Big Data Analytics: A New Perspective

A. Chudik
Federal Reserve Bank of Dallas

G. Kapetanios
King's College, London

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

Introduction

- Large data sets, "Big Data", are becoming increasingly available.
- Data may be text (emails), numeric, or video and take forms such as
  - spatiotemporal data sets covering zip codes, MSAs, counties or countries over long time periods;
  - scanner data from stores;
  - real-time financial data from computerised exchanges;
  - records of online searches;
  - records of traffic monitoring or CCTV cameras.

Challenges

- Dimensionality has many implications
  - Dimensions of the data (e.g., relative size of N and T in panels), dimensions of the parameter space (n); dimensions of the objects of interest (k).
  - Neyman & Scott (1948), consider the case where the number of unknown parameters rises with the sample size, and distinguish between a finite set of parameters of interest (k) and an infinite set of incidental parameters (n).
- But in many applications such a clear-cut division of the parameter space is not available.
- In linear panel models, one can often deal with incidental intercepts easily (e.g., demean or difference). This does not extend to non-linear models.

Focus of the analysis

With large data, the objects of interest may be

- particular subset of parameters: for instance, the focus could be a treatment effect, and one just needs to control for a large number of other **incidental parameters**. See, for example, Belloni et al. (2013).
- particular subset of variables: for instance, the focus is on forecasting a few variables, and one just needs to control for a large number of other **incidental variables**.
- the structure of interconnections between the elements of the network (e.g., variables, units).

Standard techniques

- Standard procedures suitable for models with a "small" number of parameters do not carry over to high-dimensional systems.
- The need to provide structure for the underlying relationships becomes more (and not less) important.
- Given that there are costs associated with using data, with large datasets we also need to consider:
  - the optimal amount of data to use, e.g., which variables to ignore
  - the frequency of data and model updates.

Techniques for the analysis of large data sets

- Methods for the analysis of large data sets can be classified into machine learning techniques (that are primarily applicable to observations that are distributed iid, such as random samples at a point in time) and time series econometric techniques that are also applicable to dependent observations.
- **Machine learning techniques:**
  - Classification and regression trees,
  - Penalised regressions,
  - Boosting,
  - Partial Least Squares (PLS).

Temporal dependence

- Once we allow for temporal dependence, the curse of dimensionality becomes even more serious.
- Structural breaks
  - In the case of temporal observations, the use of random sampling and cross-validation techniques are no longer appropriate. We need to develop techniques that do not depend on cross-validation.
  - Hal Varian in his recent JEL survey paper notes that most machine learning uses iid data, but economic data is rarely iid.

Econometric techniques

- There are a number of econometric techniques for the analysis of large data sets in the literature, including:
  - Bayesian shrinkage techniques (e.g. Banbura et al., 2010),
  - Factor models,
  - Spatio-temporal models,
  - High-dimensional VARs and global VARs (GVARs).

Penalised regressions

- Penalised (or regularised) 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 regressors (predictor variables), x_{it} for $i = 1, 2, .., n$, are typically
  - standardised and in some cases also orthogonalised (when PCs are used),
  - 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:

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

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

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

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 $[\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)']$

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

  $\mathbf{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})$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 computed using cross-validation techniques.

Pros of penalised regressions

- Penalised regressions, particularly Lasso, are easy to apply and have been shown to work well in the context of independently distributed observations.
- Although linear in structure, non-linear effects can also be included as predictors - such as threshold effects.

Cons of penalised regressions

- The use of cross-validation for the estimation of the penalty parameter, λ, presumes the underlying parameters and the covariance of the predictors, $Cov(\mathbf{x}_{nt})$, are stable over t.
- For large data sets arising in finance and macroeconomics, penalised least squares need not be appropriate.
- We need Big Data techniques that allow for temporal dependence and possible structural breaks.

Selected alternatives to penalised regressions

- **Boosting** (most widely known "Greedy method"), Friedman, Hastie, and Tibshirani (2000) and Friedman (2001): Regressors are chosen sequentially based on their individual ability to explain the dependent variable.
- **Stepwise regression** (Hocking, 1976). Two main approaches are common: (i) forward selection and (ii) backward elimination.
- **'General-to-Specific' modeling**, denoted as *Gets*, developed by David Hendry and coauthors, and its extension to the 'saturated' case when $n > T$ (Hendry and Krolzig, 2005, Section 7, and Santos, Hendry and Johansen, 2008)

Multiple Testing (MT) approach - a new method

- We propose a new method where regressors are selected **one at a time**, based on a t-test.
- The selected regressors from this stage are then used, in a second stage multiple regression, to provide the final coefficient estimate.
- In carrying out the t-tests we adjust the critical values to take into account the multiple testing nature of the problem.

- We refer to this new method as the multiple testing (**MT**) approach to variable selection.
- **MT** stands at the other extreme to the penalised regression technique that considers all regressors simultaneously. It has some similarity to boosting.
- Also unlike penalised regression and boosting, **MT** has an important inferential element which helps in providing a bridge between large and small dimensional analysis and inference.
- **MT** is very simple to apply and takes a fraction of the time needed to compute penalised regressions, particularly in the case of non-convex penalties.

- Assumptions that underlie **MT** are in many ways less strict that those for penalised regression and can be relaxed in a more transparent manner given **MT**'s roots in classical inference.
- **MT** can deal with 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 of noise variables.
- But like penalised and boosting techniques, **MT** only applies when the underlying DGP is sparse.
- Also, it is possible for **MT** to select some 'noise' variables if they are correlated with the 'signal' variables. In such cases one could apply standard model selection criteria (such as the Schwarz criterion) to the set of regressors selected by **MT** (which is likely to be a lot less than T, in practice).

