.025	0.690
Lasso	100	0.9542	0.1391	0.7332	1.056	6.085	0.882	0.002	0.000
	300	0.9338	0.0691	0.8013	1.081	8.387	0.837	0.000	0.000
A-Lasso	100	0.9080	0.0130	0.1732	1.036	4.307	0.810	0.477	0.000
	300	0.8883	0.0131	0.3222	1.055	5.565	0.779	0.301	0.000
Sica	100	0.8273	0.0059	0.1212	1.051	9.685	0.610	0.493	0.000
	300	0.7715	0.0016	0.1157	1.070	11.308	0.544	0.472	0.000
Hard	100	0.8930	0.0030	0.0630	1.028	3.829	0.771	0.705	0.000
	300	0.8594	0.0011	0.0748	1.041	4.632	0.728	0.673	0.000
Boosting	100	0.9884	0.3623	0.8961	1.079	5.678	0.958	0.000	0.000
	300	0.9774	0.2572	0.9509	1.136	9.248	0.927	0.000	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 15: Additional MC Findings for DGP-IV(b)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	FDR	rRMSFE	rRMSE$_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$
OCMT	100	0.8573	0.0074	0.1435	1.023	2.936	0.584	0.102
	300	0.8321	0.0022	0.1332	1.028	3.359	0.518	0.087
Lasso	100	0.9282	0.0977	0.6344	1.041	4.209	0.782	0.008
	300	0.9031	0.0481	0.7168	1.057	5.367	0.707	0.003
A-Lasso	100	0.8457	0.0092	0.1172	1.036	4.644	0.647	0.441
	300	0.8284	0.0094	0.2374	1.049	5.470	0.604	0.300
Sica	100	0.7315	0.0039	0.0769	1.051	7.226	0.537	0.415
	300	0.6625	0.0009	0.0637	1.064	8.685	0.449	0.371
Hard	100	0.7811	0.0018	0.0378	1.038	6.159	0.620	0.566
	300	0.7420	0.0006	0.0440	1.046	6.860	0.570	0.527
Boosting	100	0.9862	0.3820	0.9008	1.075	5.278	0.951	0.000
	300	0.9703	0.2637	0.9523	1.123	8.039	0.898	0.000

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$.

Table 16: Additional MC Findings for DGP-V

Summary statistics are averaged across T and R^2

	n	TPR*	FPR*	FDR*	rRMSFE*	$\text{rRMSE}^*_{\tilde{\beta}}$	$\hat{\pi}_{11}$	$\hat{P}$
OCMT	100	0.2822	0.0002	0.0037	0.985	0.417	0.000	1.000
	300	0.2679	0.0001	0.0050	0.985	0.418	0.000	1.000
Lasso	100	0.3455	0.0525	0.4458	1.001	0.570	0.000	-
	300	0.3115	0.0266	0.5621	1.007	0.646	0.000	-
A-Lasso	100	0.1447	0.0017	0.0298	1.011	0.888	0.000	-
	300	0.1553	0.0024	0.0869	1.011	0.928	0.000	-
Sica	100	0.1234	0.0012	0.0277	1.012	1.372	0.000	-
	300	0.1118	0.0003	0.0225	1.015	1.433	0.000	-
Hard	100	0.1292	0.0010	0.0240	1.010	1.259	0.000	-
	300	0.1220	0.0004	0.0305	1.013	1.326	0.000	-
Boosting	100	0.5764	0.3703	0.8366	1.045	1.689	0.001	-
	300	0.5112	0.2732	0.9336	1.088	2.619	0.000	-

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$, and boosting is reported for $v = 0.1$. $(^*)$ Findings are reported assuming the first 11 variables are signals.

Empirical Illustration: Forecasting GDP growth and inflation

- We present a macroeconomic forecasting exercise for US GDP growth and CPI inflation using a set of 109 macroeconomic variables from Stock and Watson (2012, JBES).
- The dataset is quarterly and individual series are transformed to achieve stationarity. The transformed series span 1960Q3 to 2008Q4.
- We consider a rolling forecasting scheme with a rolling window of 120 observations. The forecast evaluation period is 1990Q3 to 2008Q4. We also consider the pre-crisis evaluation sub-period 1990Q3-2007Q2.

