Screening High Dimensional Time Series via Tilting

Department of Statistics 2019

A thesis submitted for the degree of Master of Science

London School of Economics, University of London

Candidate: 13185

Supervisor: Professor Piotr Fryzlewicz

Abstract

Developing effective feature screening procedures when the number of features exceeds the sample size ($p \gg n$) is one of the most active research areas in modern statistics. Although much attention has been devoted to the problem of detecting causal and predictive relationships in high dimensional regimes, comparatively little attention has been given to the problem of developing general purpose screening procedures for high dimensional time series data. This is especially surprising given high dimensional time series are becoming increasingly common in many fields including economics, finance, and neuro-science.

The aim of this dissertation is to adapt tilted correlation screening, a variable selection procedure for linear models introduced by Cho & Fryzlewicz (2012), to settings in which the response, predictor, and error variables are allowed to be time series processes. Overall, the dissertation aims to make the following original contributions:

1. Show that under mild assumptions the main theoretical results in Cho & Fryzlewicz (2012) hold for time series data.

2. Introduce a generalised least squares variant of the tilted correlation which is more efficient than the original (in the Gauss-Markov / MSE sense).

3. Apply the principle of stability selection introduced by Meinshausen & Bühlmann (2010) to tilted correlations when the correlation structure of the predictor variables necessitates careful control.

The theoretical results presented are reinforced by simulation studies. Additionally, a real data application is provided in which tilted correlation screening is used to select a model for forecasting CPI inflation, chained GDP, and the Sterling effective exchange rate.

Key words: high dimensionality, mixing, sparsity, tilting, time series, variable screening.

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Professor Piotr Fryzlewicz, firstly for agreeing to supervise this project and secondly for giving up so much of his time to guide me through the research process. I would also like to thank Professor Clifford Lam who first interviewed me for the ESRC Studentship which has funded my masters. Finally I would like to thank Dr Yining Chen for introducing me to parallel computing, without which this dissertation would have take roughly seven times longer to write.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The need to analyse vast quantities of data is becoming increasingly common across most-all disciplines. By way of example, 90 percent of the world's data in 2013 had been generated in the last two years - see Åse (2013). Motivated by the growing challenges posed by 'Big Data', this dissertation studies the broad problem of selecting features for a high dimensional linear model with the additional requirement that the data may be from a time series process. The main contribution is the extension of tilted correlation screening (TCS), a variable selection procedure for linear models introduced by Cho & Fryzlewicz (2012), to the time series setting.

High dimensionality refers to the setting in which the number of features present in a dataset exceeds the number of observations ($p \gg n$). At the most fundamental level high dimensionality rules out parameter estimation by least squares (the workhorse of modern statistics), as the number of equations to be solved will by construction be smaller than the number of unknowns. Moreover, spurious collinearity among predictors and noise accumulation have been shown to degrade the properties of feature screening procedures which perform well in traditional settings; see Fan & Lv (2010) for an overview. The most popular approach to model building in high dimensions has been to invoke the sparsity assumption: a prior belief that only a subset of features ($s \ll n$) is relevant to the relationship of interest. The assumption has lead to two methodologies. These are penalised estimation methods such as the Lasso of Tibshirani (1996) and its many variants, and screening methods such

as Sure Independence Screening of Fan & Lv (2008) and its variants.

While existing methodologies have been widely adopted, most assume i.i.d. observations and are therefore poorly suited to the time series setting. High dimensional time series are nonetheless common in many fields. Examples can be found in neuro-science - see for example Valdés-Sosa et al. (2005), macroeconomics - see Stock & Watson (2002), and finance - see Choi et al. (2019). Moreover, even when the original dataset is not highly dimensional, uncertainty over the functional form with which each feature enters into the model often leads practitioners to take transformations of features (polynomials, splines, interactions, etc.) which can result in a very large number of 'technical' features, see for example Belloni et al. (2014). The need to advance the literature on general purpose screening procedures is therefore clear.

The dissertation is structured as follows. The remainder of Chapter 1 briefly reviews the existing literature on general purpose screening procedures for high dimensional time series, and introduces notation used throughout the dissertation. Chapter 2 formally introduces the TCS algorithm of Cho & Fryzlewicz (2012), and presents popular techniques from the high dimensional literature which will be used to extend tilted correlations to the time series setting. Original contributions are presented in Chapter 3: Section 3.1 investigates the properties of tilted correlations in a time series setting, while in Sections 3.2 and 3.3 propose two extensions which exploit the time series setting to improve the screening accuracy of the original procedure. Chapter 4 investigates the screening properties of tilted correlations in high dimensional time series through simulation studies and a real data example. Chapter 5 offers some concluding remarks. Finally, note that R code for simulation studies and data analysis carried out in this dissertation in available via GitHub.

1.1 Approach to screening high dimensional time series

Fan & Lv (2008) spearheaded feature screening screening for high dimensional linear models with Sure Independence Screening (SIS). The procedure functions by ranking each feature according to its marginal utility. Specifically, let $\boldsymbol{\omega} = (\omega_1, ..., \omega_p)^T = \boldsymbol{X}^T Y$ be a vector of coefficients obtained vai p component-wise regressions where the response Y and each column of the design matrix $\boldsymbol{X}$ have been standardised and de-meaned. Defining d_n to be

a stopping index which may grow with the sample size, for example $d_n = \text{floor}\{n/\log n\}$, the subset of features selected by SIS is obtained via:

$$\widehat{\mathcal{M}} = \{1 \leqslant j \leqslant p : |\omega_j| \text{ is among the } d_n \text{ largest}\}$$

Under reasonably mild assumptions the authors show that with probability approaching one the true set of relevant features $\mathcal{M}_*$ is contained within the chosen subset: $P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}\right) \to 1$ as $n \to \infty$. Fan & Lv coin this the 'sure screening property'. Although SIS as proposed by Fan & Lv (2008) assumes i.i.d. observations, subsequent papers have worked to relax the assumption. In particular, a handful of recent papers propose SIS-like procedures adapted to the time series setting.

Yousuf et al. (2018) first showed that SIS continues to exhibit the sure screening property in a time series setting. Specifically the paper assumes a linear model where random variables are required to be strictly stationary and ergodic time series, and dependence between ordered observations is quantified via the functional dependence measure proposed in Wu (2005). Extending TCS to the time series setting is nonetheless worthwhile, as for reasons explored in Section 2.1 TCS tends to outperform SIS. To better account for the correlation structure of time series data, the paper introduces an efficient extension to SIS in which marginal utilities are estimated via generalised least squares (GLSS). In this dissertation the same extension is applied to tilted correlations in Section 3.2.

Motivated by the observation that when forecasting real time series nonlinear models can often outperform linear models, see for example Teräsvirta et al. (2010), two recent extensions to SIS focus on screening nonlinear time series. Yousuf & Feng (2018) proposes model free screening methods based on ranking distance and partial distance co-variance measures introduced by Székely et al. (2007) and Székely et al. (2014) respectively. The assumptions are similar to those in Yousuf et al. (2018), and both methods have the sure screening property. Meanwhile Chen et al. (2018) suggest a kernel based variant of SIS (KSIS), which assigns marginal utilities by performing p component-wise kernel regressions; the rationale being that kernel regressions are consistent for any specification of the conditional mean as long as the target function is sufficiently smooth.

1.2 Model setup and notation

The dissertation considers the setting where $\{y_t\}_{t=1}^n$ represents a collection of scalar observations from a stationary time series process; d_n lags are potentially relevant, and additionally there are q_n exogenous predictors z_{tj}. In total there are $p_n = q_n + d_n$ potential predictors and the set of predictors is denoted by $\mathcal{J} = \{1, ..., p_n\}$. Both d_n and q_n may diverge as the sample size increases. The predictors are expressed as follows:

$$x_{tj} = \begin{cases} z_{tj} & j = 1, ..., q_n \\ y_{t-(j-q_n)} & j = q_n + 1, ..., q_n + d_n \end{cases}$$

The relationship between the predictor and response variables is assumed to be linear, and is given by $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors. To cope with potential high dimansionality the linear relationship is assumed to be sparse in the sense that $\mathcal{S} = \{j \in \mathcal{J} : \beta_j \neq 0\}$ with $|\mathcal{S}| = s < n$.

For notational convenience $\mathbf{X}$ denotes an $n \times p_n$ matrix where the j-th column is an $n \times 1$ vector of observations for the j-th predictor re-scaled according to $x_{tj}^{\text{scaled}} = \frac{x_{tj} - \overline{x}_j}{\widehat{\sigma}_j}$, with $\overline{x}_j = n^{-1} \sum_t x_{tj}$ and $\widehat{\sigma}_j = \sqrt{n^{-1} \sum_t (x_{tj} - \overline{x}_j)^2}$; entries in Y are scaled accordingly. Such re-scaling is not typically applied to time series as it fundamentally alters the dependence structure of the data - in a sense each observation now contains the whole sample. The re-scaling is only used for variable screening, and the difficulties introduced by re-scaling are dealt with explicitly in Section 3.1. This being said, note that standardised and de-meaned time series have been directly used for forecasting by, among others, Song & Bickel (2011) and De Mol et al. (2008), who simply re-attribute mean and variance to the forecast.

The j-th column of $\mathbf{X}$ is an $n \times 1$ vector denoted by X_j. The sub-matrix of $\mathbf{X}$ whose columns are predictors in the set $\mathcal{D}$ is expressed as $\mathbf{X}_{\mathcal{D}}$. A $p \times 1$ vector of predictors observed at time t and not re-scaled is expressed as $\boldsymbol{x}_t$. Finally, $\mathring{\mathbf{Z}}$ denotes a matrix scaled by $\frac{1}{\sqrt{n}}$. The sample correlation matrix can therefore be written as $\mathbf{C} = \mathring{\mathbf{X}}^T \mathring{\mathbf{X}} = (c_{j,k})_{j,k=1}^{p_n}$.

Chapter 2

Techniques

2.1 Tilted correlation screening

The tilted correlation of Cho & Fryzlewicz (2012) is a measure of association designed to remedy two pitfalls of marginal correlation screening algorithms identified by Fan & Lv (2008), specifically:

1. Irrelevant variables which are highly correlated with relevant ones can have high priority to be selected in marginal correlation screening.
2. A relevant variable can be marginally uncorrelated but jointly correlated with the response.

The following decomposition of the marginal correlation between an arbitrary predictor X_j and the response shows that these pitfalls can be attributed to a 'bias term' which depends on predictors in the set $\mathcal{S}\setminus\{j\}$:

$$\mathring{X}_j^T\mathring{Y} = \mathring{X}_j^T\left(\sum_{k=1}^{p_n}\beta_k\mathring{X}_k + \mathring{\boldsymbol{\epsilon}}\right) = \beta_j + \underbrace{\sum_{k\in\mathcal{S}\setminus\{j\}}\beta_k\mathring{X}_j^T\mathring{X}_k}_{\text{bias}} + \mathring{X}_j^T\mathring{\boldsymbol{\epsilon}}$$

For completeness I note that Fan & Lv address these pitfalls through Iterative Sure Independence Screening (ISIS): a procedure which repeatedly applies SIS to residuals obtained

by regression the response variable on predictors selected by SIS. This however is computationally expensive.

The tilting procedure

Tilted correlations work by transforming each predictor so that the associated bias term is negligible or zero. Since knowledge of the set $\mathcal{S}$ is impossible, the authors note that the bias contribution is largest for those predictors which attain a large sample correlation with X_j. For each predictor, a set of highly correlated predictors is constructed as $\mathcal{C}_j = \left\{k \neq j : \left|\mathring{X}_j^T \mathring{X}_k\right| > \pi_n\right\}$, where π_n effectively acts as a hard threshold on the sample correlation matrix. A de-biased or tilted predictor can then be constructed by projecting onto the space orthogonal to the space spanned by members of the set $\mathcal{C}_j$.

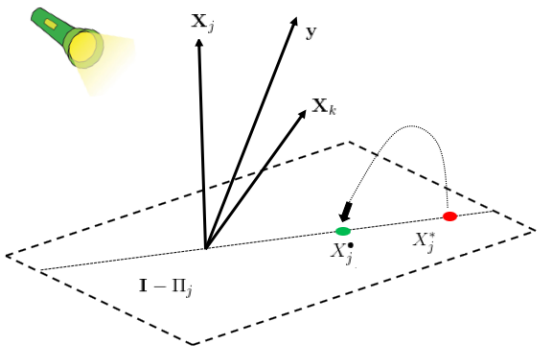

Figure 2.1: The tilting procedure visualised in R^3. With $\mathcal{C}_j = \{k\}$, X_j is projected onto the orthogonal compliment of the space spanned by X_k and re-scaled.

Defining the projection matrix $\Pi_j \equiv \boldsymbol{X}_{\mathcal{C}_j}\left(\boldsymbol{X}_{\mathcal{C}_j}^T \boldsymbol{X}_{\mathcal{C}_j}\right)^{-1} \boldsymbol{X}_{\mathcal{C}_j}^T$, the tilted counterpart of each j-th predictor is $X_j^* \equiv (\boldsymbol{I}_n - \Pi_j) X_j$. Provided the set $\mathcal{C}_j$ is non-empty it holds that $\left\|\mathring{X}_j^*\right\|_2^2 = \mathring{X}_j^T (\boldsymbol{I}_n - \Pi_j) \mathring{X}_j < \mathring{X}_j^T \mathring{X}_j = 1$, hence $\left(\mathring{X}_j^*\right)^T \mathring{Y}$ cannot be used directly as a measure of association. To address this the authors propose two re-scaling methods. The first, on which this dissertation shall focus, is motivated by the decomposition below:

$$\left(\mathring{X}_j^*\right)^T \mathring{Y} = \mathring{X}_j^T \left(\boldsymbol{I}_n - \Pi_j\right) \left\{ \sum_{k=1}^{p_n} \beta_k \mathring{X} + \mathring{\boldsymbol{\epsilon}} \right\}$$
$$= \beta_j \mathring{X}_j^T \left(\boldsymbol{I}_n - \Pi_j\right) \mathring{X}_j + \sum_{k \in \mathcal{S} \setminus \mathcal{C}_j, k \neq j} \beta_k \mathring{X}_j^T \left(\boldsymbol{I}_n - \Pi_j\right) \mathring{X}_k + \mathring{X}_j^T \left(\boldsymbol{I}_n - \Pi_j\right) \mathring{\boldsymbol{\epsilon}}$$

Provided the second and third terms are small, re-scaling by $\mathring{X}_j^T(\boldsymbol{I}_n - \Pi_j)\mathring{X}_j$ isolates β_j, which amounts to isolating the contribution of the j-th predictor to the response. The j-th tilted correlation is therefore defined as:

$$c_j^* = \left(\mathring{X}_j^*\right)^T \mathring{Y} / \left(\mathring{X}_j^T (\boldsymbol{I}_n - \Pi_j) \mathring{X}_j\right)$$

Figure 2.1 gives a graphical representation of the tilting procedure. Note that when $\mathcal{C}_j = \varnothing$ the tilted correlation between a predictor and the response collapses to the marginal correlation. In this way, tilted correlations make an adaptive choice between the traditional Pearson's (straight) correlation and a measure which corrects for the linear contribution of other highly correlated predictors.