Signal variables

- The data generating process (DGP) is given by

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

but the investigator is faced with n regressors, $\mathbf{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$, $n > k$, and possibly $n > T$. k is fixed but n and T could tend to infinity. The identity of the regressors, x_{it} for $i = 1, 2, ..., k$, which are placed at the start of $\mathbf{x}_{nt}$ for convenience, is not known. We refer to $x_{1t}, x_{2t}, ..., x_{kt}$ as 'signals'.

- More compactly we write the DGP as

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

where $\mathbf{X}$ is a $T \times k$ matrix of observations on the k signal variables, $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of ones, and $\mathbf{u}$ is a $T \times 1$ vector of errors.

Pseudo signal and noise variables

- We further assume that there are a finite number of 'pseudo signal' variables, k^*, which are not in the DGP but are correlated with the signals.
- As we shall see, it is possible to let the number of pseudo signals $\to \infty$, if we impose a certain boundedness condition on their co-variations with the signal variables.
- The remaining regressors, $x_{k^*+1,t}, x_{k^*+2,t}, \ldots, x_{nt}$, are noise variables.

Net effect coefficients

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

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

  where $\sigma_{ij} = Cov(x_{it}, x_{jt})$.
- $\theta_{i,(1)}$ measures the effect of x_{it} on y_t once its correlation with the other signal variables are taken into account.
- In what follows we will generalise this to consider the 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.

Relationship between β_i and $\theta_{i,(1)}$

- Ideally we would like to be able to base our selection decision directly on β_i and its estimate (jointly with the other factors). But when n is large such a strategy is not feasible.
- Instead we propose to base inference on $\theta_{i,(1)}$ and then decide if such an inference can help in deciding whether or not $\beta_i = 0$.
- It is important to stress that knowing $\theta_{i,(1)}$ does not imply we can determine β_i. But it is possible to identify conditions under which knowing $\theta_{i,(1)} = 0$ or $\theta_{i,(1)} \neq 0$ will help identify whether $\beta_i = 0$ or not.

The inverse mapping from $\theta_{i,(1)}$ to β_i

- There are three possibilities to consider:
  - (i) $\beta_i \neq 0$ if and only if $\theta_{i,(1)} \neq 0$,
  - (ii) $\beta_i = 0, \theta_{i,(1)} \neq 0$, and
  - (iii) $\beta_i \neq 0, \theta_{i,(1)} = 0$.
- We consider each in turn under appropriate restrictions.
- In practice case (iii) could become relevant even if $\theta_{i,(1)}$ are non-zero but very small.

- Case $\beta_i \neq 0$ if and only if $\theta_{i,(1)} \neq 0$ arises if signal and noise variables are uncorrelated.
- Case $\beta_i = 0, \theta_{i,(1)} \neq 0$ arises if the i^{th} noise 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 the boundedness condition, $\sum_{j=k+1}^{n} \left|\theta_{j,(1)}\right| < K < \infty$, is satisfied.
- It is clear that this condition is met if k^*, the number of correlated noise variables, is finite. But when $k^* \to \infty$, it is sufficient to assume

$$\sum_{j=1}^{n} |\sigma_{ij}| < K < \infty, \text{ for } i = 1, 2, ..., k. \tag{3}$$

- Case $\beta_i \neq 0, \theta_{i,(1)} = 0$ can arise for some signals (not all), and will be covered by an extension of this approach.
- To deal with this case, we generalise $\theta_{i,(1)}$ to consider a conditional "net" effect of x_{it}, with $\beta_i \neq 0$, 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 regressor i by $\theta_{i,(j)}$ for some j to be discussed later.

The MT approach

- We run the n bivariate OLS regressions of y_t on x_{it}, and compute the t-ratios

$$t_{\hat{\phi}_{i,(1)}} = \frac{\hat{\phi}_{i,(1)}}{s.e.\left[\hat{\phi}_{i,(1)}\right]} = \frac{\left(\boldsymbol{x}_i'\mathbf{M}_{(1)}\boldsymbol{x}_i\right)^{-1/2}\boldsymbol{x}_i'\mathbf{M}_{(1)}\mathbf{y}}{\hat{\sigma}_{i,(1)}}, \tag{4}$$

  for $i = 1, 2, ..., n$, where $\hat{\phi}_{i,(1)}$ denotes the estimated coefficient of x_{it} in the regression of $\mathbf{y}_t$ on $\mathbf{X}_{i,(1)} = (\boldsymbol{\tau}_T, \mathbf{x}_i)$. $\mathbf{M}_{(1)} = \mathbf{I}_T - \mathbf{X}_{(1)}\left(\mathbf{X}_{(1)}'\mathbf{X}_{(1)}\right)^{-1}\mathbf{X}_{(1)}'$, $\mathbf{X}_{(1)} = \boldsymbol{\tau}_T$, $\hat{\sigma}^2_{i,(1)} = \mathbf{y}'\mathbf{M}_{i,(1)}\mathbf{y}/T$, and $\mathbf{M}_{i,(1)} = \mathbf{I}_T - \mathbf{X}_{i,(1)}\left(\mathbf{X}_{i,(1)}'\mathbf{X}_{i,(1)}\right)^{-1}\mathbf{X}_{i,(1)}'$.

- The asymptotic analysis is simplified by basing the analysis on $z_{\hat{\phi}_{i,(1)}} = \frac{\left(\boldsymbol{x}_i'\mathbf{M}_{(1)}\boldsymbol{x}_i\right)^{-1/2}\boldsymbol{x}_i'\mathbf{M}_{(1)}\mathbf{y}}{\sigma}$ rather than $t_{\hat{\phi}_{i,(1)}}$, where $\sigma^2 = E(u_t^2)$.