Forecasting methods:

1. $AR(1)$ benchmark;
2. $AR(1)$ augmented with principal components selected in a rolling scheme by the PC_{p_1} Bai and Ng (2002) information criterion.
3. Lasso and adaptive Lasso regressions of the dependent variable y_t on y_{t-1}, lagged principal components, and $\boldsymbol{x}_{t-1}$. For these regressions, both dependent variables and regressors (including principal components) are demeaned while regressors are normalised to have unit standard deviation. Then, the regression is run. Finally, the mean of y_t is added to produce the final forecast.
4. We apply OCMT to y_t using as regressors y_{t-1}, lagged principal components, and $\boldsymbol{x}_{t-1}$. In every OCMT regression the set of lagged principal components are always included.

Table 17: RMSFE performance of the AR, factor–augmented AR, Lasso, adaptive Lasso and OCMT methods

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

Notes: OCMT is reported for $(\delta, \delta^*) = (1, 2)$ and $p = 1\%$.

Conclusion

- Model specification and selection are recurring and fundamental topics in econometric analysis.
- Both become considerably more difficult for large-dimensional datasets.
- In the context of linear regression models, the penalised regression approach has become the *de facto* benchmark in the literature. However, issues such as the choice of penalty function and tuning parameters remain contentious.
- We provided an alternative 'multiple testing' (**OCMT**) approach, which is computationally much simpler and performs well in the case of sparse regression functions.
- There are a number of avenues for future research both in extending the **OCMT** approach to other modelling contexts and in its applications.

Background Slides

Lemma 16: Statement of assumptions. Let y_t, for $t = 1, 2, ..., T$, be given by the DGP (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-dimensional vector of signals that do not belong to $\boldsymbol{q}_{\cdot t}$, with the associated coefficients, $\boldsymbol{\beta}_b$. Assume $\boldsymbol{\Sigma}_{qq} = \frac{1}{T}\sum_{t=1}^{T} E\left(\boldsymbol{q}_{\cdot t}\boldsymbol{q}'_{\cdot t}\right)$ 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 $i = 1, 2, ..., l_T$. Let $l_T = o(T^{1/3})$ and suppose 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 orthogonal projection residuals as $u_{x,t} = x_t - \gamma'_{qx,T}\boldsymbol{q}_{\cdot t}$ and $e_t = y_t - \gamma'_{yqx,T}(\boldsymbol{q}'_{\cdot t}, x_t)'$. 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 $\theta = E\left(T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{X}_b\right)\boldsymbol{\beta}_b$, where $\mathbf{X}_b = (\boldsymbol{x}_{b,1}, \boldsymbol{x}_{b,2}, ..., \boldsymbol{x}_{b,T})'$. $\exists\ \kappa > 0$ such that $n = O\left(T^{\kappa}\right)$.

Lemma 16: Statement of results. Under the assumptions listed above, 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,

$$\Pr\left[|t_x| > c_p(n) \,|\theta_T = 0\right] \leq \exp\left[\frac{-\left(1-\pi\right)^2 \sigma^2_{e,(T)}\sigma^2_{x,(T)} c_p^2\left(n\right)}{2\left(1+d_T\right)^2 \omega^2_{xe,T}}\right] \tag{12}$$

$$+\exp\left[-C_0 T^{C_1}\right],$$

where

$$t_x = \frac{T^{-1/2}\mathbf{x}'\mathbf{M}_q\mathbf{y}}{\sqrt{\left(\mathbf{e}'\mathbf{e}/T\right)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}, \tag{13}$$

$$\sigma^2_{e,(T)} = E\left(T^{-1}\mathbf{e}'\mathbf{e}\right),\ \sigma^2_{x,(T)} = E\left(T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}\right), \tag{14}$$

and

$$\omega^2_{xe,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[\left(u_{x,t}\eta_t\right)^2\right]. \tag{15}$$

Lemma 16: Statement of results (continued). Under the assumptions listed above and under $\sigma_t^2 = \sigma^2$ and/or $E\left(u_{x,t}^2\right) = \sigma_{xt}^2 = \sigma_x^2$, for all $t = 1, 2, ..., T$,

$$\Pr\left[\left|t_x\right| > c_p(n) \left|\theta_T = 0\right.\right] \leq \exp\left[\frac{-\left(1-\pi\right)^2 c_p^2(n)}{2\left(1+d_T\right)^2}\right] + \exp\left(-C_0 T^{C_1}\right). \quad (16)$$

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

$$\Pr\left[\left|t_x\right| > c_p(n) \left|\theta_T \neq 0\right.\right] > 1 - \exp\left(-C_2 T^{C_3}\right). \quad (17)$$