Threshold selection procedure

A careful choice of the set $\mathcal{C}_j$ is crucial in high dimensional settings. We would like the set to including all predictors which contribute to the bias term, however if the membership of $\mathcal{C}_j$ is too large when $p \gg n$ any vector in R^n can be closely approximated by a linear combination of predictors in the set; the tilted correlation will therefore fail to capture any relationship between a given predictor and the response. The parameter π_n determines membership of $\mathcal{C}_j$.

Cho & Fryzlewicz propose selecting π_n by performing a hypothesis test of the form $H_0 : |\Sigma_{j,k}| = 0$ on each off-diagonal entry of $\boldsymbol{C}$, the sample correlation matrix of $\boldsymbol{X}$. Their procedure is set out in Algorithm 1 below, and is an adaptation of the procedure in El Karoui et al. (2008) which controls the expected false discovery rate at a pre-set level v^*. This dissertation follows the original paper in setting $v^* = p^{-\frac{1}{2}}$.

Algorithm 1 Threshold selection procedure

1: Set $n \leftarrow \text{nrow}(\boldsymbol{X})$; $p \leftarrow \text{ncol}(\boldsymbol{X})$; $d \leftarrow p(p-1)/2$
2: Generate an $n \times p$ matrix $\boldsymbol{W}$, where each entry is distributed according to $\mathcal{N}(0,1)$
3: Obtain reference sample correlations from $\boldsymbol{W}$ of the form $\{r_{l,m} : 1 \leqslant l < m \leqslant p\}$
4: Assign p-values for each hypothesis test as $P_{j,k} \leftarrow d^{-1} \left|\{r_{l,m} : 1 \leqslant l < m \leqslant p, |r_{l,m}| \geqslant |c_{j,k}|\}\right|$
5: Sort p-values in ascending order $P_{(1)} \leqslant ... \leqslant P_{(d)}$
6: Choose the largest i such that $P_{(i)} \leqslant i/d \cdot v^*$
7: Set π_n to be the absolute value of the sample correlation corresponding to $P_{(i)}$

The separation property

The main theoretical result in Cho & Fryzlewicz (2012) is the 'separation property', which states that under appropriate conditions tilted correlations can separate between relevant and irrelevant predictors. More formally $P(\Delta) \to 0$ where the event Δ is defined as:

$$\Delta = \left\{ \frac{|c_k^*|}{\min_{j \in \mathcal{S}} |c_j^*|} \to 0 \text{ for all } k \notin \mathcal{S} \right\}$$

The authors employ the six assumptions listed below when studying the theoretical properties of tilted correlations. In what follows a deterministic sequence $\{a_n\}_{n=1}^{\infty}$ is said to be $\mathcal{O}(b_n)$, where $\{b_n\}_{n=1}^{\infty}$ is another deterministic sequence, if $\lim_{n\to\infty} \left|\frac{a_n}{b_n}\right| = c$ for some positive constant c.

A (1) *The number of non-zero coefficients satisfies* $|\mathcal{S}| = \mathcal{O}\left(n^{\delta}\right)$ *for* $\delta \in [0, 1/2)$.

A (2) *The number of predictors satisfies* $\log(p_n) = \mathcal{O}\left(n^{\theta}\right)$ *for* $\theta \in [0, 1-2\gamma)$ *with* $\gamma \in (\delta, 1/2)$.

A (3) *The threshold for membership of each set* $\mathcal{C}_j$ *satisfies* $\pi_n = \mathcal{O}(n^{-\gamma})$, *and the cardinality of each* $\mathcal{C}_j$ *satisfies* $|\mathcal{C}_j| = \mathcal{O}\left(n^{\xi}\right)$ *for* $\xi \in [0, 2(\gamma - \delta))$.

A (4) *The largest non-zero coefficient satisfies* $\max_{1 \leqslant j \leqslant p_n} |\beta_j| < M$ *for* $M \in (0, \infty)$.

A (5) *For each* $j \in \mathcal{J}$ *it holds that* $1 - \mathring{X}_j^T \Pi_j \mathring{X}_j = 1 - a_j > a$ *for* $a \in (0,1)$.

A (6) *For those* j *whose* $\mathcal{C}_j$ *satisfies* $\mathcal{S} \nsubseteq \mathcal{C}_j$, *for some* κ *satisfying* $\kappa/2 + \mu \in [0, \gamma - \delta - \xi/2)$ *it holds that* $n^{\kappa} \cdot \frac{\|(\boldsymbol{I}_n - \Pi_j)\boldsymbol{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2}{\|\boldsymbol{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2} \to \infty$

A(1) and A(2) describe how the sparsity $|\mathcal{S}|$ and dimensionality p_n of the linear model may grow with the sample size n. A(3) requires that the number of predictors in each $\mathcal{C}_j$ does not

exceed a certain polynomial rate in n, which is needed in order to guarantee the existence of each projection matrix Π_j. Finally, A(5) rules out strong collinearity among predictors, which agan guarantees the existance of each projection matrix. For a full discussion of the assumptions see *Section 2.3* of the original paper.

Tilted correlation screening algorithm

Algorithm 2 below sets out the TCS algorithm as described in *Section 3.1* of the original paper. The algorithm exploits the theoretical properties of tilted correlations, and selects a set of predictors of size $m < n$ which contains the true set $\mathcal{S}$ with high probability. The algorithm generates a solution path $\mathcal{A}_{(1)} \subset ... \subset \mathcal{A}_{(m)}$, hence the final model $\widehat{\mathcal{S}}$ is obtained either by selecting a model from the solution path or by selecting a subset of predictors from the final active set.

Algorithm 2 Tilted Correlation Screening

1: Begin with an empty active set $\mathcal{A} = \varnothing$, current residual equal to the response variable $W = Y$, and current design matrix $\boldsymbol{Z} = \boldsymbol{X}$. From $\boldsymbol{X}$ obtain a correlation threshold π_n.
2: Find the variable which achieves the maximal marginal correlation with W, and let $k = \arg\max_{j \notin \mathcal{A}} \left| Z_j^T W \right|$. Identify the set of $\mathcal{C}_k = \left\{ j \notin \mathcal{A}, j \neq k : \left| Z_j^T Z_k \right| > \pi_n \right\}$. If $\mathcal{C}_j = \varnothing$ set $k^* = k$ and go to Step 4.
3: If $\mathcal{C}_k \neq \varnothing$ screen the tilted correlation c_j^* between Z_j and W for each $j \in \{\{k\} \cup \mathcal{C}_k\}$. Set $k^* = \arg - \max_{j \in \{\{k\} \cup \mathcal{C}_k\}} \left| c_j^* \right|$.
4: Add k^* to the active set $\mathcal{A} \leftarrow \mathcal{A} \cup \{k^*\}$, and update the current residual $W \leftarrow (\boldsymbol{I}_n - \Pi_{\mathcal{A}}) Y$ and current design matrix $\boldsymbol{Z} \leftarrow (\boldsymbol{I}_n - \Pi_{\mathcal{A}}) \boldsymbol{X}$. Re-scale each column $j \notin \mathcal{A}$ to have l_2-norm 1.
5: Repeat Steps 2-4 until the cardinality of the active set reaches $|\mathcal{A}|$ a pre-specified value $m < n$.

2.2 Measures of Dependence

Beginning around the time of Rosenblatt (1956) statisticians realised the usefulness of measure of dependence for time series that do not fit any specific dependence structure but have some asymptotic notion of independence. To make the results in this dissertation as general as possible the same approach is used. Mixing conditions are one such class of measure which capture the idea that dependence between observations should decay as they get further apart.

2.2.1 Strong Mixing

Throughout the dissertation α-mixing (henceforth strong mixing) is used, however for completeness I introduce four of the most common mixing conditions. Let $\{Z_t, t \in Z\}$ be an ordered sequence of random variables equipped with a probability triple $\left(\Omega, \mathcal{F}_{-\infty}^{\infty}, P\right)$, the common mixing conditions are the following:

$$\alpha(n) := \sup_{j \in Z} \left\{ \sup_{\mathcal{A} \in \mathcal{F}_{-\infty}^{j}, B \in \mathcal{F}_{j+n}^{\infty}} \left| P\left(\mathcal{A} \cap \mathcal{B}\right) - P\left(\mathcal{A}\right) P\left(\mathcal{B}\right) \right| \right\}$$

$$\beta(n) = \sup_{\mathcal{A}, B} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \left| P\left(\mathcal{A}_i \cap \mathcal{B}_j\right) - P\left(\mathcal{A}_i\right) P\left(\mathcal{B}_j\right) \right|$$

$$\phi(n) := \sup_{j \in Z} \left\{ \sup_{\mathcal{A} \in \mathcal{F}_{-\infty}^{j}, B \in \mathcal{F}_{j+n}^{\infty}} \left| P\left(\mathcal{B} | \mathcal{A}\right) - P\left(\mathcal{B}\right) \right| \right\}$$

$$\psi(n) := \sup_{j \in Z} \left\{ \sup_{\mathcal{A} \in \mathcal{F}_{-\infty}^{j}, B \in \mathcal{F}_{j+n}^{\infty}} \left| \frac{P\left(\mathcal{A} \cap \mathcal{B}\right)}{P\left(\mathcal{A}\right) P\left(\mathcal{B}\right)} - 1 \right| \right\}$$

If $\alpha(n) \to 0$ as $n \to \infty$ the process generating $\{Z_t, t \in Z\}$ is said to be strong mixing. Clearly when $\{Z_t, t \in Z\}$ is a sequence of i.i.d. random variables all four measures are exactly zero for any $n > 0$ - i.i.d. random variables are simply a special case of mixing sequences. Indeed, as long as the mixing coefficient decays to zero sufficiently fast many properties of i.i.d. processes, in particular laws of large numbers and central limit theorems, can be established for mixing processes. For a detailed survey of mixing conditions and their properties see Bradley et al. (2005). The following inequality is well known:

$$2\alpha(n) \leqslant \beta(n) \leqslant \phi(n) \leqslant \frac{1}{2}\psi(n)$$

.

Strong mixing was chosen because it is implied by the other mixing conditions, and is therefore the easiest to satisfy. Indeed, a large class of commonly studied time series processes have been shown to be strong mixing. For example ergodic Markov chains are

strong mixing, stationary ARMA processes are geometrically strong mixing - see Mokkadem (1988), and even ARCH(∞) processes have been shown to be strong mixing with mixing coefficient depending on the rate of decay of the ARCH parameters - see Fryzlewicz et al. (2011).

2.2.2 Inequalities for strong mixing processes

Here I present some inequalities for strong mixing processes from Bosq (2012) which will be of use throughout Chapter 3. In what follows $\{Z_t, t \in Z\}$ is a zero mean strong mixing process with $S_n = Z_1 + ... + Z_n$ being a partial sum, and X and Y are real valued random variables.

B (1) *Rio's inequality. Let* $Q_X(u) = \inf\{t : P(|X| > t) \leqslant u\}$ *be the quantile function for* $|X|$ *and define* $\alpha = \alpha(\sigma(X), \sigma(Y))$*, if* $Q_X Q_Y$ *is integrable over* $(0,1)$ *it holds that:*

$$|Cov(X,Y)| \leqslant 2\int_0^{2\alpha} Q_X(u)Q_Y(u)du$$

B (2) *If* $\{Z_t, t \in Z\}$ *is bounded, i.e.* $P(\sup_t |Z_t| \leqslant b) = 1$ *for some* $b \in (0,\infty)$*, for each integer* $q \in \left[1, \frac{n}{2}\right]$ *and for any* $\epsilon > 0$ *it holds that:*

$$P(|S_n| > n\epsilon) \leqslant 4\exp\left(-\frac{\epsilon^2}{8b^2}q\right) + 22\left(1 + \frac{4b}{\epsilon}\right)^{\frac{1}{2}} q\alpha\left(\left[\frac{n}{2q}\right]\right)$$

B (3) *If the process satisfies the Cramér condition, i.e. there exists some constant* $c > 0$ *such that for all* $t \in R$ *and for integer each* $k \geqslant 3$ *it holds that* $E|Z_t|^k \leqslant c^{k-2}k!E\left(Z_t^2\right)$*, with the same* q *and* ϵ *as B(2) it holds that:*

$$P(|S_n| > n\epsilon) \leqslant a_1\exp\left(-\frac{q\epsilon^2}{25m_2^2 + 5c\epsilon}\right) + a_2(k)\alpha\left(\left[\frac{n}{q+1}\right]\right)^{\frac{2k}{2k+1}}$$

where a_1 *and* $a_2(k)$ *are defined as follows:*

$$a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\epsilon^2}{25m_2^2 + 5c\epsilon}\right) \quad \text{with } m_2^2 = \max_{1\leqslant t\leqslant n} E(Z_t^2)$$

$$a_2(k) = 11n\left(1 + \frac{5m_k^{\frac{k}{2k+1}}}{\epsilon}\right) \quad with\ m_k = \max_{1\leqslant t\leqslant n} \|Z_t\|_k$$

Exponential concentration inequalities such as B(2) and B(3) have become a mainstay of the high dimensional literature as they allow statisticians to derive bounds which tighten with n without explicitly accounting for the behaviour of p_n; see for example chapter 2 of Wainwright (2019). In particular Chen et al. (2018) use B(3) to prove that KSIS has the sure screening property.