The MT approach - continued

- Regressors for which $\widehat{I(\beta_i \neq 0)} = 1$ where

$$\widehat{I(\beta_i \neq 0)} = I\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,T)\right],\ i = 1,2,...,n, \quad (5)$$

  and $c_p(n,T) = \Phi^{-1}\left(1 - \frac{p}{2f(n)}\right)$, are selected as (possible) signals. $f(n) = n^{\delta}$, for $0 < \delta < \infty$.
- However, in cases where $\theta_{i,(1)} = 0$ even if $\beta_i \neq 0$, the tests will not be able to select the associated signal variable.

The MT approach: Assumptions

We consider the following assumptions:

Assumption

(a) The error term in DGP (1), u_t, is a martingale difference process with respect to $\mathcal{F}_{t-1}^{u} = \sigma\left(u_{t-1}, u_{t-2}, ...,\right)$. 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 $S_{nt} = \{x_{1t}, x_{2t}, ..., x_{nt}\}$, is independently distributed of the errors $u_{t'}$, for all t and t'.

Assumption

(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 (2) are nonzero for $i = 1, 2, ..., k$.

Assumption

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* $\mathcal{F}_{t}^{xn} = \cup_{j=k+k^{*}+1}^{n} \mathcal{F}_{jt}^{x}$, $\mathcal{F}_{t}^{xs} = \cup_{i=1}^{k+k^{*}} \mathcal{F}_{jt}^{x}$, *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* $E\left[x_{it}x_{jt} - E\left(x_{it}x_{jt}\right) \middle| \mathcal{F}_{t-1}^{x}\right] = 0$, *for* $i, j = 1, 2, ..., n$, *and all* t.

Assumption

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* $\mathcal{S}_{nt} = \{x_{1t}, x_{2t}, \ldots, x_{nt}\}$ *satisfy*

$$\sup_{i,t} \Pr\left(|x_{it}| > \alpha\right) \leq C_0 \exp\left(-C_1 \alpha^{s_x}\right), \text{ for all } \alpha > 0, \tag{6}$$

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

$$\sup_{t} \Pr\left(|u_t| > \alpha\right) \leq C_2 \exp\left(-C_3 \alpha^{s_u}\right), \text{ for all } \alpha > 0. \tag{7}$$

Discussion of the assumptions

- We allow for stochastic regressors, 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.

The MT approach: Theoretical results

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

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

Theorem

Under assumptions 1-4, and as long as $p > 0$ *and* $\log(n)/T \to 0$ *as* n *and* $T \to \infty$*, jointly, then*

$$\lim_{n,T\to\infty} E\left|TPR_{n,T}\right| = 1, \tag{8}$$

and

$$TPR_{n,T} \to_p 1, \text{ as } n \text{ and } T \to \infty, \text{ jointly.}$$

The MT approach: Theoretical results

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

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

Theorem

Under Assumptions 1, 3 and 4, and if $\theta_{i,(j)} = 0$, *for* $i = k + k^* + 1, \ldots, n$, *we have*

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

and $FPR_{n,T} \to_p 0$, *as* n *and* $T \to \infty$.

Note that $\exp\left[-\frac{c_p^2(n,T)}{2}\right] \to 0$ when $f(n) \to \infty$.

The MT approach: The probability of choosing the pseudo-true model

- It is instructive to define formally the concept of the pseudo-true model. In particular, 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}, $i = k + 1, ..., k + k^*$.
- The event of choosing the pseudo-true model is given by

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

The MT approach: The probability of choosing the pseudo-true model

Theorem

Under Assumptions 1-4, and $\theta_{i(j)} = 0$, *for* $i = k + k^* + 1, ..., n$, *there exist* $C_1, C_2 > 0$ *such that*

$$\Pr\left(\mathcal{A}_{true}\right) \geq 1 - C_0 \frac{n}{f\left(n\right)} - \exp\left(-C_1 T^{C_2}\right).$$

Further, for some $s = 1, ..., n - k - k^*$,

$$\Pr\left(\hat{k} - k - k^* > s\right) \leq \frac{\left(n - k - k^*\right)}{j} \left\{ \begin{array}{c} \exp\left[-\frac{\varkappa c_p^2(n)}{2}\right] + \\ O\left[\exp(-C_0 T^{C_1})\right] \end{array} \right\} \tag{9}$$

The MT approach: The norm of the in-sample error

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

Theorem

Under Assumptions 1-4, and if $\theta_i = 0$, *for* $i = k + k^* + 1, ..., n$, *and* $k^* = o\left(T^{1/3}\right)$,

$$E\left(\frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2\right) - \sigma^2 = o\left(1\right). \tag{10}$$

$$E\left(\frac{1}{T}\sum_{i=1}^{T}\tilde{u}_t^2\right) - \sigma^2 = O\left(\frac{1}{T}\right), \quad \text{if } n/f\left(n\right) = o\left(1/T\right). \tag{11}$$

The MT approach: The norm of the estimated coefficients

- Define $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_n)'$, where

$$\tilde{\beta}_i = \begin{cases} \hat{\beta}_i^{(\hat{k})}, & \text{if } I\widehat{(\beta_i \neq 0)} = 1 \\ 0, & \text{otherwise} \end{cases},$$

  and $\hat{\beta}_i^{(\hat{k})}$ is the OLS estimator of the coefficient of the i^{th} variable in a regression that includes all variables for which $I\widehat{(\beta_i \neq 0)} = 1$.

- Consider the following norm:

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

The MT approach: The norm of the estimated coefficients

Theorem

Under Assumptions 1-4, and technical conditions, and if $\theta_i = 0$, *for* $i = k + k^* + 1, \ldots, n$, *and* $k^* = o\left(T^{1/3}\right)$,

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

The Case $\beta_i \neq 0, \theta_{i,(1)} = 0$ - A simple example

As a simple example, suppose $k = 2$, (so $\beta_i \neq 0$, for $i = 1, 2$) and assume $\theta_{2,(1)} = 0$ even though $\beta_2 \neq 0$. Note that $\theta_{(1)} = (\theta_{1,(1)}, \theta_{2,(1)})' = \left[p\lim_{T\to\infty}\left(T^{-1}\mathbf{x}_1'\mathbf{M}_{(1)}\mathbf{X}\boldsymbol{\beta}\right),\ p\lim_{T\to\infty}\left(T^{-1}\mathbf{x}_2'\mathbf{M}_{(1)}\mathbf{X}\boldsymbol{\beta}\right)\right]$, and

$$\theta_{1,(1)} = \beta_1 + \sigma_{12}\beta_2 \neq 0,$$
$$\theta_{2,(1)} = \sigma_{12}\beta_1 + \beta_2 = 0.$$

Clearly, we can not have $\theta_{2,(1)} = 0$ if $\sigma_{12} = 0$. Suppose now that x_{1t} is selected in the first stage and note that a net effect in the second stage can be defined as $\theta_{2,(2)} = p\lim_{T\to\infty}\left(T^{-1}\mathbf{x}_2'\mathbf{M}_{(2)}\mathbf{X}\boldsymbol{\beta}\right)$, where $\mathbf{X}_{(2)} = (\boldsymbol{\tau}_T, \mathbf{x}_1)$ and $\mathbf{M}_{(2)} = \mathbf{I}_T - \mathbf{X}_{(2)}\left(\mathbf{X}_{(2)}'\mathbf{X}_{(2)}\right)^{-1}\mathbf{X}_{(2)}'$. Hence

$$\theta_{2,(2)} = \beta_2 p\lim_{T\to\infty}\left(T^{-1}\mathbf{x}_2'\mathbf{M}_{(2)}\mathbf{x}_2\right) = \beta_2\sigma_{22}\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right) \neq 0,$$

since x_{1t} and x_{2t} are not perfectly corrected and $\beta_2 \neq 0$.

The MT approach - The Case $\beta_i \neq 0, \theta_{i,(1)} = 0$
As we saw in the simple example, we can introduce further stages to the method to cover this case. The extended algorithm is given as follows

- Denote the number of variables selected at stage 1 by $\hat{k}^s_{(1)}$, the $T \times \hat{k}^s_{(1)}$ matrix of the observations on the $\hat{k}^s_{(1)}$ selected variables by $\mathbf{X}^s_{(1)}$. Finally, denote $\mathbf{X}_{(2)} = (\boldsymbol{\tau}_T, \mathbf{X}^s_{(1)})$, $\hat{k}_{(2)} = 1 + \hat{k}^{(s)}_{(1)}$, and the non-selected variables by $\mathbf{X}^{ns}_{(1)}$, such that $\mathbf{X}_n = \left(\mathbf{X}^s_{(1)}, \mathbf{X}^{ns}_{(1)}\right)$.
- In stages $j = 2, 3, ...,$ we proceed similarly and consider the $n - \hat{k}_{(j)}$ regressions of y_t on the variables in $\mathbf{X}_{(j)}$ and x_{it} for $i = 1, 2, ..., n - \hat{k}_{(j)}$ where $\boldsymbol{x}_i$ is a column of $\mathbf{X}^{ns}_{(1)}$, the matrix of non-selected regressors.

The MT approach - The Case $\beta_i \neq 0, \theta_{i,(1)} = 0$

Denote the number of variables selected by $\hat{k}^s_{(j)}$, the $T \times \hat{k}^s_{(j)}$ matrix of the $\hat{k}^s_{(j)}$ variables selected by $\mathbf{X}^s_{(j)}$ and the matrix containing the rest of the variables by $\mathbf{X}^{ns}_{(j)}$. Similarly, define $\mathbf{X}_{(j)} = (\mathbf{X}_{(j-1)}, \mathbf{X}^s_{(j)})$ and $\hat{k}_{(j)} = \hat{k}_{(j-1)} + \hat{k}^s_{(j)}$. Recall that $\mathbf{X}_{(1)} = \boldsymbol{\tau}_T$, $\hat{k}_{(0)} = 1$. Also define

$$\mathbf{M}_{(j)} = \mathbf{I}_T - \mathbf{X}_{(j)} \left(\mathbf{X}'_{(j)}\mathbf{X}_{(j)}\right)^{-1} \mathbf{X}'_{(j)}$$

$$\mathbf{X}_{i,(j)} = \left(\mathbf{x}_i, \mathbf{X}_{(j)}\right), \ \mathbf{M}_{i,(j)} = \mathbf{I}_T - \mathbf{X}_{i,(j)} \left(\mathbf{X}'_{i,(j)}\mathbf{X}_{i,(j)}\right)^{-1} \mathbf{X}'_{i,(j)}$$