2.3 Bootstrapping

The bootstrap of Efron (1979) is a widely used re-sampling technique which generates pseudo-samples $\boldsymbol{Z}^* = \{Z_1^*, ..., Z_n^*\}$ from an observed sample $\boldsymbol{Z} = \{Z_1, ..., Z_n\}$ by sampling with replacement from the empirical distribution. Recent papers, for example Meinshausen & Bühlmann (2010) and Yuen (2019), have successfully applied re-sampling to the problem of estimating 'stable' models for high dimensional data by performing variable selection on the basis of inclusion frequencies. In this section I briefly review extensions of the bootstrap whose pseudo-samples retain salient features of the original sample, and are therefore well suited to the high dimensional time series setting. These ideas will be useful in Section 3.3.

2.3.1 Bootstrapping stationary time series

Time series are characterised by a natural temporal ordering, as well as some structured dependence between observations. Efron's bootstrap is therefore poorly suited to the time series setting, as sampling randomly from the empirical distribution will destroy the dependence structure of the original sample. Kunsch (1989) suggest breaking the original sample into k blocks of size b_n, i.e. $\boldsymbol{Z} = \{\boldsymbol{B}_{1,b_n}, ..., \boldsymbol{B}_{n-b_n+1,b_n}\}$, and instead sampling blocks of observations. Each block is defined as: $\boldsymbol{B}_{i,b_n} = \{Z_i, Z_{i+1}, ..., Z_{i+b_n-1}\}$

If b_n is chosen so that $n = kb_n$ for some integer k with $b_n \to \infty$ and $b_n/n \to 0$ as $n \to \infty$ then each block can be treated as approximately independent from every other block.

Importantly, since within each block observations are grouped according to their natural order the 'block bootstrap' preserves the dependence structure of the original sample.

If $\boldsymbol{Z}$ is a sample from a strictly stationary time series then by construction $\boldsymbol{Z}^* = \{\boldsymbol{B}_1^*, ..., \boldsymbol{B}_k^*\}$ will be non-stationary. This is because the pseudo-sample will contain a structural break every b_n observations. To remedy this, Politis & Romano (1994) propose the 'stationary bootstrap'. Let $\{L_1, L_2, ...\}$ be i.i.d. with geometric distribution such that $P(L_1 = t) = r(1-r)^{t-1}$ for some positive integer r, and let $\{I_1, I_2, ...\}$ be i.i.d. with discrete uniform distribution on $\{1, ..., n\}$. A 'stationary bootstrap' sample is generated by setting $\boldsymbol{Z}^* = \{\boldsymbol{B}_{I_1,L_1}, \boldsymbol{B}_{I_2,L_2}, ...\}$ until the desired length is reached. The dependence structure in the original sample is preserved by the same reasoning as in the 'block bootstrap'. Additionally, structural breaks are now random so $\boldsymbol{Z}^*$ will be strictly stationary given $\boldsymbol{Z}$.

For completeness I note that there exist a large number of parametric bootstrap schemes for time series data; for a concise overview see *Chapter 9* of Shao & Tu (2012). These however assume that the form of the model generating $\boldsymbol{Z}$ is known explicitly. By the desire for generality expressed in Section 2.2 such methods are not explored.

2.3.2 Bootstrapping high dimensional data

As the limit of p_n/n grows $\boldsymbol{Z}^*$'s ability to faithfully replicate meaningful features of $\boldsymbol{Z}$ deteriorates. This is clearest from the fact that the empirical distribution of the bootstrap weights, i.e. the number of times any given Z_t is included in the bootstrap sample, has asymptotic Poisson(1) distribution; for a proof see *Proposition 4.10* in El Karoui et al. (2010). A direct consequence is that the expected number of unique Z_ts in any bootstrap sample is $n\left(1 - \frac{1}{\exp(1)}\right) \approx 0.63n$. As observed by Karoui & Purdom (2016), if $p_n/n \gg 0.63$ then each $\boldsymbol{Z}^*$ will be asymptotically rank deficient even if $\boldsymbol{Z}$ has full column rank. This problem is clearly exacerbated by the 'block bootstrap' and its variants.

Chapter 3

Extensions

3.1 Proving the separation property

The first step in applying the TCS algorithm to time series data is to show that the separation property continues to hold in the time series setting. In Cho & Fryzlewicz (2012) the separation property is proven by observing that the sample correlation between the response y_t any predictor x_{tj} projected onto the space orthogonal to the space spanned by predictors in $\mathcal{C}_j$ can be decomposed as follows:

$$\begin{aligned}
(\mathring{X}_j^*)^T\mathring{Y} &= \mathring{X}_j^T(\mathbf{I}-\Pi_j)\mathring{Y} \\
&= \mathring{X}_j^T\left\{\sum_{k=1}^{p}\beta_k(\mathbf{I}-\Pi_j)\mathring{X}_k + (\mathbf{I}-\Pi_j)\mathring{\boldsymbol{\epsilon}}\right\} \\
&= \beta_j\mathring{X}_j^T(\mathbf{I}-\Pi_j)\mathring{X}_j + \underbrace{\sum_{k\in\mathcal{S}\setminus\mathcal{C}_j,k\neq j}\beta_k\mathring{X}_j^T(\mathbf{I}-\Pi_j)\mathring{X}_k}_{1} + \underbrace{\mathring{X}_j^T(\mathbf{I}-\Pi_j)\mathring{\boldsymbol{\epsilon}}}_{2}
\end{aligned} \tag{3.1}$$

Terms 1 and 2 are shown to be negligibly small, and after appropriate re-scaling the remaining term does not depend on the sample size and is exactly zero for irrelevant predictors. The original proof relies on a well known concentration inequality for Gaussian random variables. I present a similar proof which instead uses the exponential inequalities intro-

duced in Section 2.2.2 and is therefore valid for a large class of time series processes. The proof uses assumptions A(1)-A(6) from the original paper, however to adapt to the time series setting three additional assumptions are required.

Additional assumptions

A (7) *The process* $\{(y_t, \boldsymbol{x}_t, \epsilon_t), t \in Z\}$ *is a strictly stationary* α*-mixing process with mixing coefficient having size* $\alpha(n) = \mathcal{O}(\alpha^n)$ *for* $\alpha \in (0, 1)$.

A (8) *For all* $t \in \mathbf{Z}$ *and all* $j \in \mathcal{J}$ *it holds that* $x_{tj} \perp\!\!\!\perp \epsilon_t$.

A (9) *Letting* $Z_t = x_{tj} \cdot \epsilon_t$ *for any* $j \in \mathcal{J}$ *it holds that* $EZ_t = 0$ *for all* t *and either: (i)* Z_t *is bounded as in B(2), or (ii)* Z_t *satisfies the Cramér condition as in B(3).*

A(7) is a standard assumption in the econometric literature: it allows us to quantify the degree of dependence in the data while being agnostic towards the data generating process itself. Specifically, A(7) restricts the degree of dependence to that of a geometrically strong mixing process. The same assumption is made in all of the screening methods presented in Chen et al. (2018). A(8) imposes independence between the model errors and each predictor, and allows us to split the expectation of a product into the product of expectations when using the Markov inequality in the proofs below; it may be seen as somewhat restrictive, as in general it is sufficient to assume that the errors and predictors are contemporaneously uncorrelated. In the high dimensional setting however it is not unreasonable to assume that the set of predictors is sufficiently rich to justify complete independence. Finally, A(9) is necessary in order to make use of the exponential inequalities introduced in Section 2.2.2. In the unbounded case it implies all moments of the errors and predictors are finite.

3.1.1 A proof for scenario 1

Scenario 1 in the original paper refers to the setting in which whenever X_j is projected onto the space spanned by predictors in the associated set $\mathcal{C}_j$ any relevant predictors not in $\mathcal{C}_j$ remain remain 'far away' from the projection. The setting is expressed formally as:

C (1) $\left|\left(\Pi_j \mathring{X}_j\right)^T \mathring{X}_k\right| = \mathcal{O}\left(n^{-\gamma}\right)$ *for all* $j \in \mathcal{J}$ *and* $k \in \mathcal{S} \backslash \{\mathcal{C}_j \cup \{j\}\}$.

In a time series setting, whenever C(1) holds the tilted correlations of relevant predictors dominate those of irrelevant predictors.

Theorem 1 *Under assumptions A(1)-A(9) and the additional condition C(1) it holds that* $P(\Delta) \to 1$, *where the event* Δ *is defined as:*

$$\Delta = \left\{ \frac{|c_k^*|}{\min_{j \in \mathcal{S}} |c_j^*|} \to 0 \text{ for all } k \notin \mathcal{S} \right\}$$

Proof:

For those j whose $\mathcal{C}_j$ satisfy $\mathcal{S} \backslash \{j\} \subseteq \mathcal{C}_j$ it follows that $\mathcal{S} \backslash \{\mathcal{C}_j \cup \{j\}\} = \varnothing$ and term 1 will be exactly zero. When $\mathcal{S} \backslash \{j\} \nsubseteq \mathcal{C}_j$ term 1 typically will not be zero, however we can bound:

$$\left| \mathring{X}_j \left(\mathbf{I} - \Pi_j \right) \mathring{X}_k \right| \leqslant \left| \mathring{X}_j^T \mathring{X}_k \right| + \left| \left(\Pi_j \mathring{X}_j \right)^T \mathring{X}_k \right| \leqslant C n^{-\gamma}$$

Where the second inequality comes from the fact that A(3) guarantees that for each $k \notin \mathcal{C}_j$ $\left| \mathring{X}_j^T \mathring{X}_k \right|$ will be smaller that π_n, which in turn is bounded from above by $Cn^{-\gamma}$. Hence, term 1 can be bounded as follows:

$$\left| \sum_{k \in \mathcal{C} \backslash \mathcal{C}_j, k \neq j} \beta_k \mathring{X}_j^T (\mathbf{I} - \Pi_j) \mathring{X}_k \right| \leqslant C \left| \mathcal{S} \right| n^{-\gamma} \leqslant C n^{-(\gamma - \delta)}$$

Which holds because A(1) guarantees that $|\mathcal{S}| \leqslant C n^{\delta}$ and A(4) guarantees that the largest $|\beta_j|$ is bounded away from infinity. It remains to show that with probability approaching one term 2 does not exceed the upper bound established for term 1. Since $\mathring{X}_j^T \mathring{\boldsymbol{\epsilon}}$ is proportional to $\mathring{X}_j^T \Pi_j \mathring{\boldsymbol{\epsilon}}$ by assumptions A(3) and A(5), it is sufficient to consider the event $\Delta_1 = \left\{ \left| \mathring{X}_j^T \mathring{\boldsymbol{\epsilon}} \right| > C n^{-(\gamma - \delta)} \right\}$. However, as suggested in Section 1.2 strong mixing inequalities cannot be applied directly to $\left\langle \mathring{X}_j, \mathring{\boldsymbol{\epsilon}} \right\rangle$ since having been standardised and de-meaned the entries in $\mathring{X}_j$ are no longer strong mixing. The approach will be to upper bound $P(\Delta_1)$ by the probability of three separate events and deal with each in turn:

$$
\begin{aligned}
P(\Delta_1) &\leqslant P\left(\frac{1}{n}\left|\sum_{t=1}^{n}(x_{tj}-\overline{x}_j)\,\epsilon_t\right| > \frac{Cn^{-(\gamma-\delta)}}{M_{\sigma_j^2}}\right) + P\left(\widehat{\sigma}_j^{-1} > M_{\sigma_j^2}\right) \\
&\leqslant P\left(\frac{1}{n}\left|\sum_{t=1}^{n}(\overline{x}_j-\mu_j)\,\epsilon_t\right| > \frac{Cn^{-(\gamma-\delta)}}{2M_{\sigma_j^2}}\right) + P\left(\frac{1}{n}\left|\sum_{t=1}^{n}(x_{tj}-\mu_j)\,\epsilon_t\right| > \frac{Cn^{-(\gamma-\delta)}}{2M_{\sigma_j^2}}\right) + P\left(\widehat{\sigma}_j^{-1} > M_{\sigma_j^2}\right) \\
&= P(\Delta_{1,1}) + P(\Delta_{1,2}) + P(\Delta_{1,3})
\end{aligned}
$$

Setting $M_{\sigma_j^2} = 1/\sqrt{E(x_{tj}-\mu_j)^2 - \varepsilon}$ for some small constant $\varepsilon > 0$ imediately gives $P(\Delta_{1,3}) \to 0$ as $n \to \infty$ by standard law of large numbers for geometrically strong mixing processes. $P(\Delta_{1,1})$ can be dealt with using the familiar Markov inequality as shown below.

$$
\begin{aligned}
P(\Delta_{1,1}) &= P\left(\frac{1}{n^2}\left|\sum_{t=1}^{n}\sum_{s=1}^{n}(x_{sj}-\mu_j)\,\epsilon_t\right| > Cn^{-(\gamma-\delta)}\right) \\
&\leqslant \left(\frac{n^{2(\gamma-\delta)}}{C^2n^4}\right) E\left(\sum_{t=1}^{n}\sum_{s=1}^{n}(x_{sj}-\mu_j)\,\epsilon_t\right)^2 \\
&= \left(\frac{n^{2(\gamma-\delta)}}{C^2n^2}\right)\left(\underbrace{\frac{1}{n}\sum_{t,s=1}^{n}E(x_{tj}-\mu_j)(x_{ts}-\mu_j)}_{3} \times \underbrace{\frac{1}{n}\sum_{k,l=1}^{n}E(\epsilon_k\epsilon_l)}_{4}\right) \\
&= \mathcal{O}\left(n^{-2(1-(\gamma-\delta))}\right)
\end{aligned}
$$

The final equality relies on the fact that both terms 3 and 4 are $\mathcal{O}(1)$. Why? Notice first that terms 3 and 4 are equivalent to summing over all the entries of the $n \times n$ Toeplitz matrix $n^{-1}\Gamma(n)$ shown below. Using Rio's inequality introduced in B(1) and choosing $r > 2$, $X = \epsilon_t$, and $Y = \epsilon_{t+k}$ the k-th auto-covariance for the process $\{\epsilon_t, t \in Z\}$ can be upper bounded as follows $|\gamma_k| = |\mathrm{Cov}(\epsilon_t, \epsilon_{t+k})| \leqslant \frac{2r}{r-2}(2\alpha(k))^{1-\frac{2}{r}}(E|\epsilon_0|^r)^{\frac{2}{r}} \leqslant c \cdot \alpha^k$. The same applies to the auto-covariance sequence for $\{(x_{tj}-\mu_j), t \in Z\}$. Since $\alpha \in (0,1)$ each row of the Toeplitz matrix is $\mathcal{O}(n^{-1})$, so terms 3 and 4 must be $\mathcal{O}(1)$.

$$n^{-1}\Gamma(n) = \begin{pmatrix} \gamma_0/n & \gamma_1/n & \cdots & \gamma_{n-1}/n \\ \gamma_1/n & \gamma_0/n & \cdots & \gamma_{n-2}/n \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1}/n & \gamma_{n-2}/n & \cdots & \gamma_0/n \end{pmatrix}$$

Meanwhile, strong mixing inequalities can be applied directly to $P(\Delta_{1,2})$. When x_{tj} and ϵ_t are bounded the proof can be completed using B(1). Since $(\gamma - \delta)$ lies in the open interval $(0, 1/2)$ for sufficiently large values of n it will always be possible to find some $\lambda \in (0,1)$ which satisfies $\lambda > 2(\gamma - \delta)$ and floor $\left\{\frac{n^\lambda}{\lambda}\right\} \in \left[1, \frac{n}{2}\right]$. Hence using inequality B(1) and setting $\epsilon = Cn^{-(\gamma-\delta)}$ and $q = \frac{n^\lambda}{\lambda}$ we have:

$$\begin{aligned} P(\Delta_{1,2}) &= P\left(\left|\sum_{t=1}^{n} x_{tj}\epsilon_t\right| > n\left(Cn^{-(\gamma-\delta)}\right)\right) \\ &\leqslant \left\{4\exp\left(-\left(\frac{C^2}{8\lambda b^2}\right)n^{\lambda-2(\gamma-\delta)}\right) + 22\sqrt{\left(1+\frac{4bn^{(\gamma-\delta)}}{C}\right)\frac{n^{2\lambda}}{\lambda^2}}\cdot\exp\left(-\left(\frac{c_\alpha}{2\lambda}\right)n^{1-\lambda}\right)\right\} \end{aligned}$$

The last exponential follows from a change of base for the mixing coefficient, hence c_α is a positive constant that does not depend on the sample size; combined with the conditions on λ, δ, and γ this guarantees that $P(\Delta_{1,2}) \to 0$. When either of x_{tj} or ϵ_t is unbounded the proof can be completed with inequality B(2) as long as $x_{tj} \cdot \epsilon_t$ satisfies the Cramér condition. Setting ϵ and λ as before it therefore holds that:

$$\begin{aligned} P(\Delta_{1,2}) &= P\left(\left|\sum_{t=1}^{n} x_{tj}\epsilon_t\right| > n\left(Cn^{-(\gamma-\delta)}\right)\right) \\ &\leqslant \left(\frac{2}{\lambda}n^{1-\lambda} + 2\left(1+\frac{C^2n^{-(\gamma-\delta)}}{(25m_2^2)n^{(\gamma-\delta)}+5C'}\right)\right)\times\exp\left(-\frac{C^2n^{\lambda-2(\gamma-\delta)}}{\lambda\left(25m_2^2+5C'n^{-(\gamma-\delta)}\right)}\right) \\ &+ \left(11n\left(\left(\frac{5m_k^{\frac{k}{2k+1}}}{C}\right)n^{(\gamma-\delta)}\right)\right)\times\exp\left(-c_\alpha\left(\frac{\lambda n^{1-\lambda}}{1+\lambda n^{-\lambda}}\right)\right) \end{aligned}$$

Which guarantees that $P(\Delta_{1,2}) \to 0$ by the same reasoning as was used in the bounded

case.

3.1.2 A proof for scenario 2

Let $\mathcal{K} \subset \mathcal{J}$ denote a subset of predictors which are either relevant ($k \in \mathcal{S}$) or highly correlated with at least one relevant predictor $\left(k \in \bigcup_{j \in \mathcal{S}} \mathcal{C}_j\right)$. *Scenario 2* in the original paper refers to the setting in which for each relevant predictor X_j, if X_k is a both a predictor in $\mathcal{K}$ and not highly correlated with X_j, then there does not exist a predictor X_l with $l \neq j, k$ which attains a sample correlation greater than π_n with both X_j and X_k simmultaneously. The setting is expressed formally as:

C (2) *For each $j \in \mathcal{S}$, if $k \in \mathcal{K} \backslash \{\mathcal{C}_j \cup \{j\}\}$ then $\mathcal{C}_j \cap \mathcal{C}_k = \varnothing$.*

In a time series setting, whenever C(2) holds the tilted correlations of relevant predictors dominate those of irrelevant predictors.

Theorem 2 *Under assumptions A(1)-A(9) and the additional condition C(2) it holds that $P(\Delta) \to 1$, where the event Δ is defined as:*

$$\Delta = \left\{ \frac{|c_k^*|}{\min_{j \in \mathcal{S}} |c_j^*|} \to 0 \text{ for all } k \notin \mathcal{S} \right\}$$

Proof:

From A(3) it holds that $\xi < 2(\gamma - \delta) \Leftrightarrow \delta < \gamma - \xi/2$, hence for each $\mathring{X}_k$ satisfying A(3) we have that $\left|\mathring{X}_j^T \Pi_j \mathring{X}_k\right| \leqslant C n^{-(\gamma - \xi/2)}$. Hence, term 1 in (3.1) can be bounded as follows:

$$\begin{aligned}
\left| \sum_{k \in \mathcal{C} \backslash \mathcal{C}_j, k \neq j} \beta_k \mathring{X}_j^T (\mathbf{I} - \Pi_j) \mathring{X}_k \right| &\leqslant \sum_{k \in \mathcal{C} \backslash \mathcal{C}_j, k \neq j} \beta_k \left\{ \left| \mathring{X}_j^T \mathring{X}_k \right| + \left| (\Pi_j \mathring{X}_j)^T \Pi_j \mathring{X}_k \right| \right\} \\
&\leqslant C n^{-(\gamma - \delta)} + C' n^{-(\gamma - \delta - \xi/2)} \\
&= \mathcal{O}(n^{-(\gamma - \delta - \xi/2)})
\end{aligned}$$

It remains to show that term 2 in (3.1) does not exceed new the bound given for term 1. Define Δ_2 to be the event $\left\{\left|\mathring{X}_j^T\mathring{\boldsymbol{\epsilon}}\right| > Cn^{-(\gamma-\delta-\xi/2)}\right\}$. Using the same approach as in 3.1.1 we have that $P(\Delta_2) \leqslant P(\Delta_{2,1}) + P(\Delta_{2,2}) + P(\Delta_{2,3})$ with $P(\Delta_{2,1}) = \mathcal{O}\left(n^{-2(1-\gamma-\delta-\xi/2)}\right)$ and $P(\Delta_{2,3}) \to 0$. Finally, $P(\Delta_{2,2})$ can now be bounded with either inequality B(1) or B(2); here I use B(1) for brevity. Using the fact that $(\gamma-\delta-\xi/2)$ lies in the open interval $(0,1/2)$ it is again possible to find some $\lambda \in (0,1)$ satisfying both $\lambda > 2(\gamma-\delta-\xi/2)$ and $\text{floor}\left\{\frac{n^\lambda}{\lambda}\right\} \in \left[1, \frac{n}{2}\right]$:

$$
\begin{aligned}
P(\Delta_{2,2}) &= P\left(\left|\sum_{t=1}^{n} x_{tj}\epsilon_t\right| > n\left(Cn^{-(\gamma-\delta-\xi/2)}\right)\right) \\
&\leqslant \left\{4\exp\left(-\left(\frac{C^2}{8\lambda b^2}\right)n^{\lambda-2(\gamma-\delta-\xi/2)}\right) + 22\sqrt{\left(1+\frac{4bn^{(\gamma-\delta-\xi/2)}}{C}\right)\frac{n^{2\lambda}}{\lambda^2}}\cdot\exp\left(-\left(\frac{c_\alpha}{2\lambda}\right)n^{1-\lambda}\right)\right\}
\end{aligned}
$$

Which guarantees that $P(\Delta_{2,2}) \to 0$ by the same reasoning as used in 3.1.1.

3.1.3 Simulation study

To illustrate the separation property I present the following small simulation. The idea is to show the superior screening property of tilted correlations as compared to straight correlations in low dimensions ($p = 50, n = 500$), moderate dimensions ($p = n = 500$), and high dimensions ($p = 1000, n = 500$).

Simulation Setup

Predictors were generated according to a VAR(1) process given by $\mathbf{x}_t = \Phi\mathbf{x}_{t-1} + \boldsymbol{\nu}_t$ with $\Phi = diag(0.4)$ and $\boldsymbol{\nu}_t \sim_{i.i.d} \mathcal{N}_p(\mathbf{0}, \Sigma)$ where $\Sigma = \left\{0.8^{|i-j|}\right\}_{i,j\leqslant p}$. The response variable was generated according to $y_t = \boldsymbol{\beta}^T\mathbf{x}_t + e_t$ where e_t followed an AR(1) process with parameter $\phi = 0.6$ and standard Normal innovations. Five predictors spaced maximally apart were chosen to be relevant and their entry in $\boldsymbol{\beta}$ was set to either -1 or 1. All other entries in $\boldsymbol{\beta}$ were set to zero. This setup is particularly challenging because:

1. Setting $\Sigma = \left\{0.8^{|i-j|}\right\}_{i,j\leqslant p}$ means neighbouring predictors are very highly correlated.

2. Maximal spacing ensures the largest possible number of irrelevant predictors are highly correlated with relevant predictors.

Simulation results

The predictor and response variables were first generated for the stated dimensions, then tilted and straight correlations were estimated on sub-samples of size $n = 100, 110, ..., 500$. Figure 3.1 shows plots of the simulation results. In low and medium dimension the separation property is clear: many irrelevant predictors attain straight correlations with the response in a neighbourhood of those attained by relevant predictors, however when tilted correlations are used relevant predictors are easy to distinguish. In the high dimensional setting the separation property is less clear visually. However, observing plots (c) and (f) note that with $p = 1000$ and $n = 500$ straight correlations of three out of the five relevant predictors are dominated by those of irrelevant predictors. Meanwhile, when tilted correlations are used the five relevant predictors are also the five predictors most highly correlated with the response.

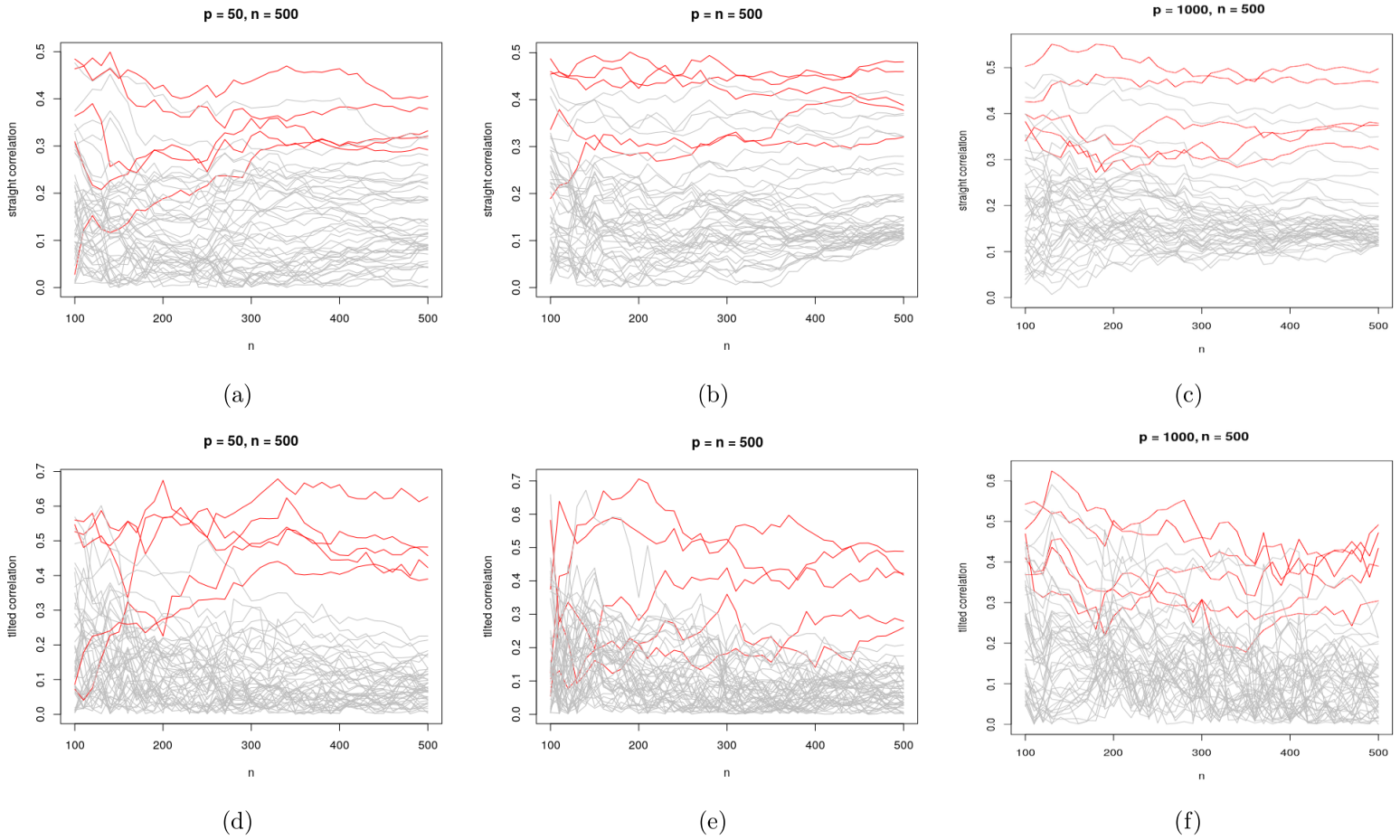

Figure 3.1: Plots of absolute tilted and straight correlations for relevant predictors (red) along with 45 irrelevant predictors most correlated with the response (grey) against sample size. Plots (a)-(c) show straight correlations while (d)-(f) show tilted correlations.