The MT approach- The Case $\beta_i \neq 0, \theta_{i,(1)} = 0$

- In general, the t-ratios at stage j are given by

$$t_{\hat{\phi}_i,(j)} = \frac{\hat{\phi}_{i,(j)}}{s.e.\left[\hat{\phi}_{i,(j)}\right]} = \frac{\left(\boldsymbol{x}_i'\mathbf{M}_{(j)}\boldsymbol{x}_i\right)^{-1/2}\boldsymbol{x}_i'\mathbf{M}_{(j)}\mathbf{y}}{\hat{\sigma}_{i,(j)}}, \tag{12}$$

  where $\hat{\phi}_{i,(j)}$ is the estimated coefficient of $\mathbf{x}_i$ in the regression of $\mathbf{y}$ on $\mathbf{X}_{i,(j)}$, and $\hat{\sigma}^2_{i,(j)} = T^{-1}\boldsymbol{y}'\mathbf{M}_{i,(j)}\mathbf{y}$, for $j = 1, 2, ...$

- Regressors for which $\widehat{I\left(\beta_i \neq 0\right)} = I\left[\left|t_{\hat{\phi}_i,(j)}\right| > c_p(n,T)\right] = 1$ are selected as signals.

The MT approach- The Case $\beta_i \neq 0, \theta_{i,(1)} = 0$

- The procedure stops when no regressors are selected at a given stage, which we denote by stage J.
- Then $I\left(\widehat{\beta_i \neq 0}\right) = 1$ as long as $I\left[\left|t_{\hat{\phi}_i,(j)}\right| > c_p(n,T)\right] = 1$ for some $j = 1, ..., J$.
- It can be shown (see relevant Lemma in the paper) that if there are no pseudo signal variables then $J \leq k$.

An extension to serially correlated regressors

- A crucial assumption made in the previous exposition is that x_{it}, $i = 1, ..., n$ are martingale difference sequences.
- The idea behind the **MT** approach can apply in more general settings with some modifications.
- It is shown, in the paper, that all of our results can follow for mixing noise variables with exponentially declining mixing coefficients, such as autoregressive processes.
- Further, as long as $c_p(n, T)$ is allowed to grow faster than under the assumption of a martingale difference, signal variables can also be mixing. Again this is formally analysed in the paper.

A Monte Carlo study

- We compare the small sample performance of the **MT** method with a number of competing methods:
  - Penalised regression methods (described in the next slide): the Lasso (e.g. Tibshirani, 1996, JRRS), SICA (Lv and Fan, 2009, Ann. Statist.), and Hard thresholding (Zheng, Fan and Lv, 2014 JRSS). 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, $\nu = 0.1$ (recommended by Buhlmann), and $\nu = 1$.

Penalties

- Consider the penalised least squares

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

  where we use the compact notation
  $P_\lambda(\beta) = P_\lambda(|\beta|) = [p_\lambda(|\beta_1|), p_\lambda(|\beta_2|), \ldots, p_\lambda(|\beta_n|)]'$.

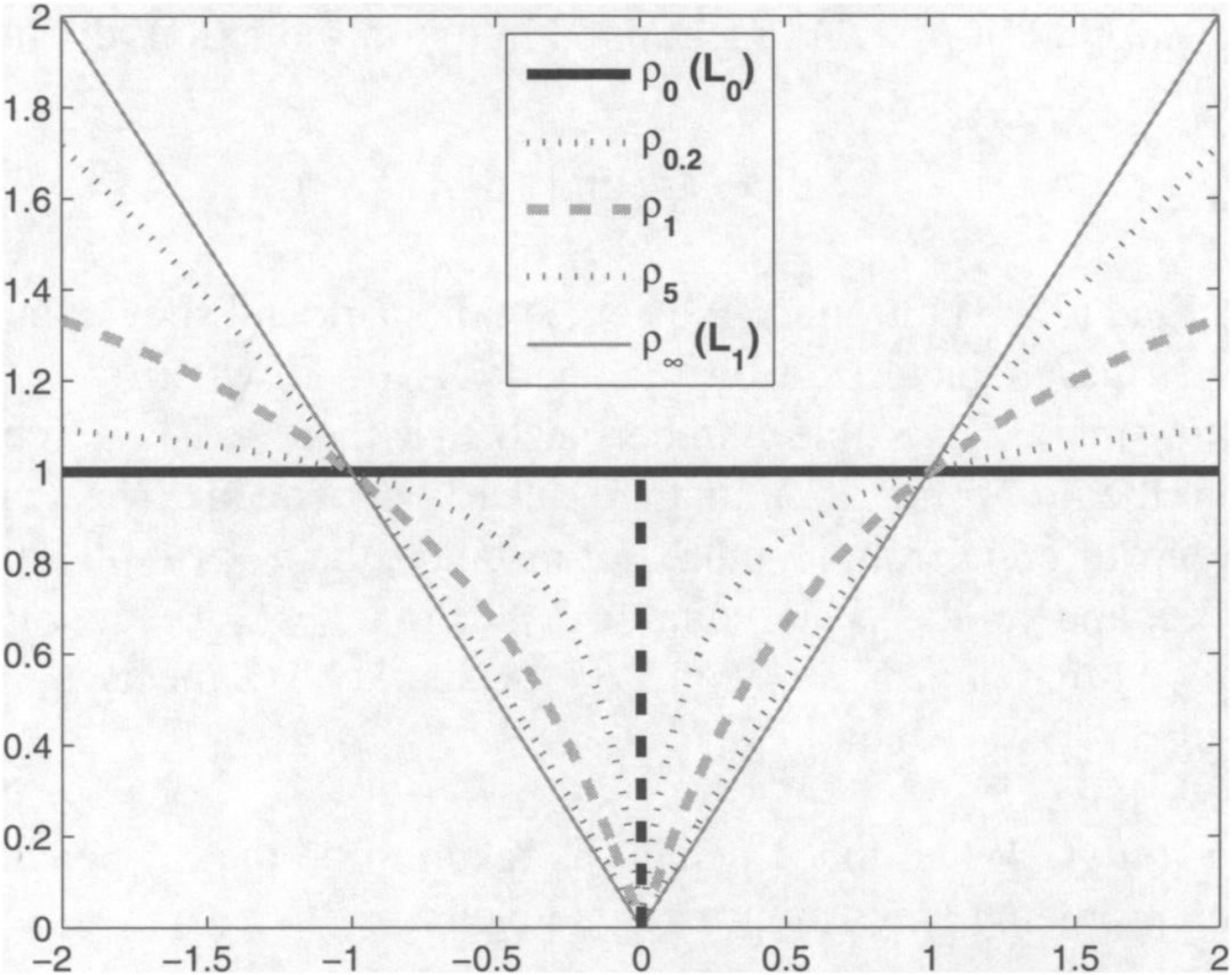
- Depending on the choice of the penalty function, we obtain:

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

$$\text{SICA: } p_\lambda(b) = \lambda(a+1)b/(a+b) \text{ with } a = 10^{-4}$$

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

- 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 .



Implementation of the Lasso, SICA and Hard thresholding

- We use **standardised regressors** $\tilde{x}_{it} = (x_{it} - \bar{x}_i)/s_{xi}$ and **de-meaned dependent variable** $\tilde{y}_t = y_t - \bar{y}$.
- We consider the same set of possible values for the penalisation parameter λ as in Zheng, Fan and Lv (2014, JRSS), namely
  $\lambda \in \Lambda \equiv \{\lambda_{\min}, \lambda_{\min} + \lambda_\epsilon, \lambda_{\min} + 2\lambda_\epsilon, ..., \lambda_{\max}\}$, where

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

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

  and $\lambda_\epsilon = (\lambda_{\max} - \lambda_{\min})/(K-1)$, with $K = 50$.
- We select λ using 10-fold cross-validation (in contrast with Zheng, Fan and Lv, 2014, JRSS)

Computational demands: MT vs. other methods

- Computational demands of data-rich methods can be a problem in certain applications.
- **MT** is simple and fast. It takes less than 0.01 seconds to apply the **MT** 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 **MT** 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 **MT**.

Monte Carlo designs

- 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 $k = n$ signal variables (**design set V**).

First set of designs (all signals have nonzero net effects)

- y_t is given by the following model:

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

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

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

$$\text{signals: } x_{it} = (\varepsilon_{it} + g_t)/\sqrt{2}, \text{ for } i = 1, 2, 3, 4, \tag{13}$$

$$\text{noise: } x_{5t} = \varepsilon_{5t},\ x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2}, \text{ for } i > 5, \tag{14}$$

where $g_t \sim IIDN(0,1)$ and $\varepsilon_{it} \sim IIDN(0,1)$.

DGP-I(b) *Temporally correlated and weakly collinear regressors*: Variables are generated according to (13)-(14) 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 (13) 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*:

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

and noise variables are generated according to (14), 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.

Second set of designs (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:*

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

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 **MT** procedure is expected to pick up the pseudo-signals in DGP-II(a) with a high probability, but $\tilde{\beta}$ remains consistent in the sense that $\left\|\tilde{\beta} - \beta\right\|_F^2 \to 0$ (see Theorem 5). $\tilde{\beta}$ will be asymptotically less efficient than the estimates of the true model due to the presence of pseudo-signals.

Third set of designs (featuring signals with zero net effects)

- Signal and noise variables $\{x_{it}\}$ are as in DGP-I(a) (see (13)-(14)), 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 $i \geq 4$.

Fourth set of designs (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 $i > 1$.

Fifth set of designs (featuring $k = n$ signal variables)

- In the fifth set of experiments, we consider $k = n$ signal variables.

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

- In all DGPs we set $\varkappa$ so that $R^2 = 30\%$, 50% or 70%.
- We consider $n = 100, 200$ and 300, $T = 100, 300$ and 500, and compute $R_{MC} = 2000$ Monte Carlo replications.
- We report:
  - true and false positive rates (TPR and FPR, respectively),
  - the out-of-sample root mean square forecast error *relative* to the true benchmark model (rRMSFE),
  - the root mean square error of $\tilde{\boldsymbol{\beta}}$ *relative* to the true benchmark model (rRMSE$_{\tilde{\boldsymbol{\beta}}}$),
  - the probability that regressors $1, 2, ..., k$ are among the selected ($\hat{\pi}_k$), and the probability of selecting the correct model ($\hat{\pi}$).

MC findings for the first set of experiments

- To save on space, we average reported statistics across $R^2 = 30, 50, 70\%$ and across $T = 100, 300, 500$, and report these averages only for $n = 100$ and 300. We report MT method for $p = 0.01$ and $\delta = 1$ only.
- In the case of MT, we also report the average number of iterations before convergence (r).
- Complete set of findings is provided in the MC Supplement.