3.2 Efficient time series tilting

The tilted correlation between the response variable and the j-th predictor is equivalent to the ordinary least squares estimate β_j when regressing y_t onto $\boldsymbol{x}_{t,\mathcal{D}_j}$ where $\mathcal{D}_j$ is the set $\{\{j\} \cup \mathcal{C}_j\}$. This implies a marginal regression model:

$$Y = \mathbf{X}_{\mathcal{D}_j}\boldsymbol{\beta}_{\mathcal{D}_j} + \boldsymbol{u}_{\mathcal{D}_j}$$

Following the intuition from *Section 4* of Yousuf et al. (2018) observe that the error term in the marginal model is likely to be auto-correlated even if the error term in the full model is independently distributed. In this setting, by the main result in Aitkin (1935), a more efficient estimator in the MSE sense can be obtained by accounting for the serial correlation in the errors $\left\{\left(u_{t,\mathcal{D}_j}\right) t \in Z\right\}$.

An in-feasible estimator

Let Σ_j be the $n \times n$ auto-covariance matrix for the unobservable errors $u_{1,\mathcal{D}_j}, ..., u_{n,\mathcal{D}_j}$. Then a vector of weighted predictors can be defined as $\widetilde{X}_j = \Sigma_j^{-\frac{1}{2}} X_j$, and an efficient estimator for the j-th tilted correlation is as follows:

$$\widetilde{c_j^*} = \left(\mathring{\widetilde{X}}_j^*\right)^T \mathring{\widetilde{Y}} / \left(\mathring{\widetilde{X}}_j^T \left(\mathbf{I} - \widetilde{\Pi}_j\right) \mathring{\widetilde{X}}_j\right)$$

Clearly both $\widetilde{c_j^*}$ and c_j^* converge in probability to the same object, hence the separation property also holds for $\widetilde{c_j^*}$. Since the latter is efficient in the MSE sense this suggests that the performance of the TCS algorithm will improve (or at the very least will not deteriorate) when $\widetilde{c_j^*}$ is used.

A feasible estimator

Since Σ_j is never observed $\widetilde{c_j^*}$ is in-feasible. Given any consistent estimator $\widehat{\Sigma}_j$ for Σ_j a feasible estimate for the j-th tilted correlation is as follows:

$$\widehat{c}_j^* = \left(\mathring{\widehat{X}}_j^*\right)^T \mathring{\widehat{Y}} / \left(\mathring{\widehat{X}}_j^T \left(\mathbf{I} - \widehat{\Pi}_j\right) \mathring{\widehat{X}}_j\right)$$

where $\widehat{X}_j = \widehat{\Sigma}_j^{-\frac{1}{2}} X_j$. Section 3.2.1 presents a proof for the equivalence of the limiting distribution of the feasible and in-feasible estimators. In addition to A(1)-A(9) the proof requires three assumptions listed below. If these conditions are met we can therefore expect the performance of the TCS algorithm to improve when $\widehat{c}_j^*$ is used.

Additional assumptions

A (10) *The process* $\left\{\left(y_t, \boldsymbol{x}_{t,\mathcal{D}_j}, \boldsymbol{u}_{t,\mathcal{D}_j}\right), t \in Z\right\}$ *is a strictly stationary* α*-mixing process with mixing coefficient given by A(7).*

A (11) $\{(u_{t,\mathcal{D}})\, t \in Z\}$ *is a stationary AR(p) process, where* $p \in [0, Cn^{\theta})$ *for some constant* $C > 0$.

A (12) *The estimator* $\widehat{\Sigma}_j$ *is consistent in spectral norm, i.e.* $P\left(\left\|\widehat{\Sigma}_j - \Sigma_j\right\|_2 > \varepsilon\right) \to 0$ *as* $n \to \infty$ *for any* $\varepsilon > 0$.

Assumption A(10) follows naturally from assumption A(7). Assumption A(11) imposes additional structure on Σ_j which will be necessary to construct an estimator which converges to its population counterpart sufficiently fast. In practice the assumption is not too restrictive, as any stationary process with continuous spectral density can be approximated arbitrarily well by a finite order linear AR process; see for example *Corollary 4.4.2* in Brockwell et al. (1991). Additionally, some structure will always need to be imposed on Σ_j as the wholly unstructured problem requires the estimation of $n(n+1)/2$ parameters from a sample of only n observations.

Under A(11) the error term is of the form $u_{t,\mathcal{D}_j} + \sum_{k=1}^{p} \alpha_k(j) u_{t-k,\mathcal{D}_j} = e_{t,\mathcal{D}_j}$ with $e_{t,\mathcal{D}_j} \sim_{i.i.d} \mathcal{N}(0, \omega_j^2)$ for some constant $\omega_j > 0$. An intuitive estimator $\widehat{\Sigma}_j$ can therefore be constructed by using residuals $\widehat{u}_{1,\mathcal{D}_j}, ..., \widehat{u}_{n,\mathcal{D}_j}$ from a first stage regression to estimate $\widehat{\alpha}_1(j), ..., \widehat{\alpha}_p(j)$, whereupon the estimates are substituted into the $n \times n$ auto-covariance matrix for an AR(p) process. This type of parametric approach is popular in the econometric literature and has been shown to give efficient parameter estimates by Amemiya (1973), Wickens (1969), and

Koreisha & Fang (2001), among others. Note that in practice the order of the AR process will be unknown; the order can be approximated by choosing the value of p which minimises the AIC.

Assumption A(12) requires that the estimator $\widehat{\Sigma}_j$ be more than point-wise consistent. This is necessary because in general point-wise consistency is not sufficient to guarantee that the feasible estimator will be more efficient than the least squares estimator; see for example the elegant counter-example in *Section 2.5.12* of Schmidt (1976). In practice A(12) rules out the naive estimator $\widetilde{\Sigma}_j = \left(\widehat{\gamma}_{|i-j|}\right)_{1\leqslant i,j\leqslant n}$ with $\widehat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-|k|} \widehat{u}_{t,\mathcal{D}_j}\widehat{u}_{t+|k|,\mathcal{D}_j}$; as shown by Wu & Pourahmadi (2009) the largest eigenvalue of $\left(\widetilde{\Sigma}_j - \Sigma_j\right)$ does not go to zero as the sample size goes to infinity, hence the estimator cannot be consistent in spectral norm.

3.2.1 Proof of efficiency

Theorem 3 *For the j-th tilted correlation, under assumptions A(1)-A(12) and using the intuitive estimator suggested in Section 3.2 it holds that:*

$$\sqrt{n}\left(\widehat{c}_j^* - \beta_{(j)}\right) \to_d \sqrt{n}\left(\widetilde{c}_j^* - \beta_{j\,(j)}\right)$$

Proof:

By *Theorem 15* in Schmidt (1976) sufficient conditions for the feasible and in-feasible tilted correlations to have the same limiting distribution are the following:

$$\lim_{n\to\infty} P\left(\Delta_3\right) := \lim_{n\to\infty} P\left(\frac{1}{n}\left|X_j^T\left(\widehat{\Sigma}_j^{-1} - \Sigma_j^{-1}\right)X_j\right| > \varepsilon\right) = 0$$
$$\lim_{n\to\infty} P\left(\Delta_4\right) := \lim_{n\to\infty} P\left(\frac{1}{\sqrt{n}}\left|X_j^T\left(\widehat{\Sigma}_j^{-1} - \Sigma_j^{-1}\right)\boldsymbol{\epsilon}_{\mathcal{D}}\right| > \varepsilon\right) = 0$$

where $\varepsilon > 0$ is a small positive constant. To show that $P(\Delta_3)$ gos to zero it will be useful to have some link between convergence of $\widehat{\Sigma}_j$ in spectral norm as guaranteed by A(12) and convergence of the inverse. Such a link is provided by the following argument, which is

taken from the proof of *Theorem 3* in Yousuf et al. (2018). Let $a_1 \geqslant a_2 \geqslant ... \geqslant a_n$ be the ordered eigenvalues of the matrix $\Sigma_j^{-\frac{1}{2}} \widehat{\Sigma}_j \Sigma_j^{-\frac{1}{2}}$. Using the eigenvalue-norm inequality, see for example *Proposition 4.4* in Gallier (2019), we have that:

$$\begin{aligned}
\lambda_{\min}\left(\Sigma_j\right) \left\| \widehat{\Sigma}_j^{-1} - \Sigma_j \right\|_2 &\leqslant \left\| \Sigma_j^{\frac{1}{2}} \left(\widehat{\Sigma}_j^{-1} - \Sigma_j^{-1} \right) \Sigma_j^{\frac{1}{2}} \right\|_2 \\
&= \left\| \Sigma_j^{\frac{1}{2}} \widehat{\Sigma}_j^{-1} \Sigma_j^{\frac{1}{2}} - \mathbf{I}_n \right\|_2 \\
&= \max_{i \leqslant n} \left| \frac{1 - a_i}{a_i} \right|
\end{aligned}$$

Going in the other direction we have that:

$$\begin{aligned}
\max_{i \leqslant n} \left| a_i - 1 \right| &= \left\| \Sigma_j^{-\frac{1}{2}} \widehat{\Sigma}_j \Sigma_j^{-\frac{1}{2}} - \mathbf{I}_n \right\|_2 \\
&\leqslant \lambda_{\max}(\Sigma_j^{-1}) \left\| \widehat{\Sigma}_j - \Sigma_j \right\|_2
\end{aligned}$$

Finally, define $a_j = \min_{i \leqslant n} \left| a_i^{-1} \right|$. By assumption A(11) the eigenvalues of Σ_j are guaranteed to be bounded away from zero and infinity, and the link is therefore valid. This is because the spectral density of a finite order AR process is bounded away from zero and infinity. Using the fact that with standardised and de-meaned entries $\|X_j\|_2^2 = n$ we have that:

$$\begin{aligned}
P\left(\Delta_3\right) &\leqslant P\left(\left\| \widehat{\Sigma}_j^{-1} - \Sigma_j^{-1} \right\|_2 > \varepsilon \right) \\
&\leqslant P\left(\max_{i \leqslant n} \left| \frac{1 - a_i}{a_i} \right| > \varepsilon' \right) \\
&\leqslant \underbrace{P\left(\left\| \widehat{\Sigma}_j - \Sigma_j \right\|_2 > \varepsilon'' / M_\varepsilon \right)}_{5} + \underbrace{P\left(a_j < M_\varepsilon \right)}_{6}
\end{aligned}$$

By setting $M_\varepsilon = 1 - \varepsilon$ convergence of term 5 is guaranteed under assumption A(12). Convergence of term 6 can be shown through its connection to the spectral norm of the

approximation error as follows:

$$\begin{aligned} P\left(a_j < M_\varepsilon\right) &\leqslant P\left(|a_j - 1| > M_\varepsilon - 1\right) \\ &= P\left(\max_{i \leqslant n} |a_i - 1| > \varepsilon\right) \\ &\leqslant P\left(\left\|\widehat{\Sigma}_j - \Sigma_j\right\|_2 > \varepsilon'\right) \end{aligned}$$

To show $P\left(\Delta_4\right)$ goes to zero we can make use of the additional structure provided by assumption A(10). The next argument closely follows the proof of *Condition 3.10* in Wickens (1969). Note that the error terms can be can be expressed in matrix form as $\boldsymbol{J}\boldsymbol{u}_{\mathcal{D}_j} = \boldsymbol{e}^*$, where e^*_{t,D_j} is defined as $e_{t,D_j} - \sum_{s=t-p}^{0} \alpha_{t-s}(j) u_{s,\mathcal{D}_j}$ for $t \leqslant p$ and e_{t,D_j} otherwise, and $\boldsymbol{J} = \boldsymbol{I}_n + \alpha_1(j)\boldsymbol{L} + ... + \alpha_p(j)\boldsymbol{L}^p$ with $\boldsymbol{L}$ being an augmented $\boldsymbol{0}_{n\times n}$ matrix given by:

$$\boldsymbol{L} = \begin{pmatrix} 0 & \boldsymbol{I}_{n-1} \\ 0 & 0 \end{pmatrix}$$

It therefore follows that:

$$\begin{aligned} E\left(\boldsymbol{J}\boldsymbol{u}_{\mathcal{D}_j}\left(\boldsymbol{J}\boldsymbol{u}_{\mathcal{D}_j}\right)^T\right) &= \boldsymbol{J}E\left(\boldsymbol{u}_{\mathcal{D}_j}\boldsymbol{u}_{\mathcal{D}_j}^T\right)\boldsymbol{J}^T \\ &= \boldsymbol{J}\Sigma_j\boldsymbol{J}^T \\ &= \omega_j^2 \begin{pmatrix} \boldsymbol{I}_{n-p} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G} \end{pmatrix} \end{aligned}$$

where $\boldsymbol{G}$ is a $p \times p$ positive definite matrix. Using the above, the inverse auto-covariance matrix can be written as:

$$\Sigma_j^{-1} = \omega_j^{-2} \boldsymbol{J}^T \begin{pmatrix} \boldsymbol{I}_{n-p} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} \end{pmatrix} \boldsymbol{J}$$

Let a wide-hat indicate that AR coefficients have been replaced by consistent sample estimates. As long as $\widehat{\omega}_j^{-2}$ is consistent for ω_j^{-2} we have that $\left(\widehat{\Sigma}_j^{-1} - \Sigma_j^{-1}\right)$ in Δ_4 can be replaced with $\left(\widehat{\Sigma}_j^{*} - \Sigma_j^{*}\right)$ where the relevant matrices are defined below:

$$\Sigma_j^{*} = \boldsymbol{J}^T \begin{pmatrix} \boldsymbol{I}_{n-p} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} \end{pmatrix} \boldsymbol{J}$$

$$\widehat{\Sigma}_j^{*} = \widehat{\boldsymbol{J}}^T \begin{pmatrix} \boldsymbol{I}_{n-p} & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{G}}^{-1} \end{pmatrix} \widehat{\boldsymbol{J}}$$

Partitioning $\boldsymbol{J}^T = \left(\boldsymbol{J}_1^T, \boldsymbol{J}_2^T\right)$ and $\widehat{\boldsymbol{J}}^T = \left(\widehat{\boldsymbol{J}}_1^T, \widehat{\boldsymbol{J}}_2^T\right)$, where the first sub-matrix has dimensions $(n-p) \times n$ and the second has dimensions $p \times n$, the left hand side of event Δ_4 can then be re-written as the following:

$$\begin{aligned} \frac{X_j \left(\widehat{\Sigma}_j^{*} - \Sigma_j^{*}\right) \boldsymbol{u}_{\mathcal{D}_j}}{\sqrt{n}} = & \underbrace{n^{-\frac{1}{2}} X_j^T \widehat{\boldsymbol{J}}_2^T \left(\widehat{\boldsymbol{G}}^{-1} - \boldsymbol{I}_p\right) \widehat{\boldsymbol{J}}_2^T \boldsymbol{u}_{\mathcal{D}_j} - n^{-\frac{1}{2}} X_j^T \boldsymbol{J}_2^T \left(\boldsymbol{G}^{-1} - \boldsymbol{I}_p\right) \boldsymbol{J}_2^T \boldsymbol{u}_{\mathcal{D}_j}}_{7} \\ & + \underbrace{n^{-\frac{1}{2}} X_j^T \left(\widehat{\boldsymbol{J}} - \boldsymbol{J}\right)^T \left(\widehat{\boldsymbol{J}} - \boldsymbol{J}\right) \boldsymbol{u}_{\mathcal{D}_j}}_{8} \\ & + \underbrace{n^{-\frac{1}{2}} X_j^T \left(\widehat{\boldsymbol{J}} - \boldsymbol{J}\right)^T \boldsymbol{J} \boldsymbol{u}_{\mathcal{D}_j}}_{9} + \underbrace{n^{-\frac{1}{2}} X_j^T \boldsymbol{J}^T \left(\widehat{\boldsymbol{J}} - \boldsymbol{J}\right)^T \boldsymbol{u}_{\mathcal{D}_j}}_{10} \end{aligned}$$

A sequence of random variables $\{a_n\}_{n=1}^{\infty}$ is said to be $\mathcal{O}_p(b_n)$, where $\{b_n\}_{n=1}^{\infty}$ is a deterministic sequence, if for any $\varepsilon > 0$ there exist constants $c > 0$ and $n_0 > 0$ such that $P\left(|a_n| > c \cdot b_n\right) < \varepsilon$ whenever $n \geqslant n_0$. By the specialised form of the $\boldsymbol{G}$ matrix, and the

element-wise consistence of $\widehat{\boldsymbol{J}}$ and $\widehat{\boldsymbol{G}}$, term 7 is at least $\mathcal{O}_p(n^{-\frac{1}{2}})$. By the central term being in quadratic form, term 8 goes to zero as long as $(\widehat{\alpha}_s(j) - \alpha_s(j)) = \mathcal{O}_p(n^{-\frac{1}{2}})$ for each $s \in [1, p]$, which is a condition generally satisfied by consistent estimators. Finally, terms 9 and 10 involve p^2 sums of the form $\alpha_s(j)\,(\widehat{\alpha}_r(j) - \alpha_r(j)) \times n^{-\frac{1}{2}}\,(\boldsymbol{L}^r X_j)^T\,\boldsymbol{L}^s \boldsymbol{u}_{\mathcal{D}_j}$ for $r, s \in [1, p]$. The desired result therefore holds whenever $n^{-\frac{1}{2}}\,(\boldsymbol{L}^r X_j)^T\,\boldsymbol{L}^s \boldsymbol{u}_{\mathcal{D}_j} = \mathcal{O}_p(1)$, which is clearly the case under assumption A(10).

3.3 Stable time series tilting

Construction of the set $\mathcal{C}_j$ is crucial to the performance of the TCS algorithm. Both the number of predictors incorrectly included in $\mathcal{C}_j$ (the false positives - FP) and the number incorrectly excluded from the set (the false negatives - FN) will reduce the effectiveness of the tilted correlation as a screening device. In Section 2.1 it was proposed that the choice of π_n controls the expected FDR at $p_n^{-\frac{1}{2}}$; this can be improved by choosing a larger π_n, hover a fall in the type I error necessitates a trade-off in terms of the type II error. In this section I propose a re-sampling approach based on the work of Meinshausen & Bühlmann (2010) which, for a given π_n, seeks to reduce the number of false positives with minimal impact on the number of false negatives.

Primer on stability selection

The stability selection of Meinshausen & Bühlmann (2010) is a wrap-around procedure designed to enhance any feature screening algorithm that selects a subset of features, say $\widehat{\mathcal{S}}$, via regularisation. The authors point out that choosing the right amount of regularisation is often extremely difficult. Stability selection uses a grid of regularisation parameters Λ and a set of pseudo samples $\boldsymbol{Z}_1^*, ..., \boldsymbol{Z}_N^*$ obtained by randomly sampling $n/2$ observations from the original data without replacement (i.e. sub-sampling). Let $\widehat{\mathcal{S}}_i^\lambda$ be the set of features selected by the base algorithm using pseudo-sample $\boldsymbol{Z}_i^*$ under some level of regularisation λ. For each feature k the selection probability under $\lambda \in \Lambda$ is calculated as $\widehat{\Pi}_k^\lambda = \frac{1}{N}\sum_{i=1}^N \mathcal{I}\left(k \in \widehat{\mathcal{S}}_i^\lambda\right)$, where $\mathcal{I}(\cdot)$ is the indicator function. The set of stable features is then defined as follows:

$$\widehat{\mathcal{S}}^{\text{stable}} = \left\{ k : \max_{\lambda \in \Lambda} \left(\widehat{\Pi}_k^{\lambda} \right) \geqslant \upsilon_{th} \right\}$$

Intuitively, stability selection injects noise into the original data by re-sampling, and selects only those features which survive the screening process with high probability. The authors claim that stability selection works well for any choice of υ_{th} in the interval $(0.6, 0.9)$. A theorem is also provided which bounds the expected number of falsely selected features (the family-wise type I error - FWE) based on the choice of υ_{th}, however this requires knowledge of the average number of features selected by the base algorithm.

Extension to stable tilting

Here I introduce stable tilting, in which for each predictor a 'stable' conditioning set $\mathcal{C}_j^{stb}$ is constructed according to the principle of stability selection. The procedure begins with calculating π_n using the original sample. A grid of thresholds is therefore defined as $\Lambda = \left\{ \pi_{(1)}, ..., \pi_{(G)} \right\}$, with $\pi_n = \pi_{(1)} < \pi_{(2)} < ... < \pi_{(G)} = c \cdot \pi_n$ for some $c \gg 1$. In keeping with stability selection pseudo-samples $\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*$ are generated from the original sample, and a stable conditioning set for the j-th predictor is constructed according to $\mathcal{C}_j^{stb} = \left\{ k : \max_{\pi \in \Lambda} \left(\widehat{\Pi}_k^{\pi} \right) \geqslant \upsilon_{th} \right\}$. Figure 3.2 provides a graphical representation of the procedure.

Figure 3.2: Diagram of stable tilting being used to construct a set of highly correlated predictors for the j-th predictor.

The construction of Λ ensures that for each pseudo-sample $\boldsymbol{X}_i^*$ and for every $\pi \in \Lambda$ the highest expected FDR is around $p_n^{-\frac{1}{2}}$, and in general will be smaller. Since selection is made on the basis of the highest selection probability among all thresholds, we can expect predictors which are correlated at the population level to survive the screening process. Meanwhile, predictors which attain a high correlation spuriously are less likely to do so over multiple pseudo-samples. A more refined approach would involve re-calculating π_n for each pseudo-sample to ensure the expected FRD is exactly $p_n^{-\frac{1}{2}}$. This however is computationally expensive, and simulation studies confirm that in practice the difference in selected variables is negligible.

The original stability selection procedure was developed for i.i.d. data. To adapt to the time series setting I suggest generating $\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*$ using the stationary bootstrap described in Section 2.3.1 as the method is well suited to replicating the dependence structure in stationary time series. Sub-sampling can also be used to generate pseudo samples from stationary time series using blocks of data, see *Chapters 2-3* of Politis et al. (1999) for a detailed explanation. However, there is very little literature on using random block sizes to preserve strict stationarity as in the case of the stationary bootstrap. Additionally, there are subtle differences between the two procedures: sub-sampling treats each block as its own "mini time series" whereas the bootstrap uses blocks as "building stones to construct a new pseudo time series"; see *Section 3.9* of Politis et al. for a comparison. In the context of selecting a stable conditioning set, the latter seems preferable. In the i.i.d. setting Meinshausen and Bühlmann opt for sub-samples of size $n/2$ as they "resembles most closely the bootstrap while allowing computationally efficient implementation". However open source implementations of stability selection in the R and Python communities, see for example Huijskens (2018), allow for the use of the bootstrap as in practice the computational gains from sub-sampling in terms or run-time are modest - this further supports the use of the stationary bootstrap.

Finally, I note that stable tilting can be seen as a first step in addressing Professor Howell Tong's question raised in the discussion section of Meinshausen & Bühlmann (2010), namely whether stability selection may be used in time series modelling. This question, to my knowledge, has not yet been explored in the literature.

3.3.1 Simulation study

To illustrate the effect of stable tilting on membership of the conditioning set I present the following three simulation studies. Here the idea is to show that stable tilting can be used to construct a more parsimonious conditioning set than tilting alone. In each simulation the sample size was fixed at $n = 500$ and the number of predictors was varied over $p = \{50, 500, 1000\}$. For ease of comparison I focus on $\mathcal{C}_1$, the conditioning set for the first predictor. To construct a stable conditioning set $N = 50$ bootstrap replicates were generated using the stationary bootstrap of Politis & Romano (1994) with the average

block size set to $l_n = 25$, and v_{th} was set to 0.6.

Simulation no. 1

Data was generated according to a VAR(1) process given by $\mathbf{x}_t = \Phi \mathbf{x}_{t-1} + \boldsymbol{\nu}_t$ as in Section 3.1.3. The vector of innovations was generated according to $\boldsymbol{\nu}_t \sim_{i.i.d} \mathcal{N}_p(\mathbf{0}, \Sigma)$ where $\Sigma = \left\{\rho^{|i-j|}\right\}_{i,j \leqslant p}$ and ρ was varied over $\{0.7, 0.8, 0.9\}$. The first row of Σ was replaced by the vector $(1, 0, ..., 0)$. The coefficient matrix was set to $\Phi = diag(0.4)$ and the first row was replaced by a zero vector with the first 10 entries set to 0.5.

We can express $x_{t,1} = 0.5 \sum_{k=2}^{10} x_{t,k} + \nu_{t,1}$, and conditional on the set $\mathcal{C}_1^* = \{2, ..., 10\}$ predictor one is independent of the remaining predictors. I therefore count a false negative as excluding a member of $\mathcal{C}_1^*$ and a false positive as including a member of $\mathcal{J} \setminus \{\mathcal{C}_1^* \cup \{1\}\}$. Clearly, not accounting for $\mathcal{C}_1^*$ predictor one is correlated with every other predictor. The test is therefore more challenging than the setting described in Section 3.3, where the aim was simply to filter out uncorrelated predictors.

Simulation no. 2

The setup is identical to the previous simulation, with the additional condition that each predictor is observed with additive i.i.d. $\mathcal{N}(0, 3)$ noise. This replicates the realistic scenario in which a time series is observed with some non-systematic measurement error. Since the variance of the contaminating noise is nine times larger than that of marginal innovation in the VAR process, this setting is considerably more challenging than the last.

Simulation no. 3

The setup is identical to the first simulation, with the exception that the first row of Φ is instead replaced by a zero vector where the first entry along with nine more entries spaced maximally apart are set to 0.5. For example, with $p = 500$ the set of predictors we aim to identify is $\mathcal{C}_1^* = \{56, 112, 167, 223, 278, 334, 389, 445, 500\}$. This setting is more challenging than the previous simulations: where predictors in $\mathcal{C}_j^*$ were highly correlated among themselves and weakly correlated with irrelevant predictors, now predictors in the target set are attain the minimal possible correlation among themselves and are highly

correlated with 'irrelevant' predictors.

Simulation results

Simulation results are reported in Tables 3.1, 3.2, and 3.3, where FP and FN were calculated as the the average of 100 trials. The improved performance of stable tilting is immediately clear: in each scenario stable tilting achieves a lower value for FP+FN than tilting alone. The result is driven, as expected, by the fact that stable tilting commits far fewer type II errors. stable tilting does tends to commit slightly more type I errors; this is because using a grid of parameters on bootstrap samples destroys the expected FDR control enjoyed by the original procedure. However, for sufficiently large ρ the effect is negligible. This suggests that the new procedure will work best when the design matrix is highly correlated. All things being equal increasing the number of predictors has a greater effect on the number of type II errors committed, suggesting the new procedure should be used when p_n is large. Finally, it should be noted that stable tilting is highly computationally intensive. With the parameters chosen in these simulation each $\mathcal{C}_1^{stb}$ is around 1,000 times more expensive to compute than $\mathcal{C}_1$. Since the TCS algorithm computes $\mathcal{O}(m \cdot n^{\xi})$ tilted correlations, stable tilting should only be used when it is suspected that doing so will improve significantly the performance of the TCS algorithm.

	(p = 50)						(p = 500)						(p = 1,000)					
	Tilting			Stability			Tilting			Stability			Tilting			Stability		
ρ	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN
0.7	1.9	0.6	2.5	1.3	1.1	**2.4**	63.1	0.1	63.2	43.0	0.3	**43.2**	238.0	0.1	238.1	211.0	0.2	**211.1**
0.8	3.3	0.2	3.5	2.1	0.4	**2.5**	63.1	0.0	63.1	44.1	0.1	**44.2**	244.9	0.0	244.9	219.4	0.0	**219.4**
0.9	7.6	0.0	7.6	5.9	0.0	**5.9**	71.8	0.0	71.8	52.3	0.0	**52.3**	244.6	0.0	244.6	215.3	0.0	**215.3**

Table 3.1: Screening accuracy of tilting versus stabilised tilting for the set $\mathcal{C}_1$ with nine consecutive relevant predictors in a VAR(1) setting; lowest values for FP+FN are in bold.

	(p = 50)						(p = 500)						(p = 1,000)					
	Tilting			Stability			Tilting			Stability			Tilting			Stability		
ρ	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN
0.7	1.9	0.6	2.5	1.1	1.3	**2.4**	58.7	0.2	58.8	40.2	0.3	**40.5**	230.5	0.1	230.6	202.2	0.2	**202.4**
0.8	3.2	0.2	3.4	2.4	0.3	**2.7**	59.7	0.0	59.7	39.1	0.1	**39.2**	232.7	0.0	232.7	203.1	0.0	**203.1**
0.9	7.5	0.0	7.6	5.7	0.1	**5.8**	69.2	0.0	69.2	49.2	0.0	**49.2**	238.2	0.0	238.2	210.1	0.0	**210.1**

Table 3.2: Screening accuracy in a VAR(1) setting with nine consecutive relevant predictors where each predictor is observed with noise; lowest values for FP+FN are in bold.