Table 1: MC Findings for DGP-I(a)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	r
MT	100	0.9769	0.0003	1.002	1.084	0.95	0.92	0.012
$(p = 0.01, \delta = 1)$	300	0.9681	0.0001	1.003	1.129	0.93	0.91	0.012
Lasso	100	0.9723	0.0541	1.021	1.513	0.91	0.09	-
	300	0.9669	0.0282	1.029	1.715	0.89	0.06	-
Sica	100	0.6818	0.0016	1.050	5.692	0.40	0.36	-
	300	0.6440	0.0005	1.059	6.551	0.36	0.33	-
Hard	100	0.6805	0.0050	1.054	5.511	0.34	0.23	-
	300	0.6221	0.0011	1.065	6.695	0.27	0.21	-
Boosting	100	0.9850	0.3360	1.062	3.726	0.94	0.00	-
$(v = 0.1)$	300	0.9813	0.2750	1.115	6.691	0.93	0.00	-

Table 2: MC Findings for DGP-I(b)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	r
MT	100	0.9768	0.0003	1.002	1.087	0.94	0.92	0.010
$(p = 0.01, \delta = 1)$	300	0.9663	0.0001	1.004	1.140	0.93	0.89	0.013
Lasso	100	0.9710	0.0557	1.021	1.501	0.90	0.08	-
	300	0.9675	0.0296	1.028	1.705	0.89	0.05	-
Sica	100	0.6731	0.0017	1.055	6.019	0.39	0.35	-
	300	0.6363	0.0006	1.065	6.728	0.35	0.32	-
Hard	100	0.6727	0.0054	1.058	5.682	0.33	0.23	-
	300	0.6141	0.0012	1.070	6.846	0.26	0.20	-
Boosting	100	0.9835	0.3224	1.064	3.629	0.94	0.00	-
$(v = 0.1)$	300	0.9807	0.2581	1.118	6.419	0.93	0.00	-

Table 3: MC Findings for DGP-I(c)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	r
MT	100	0.9761	0.0002	1.002	1.159	0.94	0.93	0.007
$(p = 0.01, \delta = 1)$	300	0.9682	0.0001	1.003	1.297	0.93	0.91	0.009
Lasso	100	0.9737	0.0415	1.018	1.453	0.91	0.12	-
	300	0.9711	0.0211	1.024	1.598	0.90	0.08	-
Sica	100	0.6895	0.0016	1.049	5.843	0.41	0.37	-
	300	0.6546	0.0005	1.057	6.454	0.37	0.34	-
Hard	100	0.7103	0.0051	1.048	5.134	0.38	0.26	-
	300	0.6515	0.0012	1.060	6.078	0.30	0.24	-
Boosting	100	0.9869	0.3277	1.059	5.258	0.95	0.00	-
$(\nu = 0.1)$	300	0.9835	0.2125	1.091	6.949	0.94	0.00	-

Table 4: 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	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	r
MT	100	0.9183	0.0003	1.015	1.711	0.84	0.82	0.020
$(p = 0.01, \delta = 1)$	300	0.8984	0.0001	1.020	1.968	0.81	0.79	0.024
Lasso	100	0.9848	0.0791	1.029	2.576	0.95	0.03	-
	300	0.9799	0.0404	1.041	3.170	0.94	0.02	-
Sica	100	0.8770	0.0021	1.030	3.420	0.70	0.63	-
	300	0.8512	0.0008	1.038	3.912	0.65	0.60	-
Hard	100	0.8794	0.0033	1.032	3.459	0.70	0.60	-
	300	0.8399	0.0009	1.043	4.365	0.63	0.56	-
Boosting	100	0.9951	0.3399	1.065	5.391	0.98	0.00	-
$(v = 0.1)$	300	0.9914	0.2699	1.119	9.648	0.97	0.00	-

MC findings for the second set of designs

- In the second set of experiments, we allow for pseudo-signals (i.e. variables with $\beta_i = 0$ and $\theta_i \neq 0$). The **MT** procedure picks up all variables with θ_i sufficiently large with a high probability (we require $T\theta_i^2/ln(n) \to \infty$). Nevertheless, $\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_F^2$ will converge to zero and the RMSE of $\tilde{\boldsymbol{\beta}}$ *relative* to the true model remains bounded as $T \to \infty$.
- 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).
- We report findings for both DGP-II(a) and DGP-II(b) below.

Table 5: MC Findings for DGP-II(a)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\hat{\pi}^*$	r
MT	100	0.9768	0.0194	1.006	1.862	0.95	0.01	0.85	0.010
$(p = 0.01, \delta = 1)$	300	0.9667	0.0061	1.007	1.842	0.93	0.02	0.83	0.014
Lasso	100	0.9650	0.0577	1.022	1.807	0.88	0.06	0.00	-
	300	0.9604	0.0293	1.029	1.947	0.87	0.05	0.00	-
Sica	100	0.6685	0.0020	1.052	6.129	0.38	0.35	0.00	-
	300	0.6303	0.0006	1.061	6.979	0.34	0.32	0.00	-
Hard	100	0.6650	0.0057	1.055	6.320	0.31	0.22	0.00	-
	300	0.6077	0.0012	1.067	7.421	0.25	0.20	0.00	-
Boosting	100	0.9788	0.3377	1.062	3.984	0.92	0.00	0.00	-
$(v = 0.1)$	300	0.9743	0.2760	1.116	6.860	0.91	0.00	0.00	-

Table 6: MC Findings for DGP-II(b)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	r
MT	100	0.9514	0.0059	1.007	1.349	0.88	0.39	0.013
$(p = 0.01, \delta = 1)$	300	0.9376	0.0017	1.009	1.417	0.86	0.41	0.016
Lasso	100	0.9737	0.0644	1.025	1.843	0.91	0.05	-
	300	0.9679	0.0334	1.034	2.148	0.90	0.03	-
Sica	100	0.7402	0.0016	1.041	5.408	0.47	0.43	-
	300	0.7054	0.0006	1.049	6.249	0.42	0.39	-
Hard	100	0.7207	0.0038	1.047	5.849	0.39	0.30	-
	300	0.6656	0.0009	1.059	7.175	0.32	0.27	-
Boosting	100	0.9884	0.3695	1.068	4.618	0.96	0.00	-
$(\nu = 0.1)$	300	0.9833	0.2715	1.114	7.153	0.94	0.00	-

MC findings for the third set of designs

- In the third set of experiments, we allow for signal variables with zero net effects, namely $\beta_i \neq 0$ and $\theta_i = 0$.