	(p = 50)						(p = 500)						(p = 1,000)					
	Tilting			Stability			Tilting			Stability			Tilting			Stability		
ρ	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN	FP	FN	FP+FN
0.7	8.7	5.3	14.0	5.5	6.4	**12.0**	66.*o*	3.8	69.8	45.4	4.8	**50.2**	243.0	2.7	245.7	214.9	3.4	**218.3**
0.8	17.6	3.8	21.4	13.0	5.0	**18.0**	77.0	3.6	80.6	53.6	4.6	**58.2**	250.7	2.5	253.2	223.0	3.2	**226.2**
0.9	31.5	1.6	33.1	27.5	2.5	**30.0**	96.4	3.8	100.2	68.4	4.7	**73.1**	271.9	2.6	274.5	238.3	3.3	**241.6**

Table 3.3: Screening accuracy in a VAR(1) setting with nine maximally spaced relevant predictors (observed without noise); lowest values for FP+FN are in bold.

Chapter 4

Numerical Studies

4.1 Extended simulation studies

In this section simulation studies are used to compare the screening accuracy of the TCS algorithm executed with standard tilted correlations against that of same algorithm executed with extensions proposed in Sections 3.2 and 3.3. For each simulation data was generated according to a VAR(1) process given by $\mathbf{x}_t = \Phi\mathbf{x}_{t-1} + \boldsymbol{\nu}_t$ as in Section 3.1.3. Similarly the response variable was generated according to $y_t = \boldsymbol{\beta}^T\mathbf{x}_t + e_t$ where e_t followed an AR(1) process, and the five non-zero entries in $\boldsymbol{\beta}$ were set to either -1 or 1. Each time the TCS algorithm was run the stopping index was set to $m = 20$. Stable tilted correlations were estimated as set out in Section 3.3.1. For ease of comparison results are presented in the form of Receiver Operator Characteristic (ROC) curves, which plot the True Positive Rate (TPR) against the False Positive Rate (FPR).

4.1.1 General comparison of screening performance

For a general comparison of the three methods I first consider a setting which conforms to assumptions A(1)-A(12), and record the effect of varying the population noise level and sample size. In keeping with *Section 4* of Cho & Fryzlewicz (2012) the noise level is quantified via $R^2 = V\left(\boldsymbol{\beta}^T\mathbf{x}_t\right)/V\left(y_t\right)$, where $V(\cdot)$ gives the population variance. To gauge sensitivity to low sample sizes the number of predictors was held constant at $p = 1,000$

and the sample size was varied over $n = \{500, 200, 100\}$. In addition, to gauge sensitivity to the noise level innovations in the AR(1) process were drawn from $\mathcal{N}(0, r^2)$ and r was varied in order to attain $R^2 = \{0.9, 0.6, 0.3\}$.

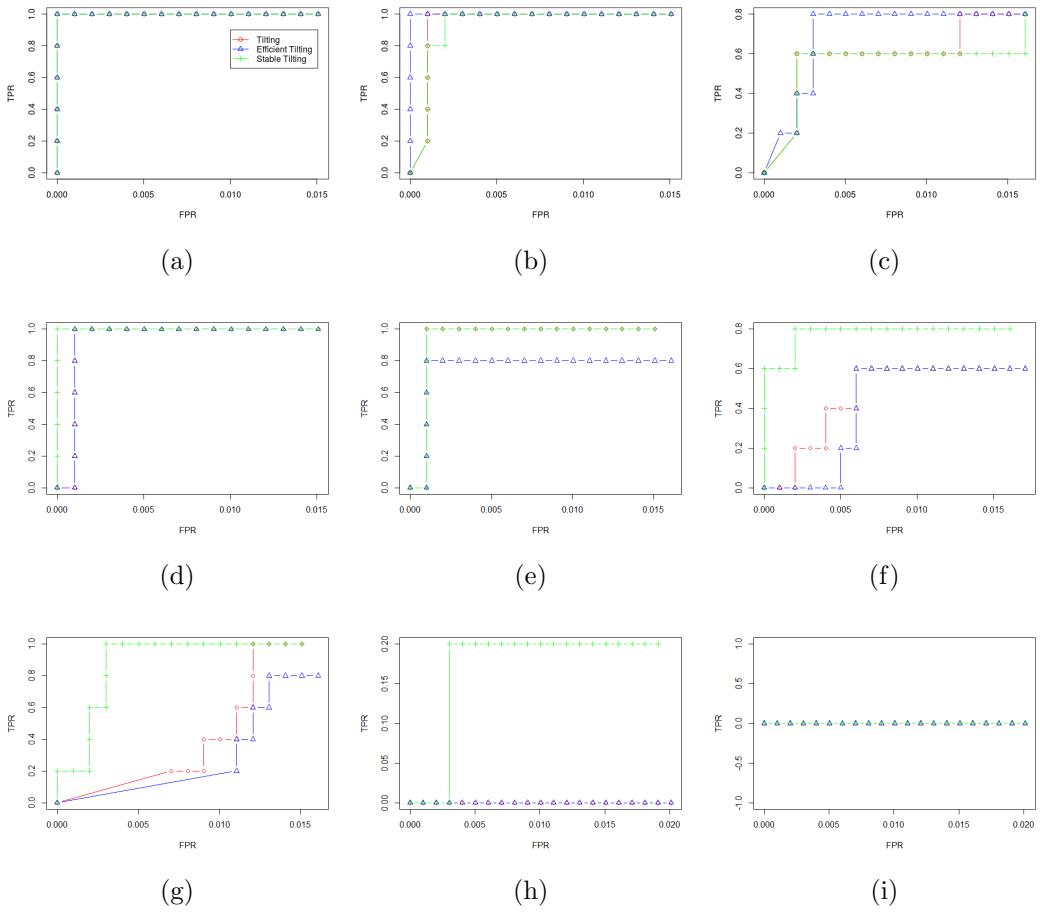

Figure 4.1: ROC curves for three variants of the TCS algorithm in a VAR setting. Along each column $R^2 = \{0.9, 0.6, 0.3\}$ ans long each row $n = \{500, 200, 100\}$

.

ROC curves for the above simulation are presented in Figure 4.1. When the sample size is large all three methods are robust to the noise level. With $n = 500$ efficient tilting outperforms the other the other two methods, however its screening performance deteriorates as the sample size falls. This is likely because parameter estimates for the AR process used to used to estimate the error auto-covariance matrix become unstable in smaller samples. Stable tilting is overall robust to both the sample size and the noise level. For example in

plot (f), where $n = 200$ and $R^2 = 0.6$, stable tilting finds 80 percent of relevant predictors whereas the remaining method find only 60 percent. It should be noted that for moderate sample sizes and $R^2 \geqslant 0.6$ stable tilting does not perform much better than the original procedure. Given the high computational cost of stable tilting this suggests that in most settings regular tilting is sufficient. Unsurprisingly, with $n = 100$ and $R^2 = 0.3$ all three methods break down completely.

4.1.2 Robustness to additional assumptions

As expressed in Section 3.1 some assumptions used to extend tilted correlations to the time series setting are quite restrictive. This section explores the effect of relaxing two assumptions which in real datasets may not be satisfied. To isolate the effect of relaxing the assumptions the sample size was varied over $n = \{500, 200\}$ only, as Figure 4.1 shows that as all three methods performed well in this range.

Relaxing assumption A(8)

Assumption A(8) requires that each predictor be independent of the error term e_t. To test the effect of relaxing the assumption innovations in the AR(1) error process were modified as shown below. The value of r was again varied to attain $R^2 = \{0.9, 0.6, 0.3\}$.

$$
\begin{aligned}
e_t &= 0.6e_{t-1} + \iota_t \\
\iota_t &= w_t \cdot \sum_{j \in \mathcal{S}} x_{tj} \quad \text{with} \quad w_t \sim_{i.i.d.} \mathcal{N}\left(0, r^2\right)
\end{aligned}
$$

Simulation results are presented in Figure 4.2. With $n = 500$ all three methods seem to perform well, although strangely stable tilting performs worse than the other two methods in plot (c), in which $n = 500$ and $R^2 = 0.3$. The performance of efficient tilting is sensitive to high noise levels when the sample size is low. This may be explained by the fact that in the presence of heteroskedasticity AIC selects a higher order AR model, leading to even more unstable parameter estimates.

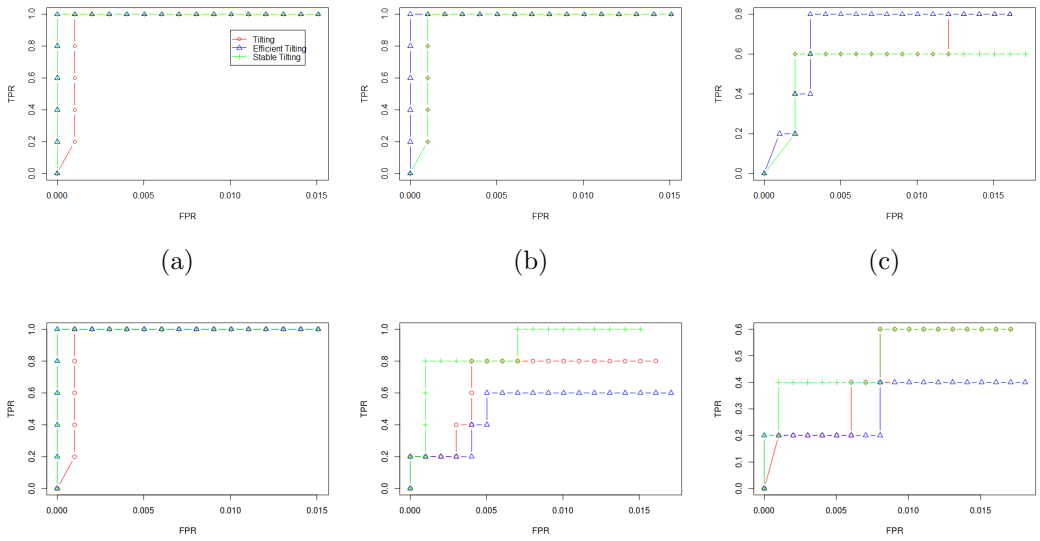

Figure 4.2: ROC curves for three variants of the TCS algorithm in a VAR setting with heteroskedasticity. Along each column $R^2 = \{0.9, 0.6, 0.3\}$ ans long each row $n = \{500, 200\}$

Relaxing assumption A(9)

Assumption A(9) requires either that the predictors be bounded or that all of their moments exist. To test the effect of relaxing the assumption innovations in the VAR(1) process were genertaed according to $\boldsymbol{\nu}_t \sim_{i.i.d.} \boldsymbol{t}_3(\mathbf{0}, \Sigma)$, where $\boldsymbol{t}_3$ represents a multivariate t-distribution with 3 degrees of freedom. The predictors are therefore unbounded and only the first two moments are finite. Note however that this setting is sufficient to guarantee consistency of regression parameters estimated via OLS. The shape matrix Σ was set to equal the variance-covaraince in the multivariate Normal case, and values for r were recycled from Section 4.1.1.

Simulation results are presented in Figure 4.3. Results are largely similar to those in 4.1.1, with the exception that efficient tilting performs slightly worse when $n = 200$ and the noise level is high. The fact that TCS algorithm does not break down completely confirms that A(9) is a purely technical assumption which arises from the strong mixing condition. In fact, similar results to Theorems 1 and 2 could have been proven using a number of dependence measures for time series processes. For example, the functional dependence measure of Wu

(2005) used in Yousuf et al. (2018) allows for unbounded errors and predictors and only requires that a finite number of moments exist.

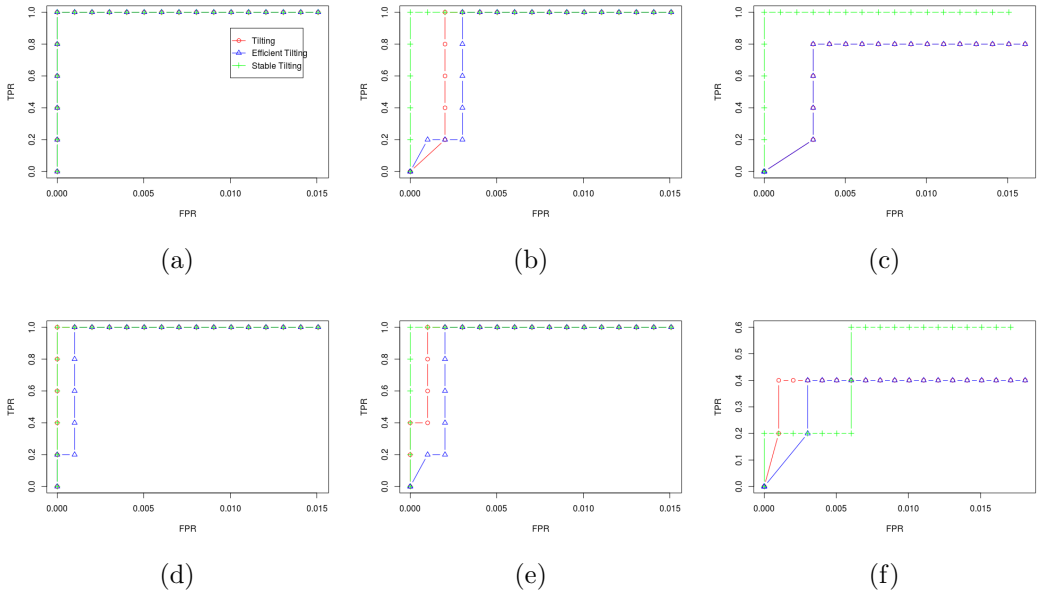

Figure 4.3: ROC curves for three variants of the TCS algorithm in a VAR setting; only the first two moments of each predictor exist. Along each column $R^2 = \{0.9, 0.6, 0.3\}$ ans long each row $n = \{500, 200\}$

Disclaimer!

Due to the high computational cost of stable tilting each of the ROC plots presented above is the result of a single simulation. The accompanying discussion is therefore only indicative of the relative performance of the three methods. Note however that results presented can be reproduced exactly using the R code provided on GitHub.