Table 7: MC Findings for DGP-III

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	r
MT	100	0.9184	0.0003	1.017	2.020	0.86	0.84	0.920
$(p = 0.01, \delta = 1)$	300	0.9015	0.0001	1.022	2.290	0.84	0.81	0.902
Lasso	100	0.9600	0.1367	1.056	5.663	0.89	0.00	-
	300	0.9394	0.0679	1.080	7.857	0.84	0.00	-
Sica	100	0.9069	0.0024	1.026	3.010	0.81	0.73	-
	300	0.8737	0.0010	1.039	3.824	0.77	0.70	-
Hard	100	0.8587	0.0045	1.045	5.140	0.71	0.57	-
	300	0.7975	0.0012	1.065	7.185	0.62	0.54	-
Boosting	100	0.9938	0.3606	1.078	5.164	0.98	0.00	-
$(\nu = 0.1)$	300	0.9821	0.2559	1.135	8.621	0.94	0.00	-

MC findings for the fourth set of designs

- In the fourth set of experiments, we allow for pseudo-signals (i.e. variables with $\beta_i = 0$ and $\theta_i \neq 0$) as well as signals with zero net effect.

Table 8: MC Findings for DGP-IV(a)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	$\hat{\pi}^*$	r
MT	100	0.9198	0.0174	1.020	2.649	0.86	0.03	0.72	0.919
$(p = 0.01, \delta = 1)$	300	0.9034	0.0055	1.024	2.904	0.84	0.03	0.69	0.903
Lasso	100	0.9544	0.1393	1.056	6.106	0.88	0.00	0.00	-
	300	0.9324	0.0689	1.081	8.301	0.83	0.00	0.00	-
Sica	100	0.8925	0.0029	1.028	3.807	0.77	0.70	0.00	-
	300	0.8600	0.0011	1.041	4.636	0.73	0.67	0.00	-
Hard	100	0.8282	0.0060	1.050	9.708	0.62	0.50	0.00	-
	300	0.7682	0.0016	1.071	11.540	0.54	0.46	0.00	-
Boosting	100	0.9894	0.3623	1.078	5.706	0.96	0.00	0.00	-
$(v = 0.1)$	300	0.9772	0.2571	1.135	9.164	0.93	0.00	0.00	-

Table 9: MC Findings for DGP-IV(b)

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	$\text{rRMSE}_{\tilde{\beta}}$	$\hat{\pi}_k$	$\hat{\pi}$	r
MT	100	0.8921	0.0076	1.016	2.325	0.72	0.16	0.730
($p = 0.01$, $\delta = 1$)	300	0.8729	0.0022	1.020	2.558	0.68	0.17	0.697
Lasso	100	0.9287	0.0982	1.042	4.237	0.79	0.01	-
	300	0.9046	0.0481	1.057	5.451	0.71	0.00	-
Sica	100	0.7829	0.0019	1.037	6.120	0.63	0.57	-
	300	0.7424	0.0007	1.047	6.762	0.57	0.53	-
Hard	100	0.7309	0.0038	1.051	7.299	0.54	0.41	-
	300	0.6646	0.0009	1.064	9.051	0.45	0.37	-
Boosting	100	0.9857	0.3826	1.075	5.335	0.95	0.00	-
($\nu = 0.1$)	300	0.9697	0.2637	1.123	8.150	0.90	0.00	-

MC findings for the fifth set of designs

- In the fifth set of experiments, we set $\beta_i = 1/i^2$, namely $n = k$ and all variables are signal variables.
- We report (using the model consisting of the first 11 variables as a benchmark):
  - true and false positive rates assuming variables $i \leq 11$ are signal variables.(TPR and FPR, respectively)
  - the out-of-sample root mean square forecast error *relative* to the benchmark model (rRMSFE)
  - RMSE of $\tilde{\beta}$ relative to the benchmark model (rRMSE$_{\tilde{\beta}}$)
  - probability of the first 11 regressors being among the selected variables, denoted as π_{11}

Table 10: MC Findings for DGP-V

Summary statistics are averaged across T and R^2

	n	TPR	FPR	rRMSFE	rRMSE_{β}	$\hat{\pi}_{11}$	r
MT	100	0.2820	0.0003	0.986	0.433	0.00	0.018
$(p = 0.01, \delta = 1)$	300	0.2691	0.0001	0.986	0.443	0.00	0.019
Lasso	100	0.3450	0.0522	1.001	0.570	0.00	-
	300	0.3121	0.0265	1.008	0.647	0.00	-
Sica	100	0.1294	0.0011	1.010	1.264	0.00	-
	300	0.1216	0.0004	1.014	1.351	0.00	-
Hard	100	0.1231	0.0012	1.012	1.374	0.00	-
	300	0.1117	0.0003	1.015	1.433	0.00	-
Boosting	100	0.5751	0.3696	1.045	1.683	0.00	-
$(v = 0.1)$	300	0.5119	0.2731	1.089	2.620	0.00	-

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 statistical and econometric analysis. However, issues such as the choice of penalty function and tuning parameters remain contentious.

- We provided an alternative 'multiple testing' (**MT**) approach, which is computationally much simpler and performs well in the case of sparse regression functions.
- Extensive theoretical and Monte Carlo results provide support for adding this method to the toolbox of the applied researcher.
- There are a number of avenues for future research both in extending the **MT** approach to other modelling contexts and in its applications to a wide variety of problems in economics, finance, as well as in other disciplines.