4.2 Forecasting UK macroeconomic time series

To illustrate the practical utility of tilted correlations in a time series setting here I present a numerical study in which tilted correlations are used to select a model with which to forecast three key macroeconomic variables for the UK economy:

$$\boldsymbol{y}_{t+1} = \begin{pmatrix} y_{t+1,1} \\ y_{t+1,2} \\ y_{t+1,3} \end{pmatrix} = 100 \times \begin{pmatrix} \log(D7BT_{t+1}/D7BT_t) \\ \log(ABMI_{t+1}/ABMI_t) \\ \log(XUQABK_{t+1}/XUQABK_t) \end{pmatrix}$$

These are roughly equal to the one period ahead percentage changes in quarterly inflation (D7BT), aggregate output (ABMI), and Sterling effective exchange rate (XUQABK67).

4.2.1 Forecasting Setup

Data

Data for this study was obtained from the Office for National Statistics (ONS) and Bank of England (BoE) websites. A large set of low frequency time series were collected by following *Appendix D* of Kapetanios et al. (2008), which explains how to reproduce the dataset used by Stock & Watson (2002) with UK data. All time series were transformed to achieve stationarity according to *Appendix D*, and a summary of the data along with transformations applied is provided in Table 4.5.

Since the Bank of England began using monetary policy to target inflation in 1992, and was granted independence in 1997, we can expect pronounced structural breaks in D7BT and XUQABK67 around this period. The data used therefore begins in the first quarter to 1998. The final set of predictors includes three factors obtained by taking principal components of the stationary predictors $\boldsymbol{x}_t$ and denoted by $\widehat{\boldsymbol{f}}_t = \left(\widehat{f}_{t,1}, \widehat{f}_{t,2}, \widehat{f}_{t,3}\right)^T$. Additionally, the first three lags of $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$ were added to the set for a total of $p = 110$ potential predictors.

Model selection with tilting

The TCS algorithm set out in Algorithm 2 was first used to reduce the number of potential predictors to an active set $\mathcal{A}$ of size $|\mathcal{A}| = m$. I followed *Section 3.1* in Cho & Fryzlewicz (2012) in setting $m = \text{floor}\{n/2\}$. This value is designed to guarantee that projections performed by the algorithm are numerically stable while allowing a sufficiently large number of predictors to enter into $\mathcal{A}$. Where projections were numerically unstable the poorly conditioned matrix was augmented by adding a small amount of noise distributed according to $\mathcal{N}(0, 0.001)$ to each entry prior to inversion.

Algorithm 2 was run with tilted correlations from the original paper (Tilt), efficient tilted correlations introduced in Section 3.2 (Tilt^E), and stable tilted correlations introduced in Section 3.3 (Tilt^S). A final model was obtained by applying the Lasso (+Lasso) to predictors in the active set $\mathcal{A}$, where the shrinkage parameter was selected via 10 fold cross-validation.

Benchmark models

The predictive performance of the TCS algorithm was bench-marked against forecasts from the four models presented in Table 4.1. These are models which are known to do well in practice, and consist of: a univariate auto-regression (AR), a vector auto-regression (VAR), and two factor augmented models (+F). Recently, forecasts from factor augmented models have garnered much attention; Boivin & Ng (2005) for example note that institutions including the Federal Reserve of Chicago, the U.S. Treasury, the European Central Bank, the European Commission, and the Centre for Economic Policy Research are all investigating the performance of factor augmented models. Such models are considered useful as they offer a means of incorporating information from a large number of related time series while avoiding the curse of dimensionality.

Model	Forecast	Parameter selection	Estimation
AR	$\widehat{y}_{t+1,j} = \widehat{c} + \sum_{\tau=1}^{p} \widehat{\phi}_\tau y_{t+1-\tau,j}$	AIC	Yule-Walker
VAR	$\widehat{\boldsymbol{y}}_{t+1} = \widehat{\boldsymbol{c}} + \sum_{\tau=1}^{p} \widehat{\Phi}_\tau \boldsymbol{y}_{t+1-\tau}$	AIC	Least Squares
AR+F	$\widehat{y}_{t+1,j} = \widehat{c} + \sum_{\tau=1}^{3} \widehat{\phi}_\tau y_{t+1-\tau,j} + \sum_{k=1}^{3} \widehat{\gamma}_k \widehat{f}_{t,k}$	3 lags + 3 factors	Least Squares
VAR+F	$\widehat{\boldsymbol{y}}_{t+1} = \widehat{\boldsymbol{c}} + \sum_{\tau=1}^{3} \widehat{\Phi}_\tau \boldsymbol{y}_{t+1-\tau} + \widehat{\Gamma}_\tau \widehat{\boldsymbol{f}}_{t+1-\tau}$	3 lags + 3 factors	Least Squares*

Table 4.1: Summary of benchmark models used in numerical study. The symbol * indicates the model was estimated according to *Section II.C* of Bernanke et al. (2005).

4.2.2 Forecasting results

Tables 4.2, 4.3, and 4.4 present the mean square forecast error (MSE), mean absolute forecast error (MAE), and average number of covariates selected (Size) for each method across each time series. Forecasts were produced using a rolling window scheme with $n = 50$

observations, and a total of 32 forecasts were made.

Excluding the inflation time series, for which AR-type models performed best, forecasts obtained via tilting outperformed all four benchmark models in terms of MSE and MAE. Among the tilting variants stable tilting consistently produced the most accurate forecasts, followed by efficient tilting. Efficient tilting consistently selected a more parsimonious model than regular tilting, hence its improved performance may be due to the fact that a model with fewer parameters enjoys a lower forecast variance. However, stable tilting occasionally selected a larger model than regular tilting. This point to the fact that by selecting predictors on the basis of inclusion frequencies stable tilting was indeed isolating predictors which contributed to the true causal relationship with the response.

Forecasts from multivariate models were consistently poor. However, the robust performance of AR type models across all three time series is impressive given the simplicity and low computational costs of these models. Finally I note that factor augmented AR models outperformed AR models in roughly half of all cases, which contradicts the results of the numerical study presented in *Section 7* of Yousuf et al. (2018).

	AR(p)	AR(3) + F(3)	Tilt + Lasso	Tilt^E + Lasso	Tilt^S + Lasso	VAR(p)	VAR(3)+F(3)
MSE	**1.32e−3**	1.81e−3	2.02e−3	1.92e−3	1.94e−3	7.34e−3	3.14e−3
MAE	2.99e−1	**2.08e−1**	3.51e−1	3.36e−1	3.36e−1	6.95e−1	4.48e−1
Size	5.84	6.00	7.00	5.97	5.56	**2.99**	9.00

Table 4.2: Average prediction errors and average number of predictors used, for seven models used to predict the quarterly percentage change in UK inflation measured via the consumer price index (CPI); lowest values for MSE, MAE, and average model size are in bold.

	AR(p)	AR(3) + F(3)	Tilt + Lasso	Tilt^E + Lasso	Tilt^S + Lasso	VAR(p)	VAR(3)+F(3)
MSE	1.89e−3	1.43e−3	1.13e−3	1.10e−3	**9.07e−4**	9.61e−3	3.17e−3
MAE	3.45e−1	2.99e−1	2.69e−1	2.66e−1	**2.22e−1**	7.69e−1	4.33e−1
Size	**2.81**	6.00	4.28	3.97	4.44	2.99	9.00

Table 4.3: Average prediction errors and average number of predictors used, for seven models used to predict the quarterly percentage change UK gross domestic product (ABMI); lowest values for MSE, MAE, and average model size are in bold.

	AR(p)	AR(3) + F(3)	Tilt + Lasso	Tilt^E + Lasso	Tilt^S + Lasso	VAR(p)	VAR(3)+F(3)
MSE	6.36e−2	8.09e−2	6.59e−2	6.43e−2	**6.14e−2**	1.87e−1	9.17e−2
MAE	1.90e0	2.16e0	1.99e0	1.95e0	**1.93e0**	3.66e0	2.47e0
Size	2.34	6.00	2.34	2.16	**1.38**	29.91	9.00

Table 4.4: Average prediction errors and average number of predictors used, for seven models used to predict the quarterly percentage change in the effective exchange rate for Sterling against a basket of representative currencies (XUQABK67); lowest values for MSE, MAE, and average model size are in bold.

Identifier	Source	Description	Transformation
UTKZ	ONS	Services price index	$\Delta^2 \log$
UTKX	ONS	Non-durable goods price index	$\Delta^2 \log$
UTLB	ONS	Semi-durable goods price index	$\Delta^2 \log$
UTKT	ONS	Durable goods price index	$\Delta^2 \log$
ABJS	ONS	Household final consumption expenditure (index)	$\Delta^2 \log$
PLLU	ONS	Output of manufactured products	$\Delta^2 \log$
IUQAAMIH	BoE	Average of 4 UK Banks' base rates	–
Rate	BoE	Official Bank Rate history	–
D7BT	ONS	Consumer price inflation, all items	$\Delta \log$
XUQAERG	BoE	Average Effective exchange rate, Euro	$\Delta \log$
XUQAUSG	BoE	Average Effective exchange rate, US $	$\Delta \log$
XUQABK67	BoE	Average Effective exchange rate, Sterling	$\Delta \log$
FBYH	ONS	Change in Inventories, Retail	$\Delta \log$
FAJM	ONS	Change in Inventories, Wholesale	$\Delta \log$
DLWX	ONS	Change in Inventories, Other Industries	$\Delta \log$
ABJR	ONS	Household final consumption expenditure (£m)	$\Delta \log$
UTID	ONS	Total consumption, Durable goods	$\Delta \log$
UTIT	ONS	Total consumption, Semi-durable goods	$\Delta \log$
UTIL	ONS	Total consumption, Non-durable goods	$\Delta \log$
UTIP	ONS	Total consumption, Services	$\Delta \log$
TMMI	ONS	Purchase of vehicles	$\Delta \log$
MGRZ	ONS	Number of People in Employment	$\Delta \log$
ABMI	ONS	Gross Domestic Product	$\Delta \log$
NRJR	ONS	Real Households' disposable income	$\Delta \log$

Table 4.5: Summary of low frequency macroeconomic data used in numerical study along with Office for National Statistics (ONS) / Bank of England (BoE) identifier and transformation applied to achieve stationarity.

Chapter 5

Conclusion

This dissertation has studied the largely unexplored problem of feature screening for high dimensional linear models in a time series setting. This was done by extending the TCS algorithm of Cho & Fryzlewicz (2012) to the time series setting while assuming strong mixing conditions. In Section 3.2 an efficient variant of the tilted correlation measure was proposed. In section 3.3 a stable variant of the tilted correlation was proposed. This extension adapted the principle of stability selection introduced by Meinshausen & Bühlmann (2010) to generate a stable conditioning set for each predictor. Finally, simulation studies and a real data example in Chapter 4 confirmed that all three methods are good at distinguishing relevant predictors from irrelevant predictors in a high dimensional time series setting, with stable tilting generally outperforming the other two methods.

It would have been interesting to compare the screening performance of the TCS algorithm to other techniques adapted to the time series setting, such as GLSS of Yousuf et al. (2018) and KSIS of Chen et al. (2018). However, code for these methods was not provided, and the emphasis of this dissertation was to develop a new screening procedure as opposed to comaring existing ones. Stable tilting was carried out using the stationary bootstrap of Politis & Romano (1994). It would certainly have have been useful to develop a data driven procedure for choosing the block length. For example, the block length could have been chosen to minimise the MSE of bootstrap replicates used to estimate the covariance between two predictors.

There are several directions in which the results presented in this dissertation may be extended. Numerical studies in Chapter 4 show that stable tilting generally outperforms the other methods considered. An advantage of stable tilting is that it is robust to small changes in the data, however the method's computational cost may be an prohibitively high price to pay. Further work could investigate the estimating tilted correlations using a partitioned median regression, which is robust to outliers and far less computationally expensive. Additionally, Chapter 4 revealed that efficient tilting was particularly sensitive to low sample sizes. It was speculated that parameter estimates for the AR process used to construct a weighting became unstable in low sample sizes. Further work could explore the possibility of choosing the order of the AR process by placing greater weight on the sample size, thus preventing highly unstable parameter estimates from being used in the weighting matrix.

Chapter 6

Bibliography

Aitkin, A. (1935), 'On least squares and linear combination of observations', *Proceedings of the Royal Society of Edinburgh* **55**, 42–48.

Amemiya, T. (1973), 'Generalized least squares with an estimated autocovariance matrix', *Econometrica: Journal of the Econometric Society* pp. 723–732.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**(2).

Bernanke, B. S., Boivin, J. & Eliasz, P. (2005), 'Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach', *The Quarterly journal of economics* **120**(1), 387–422.

Boivin, J. & Ng, S. (2005), Understanding and comparing factor-based forecasts, Technical report, National Bureau of Economic Research.

Bosq, D. (2012), *Nonparametric statistics for stochastic processes: estimation and prediction*, Vol. 110, Springer Science & Business Media.

Bradley, R. C. et al. (2005), 'Basic properties of strong mixing conditions. a survey and some open questions', *Probability surveys* **2**, 107–144.

Brockwell, P. J., Davis, R. A. & Fienberg, S. E. (1991), *Time Series: Theory and Methods: Theory and Methods*, Springer Science & Business Media.

Chen, J., Li, D., Linton, O. & Lu, Z. (2018), 'Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series', *Journal of the American Statistical Association* **113**(522), 919–932.

Cho, H. & Fryzlewicz, P. (2012), 'High dimensional variable selection via tilting', *Journal of the Royal Statistical Society: series B (statistical methodology)* **74**(3), 593–622.

Choi, Y.-G., Lim, J. & Choi, S. (2019), 'High-dimensional markowitz portfolio optimization problem: empirical comparison of covariance matrix estimators', *Journal of Statistical Computation and Simulation* **89**(7), 1278–1300.

De Mol, C., Giannone, D. & Reichlin, L. (2008), 'Forecasting using a large number of predictors', *Journal of Econometrics* **146**(2), 318–328.

Efron, B. (1979), Bootstrap methods: another look at the jackknife, *in* 'Breakthroughs in statistics', Springer, pp. 569–593.

El Karoui, N. et al. (2008), 'Spectrum estimation for large dimensional covariance matrices using random matrix theory', *The Annals of Statistics* **36**(6), 2757–2790.

El Karoui, N. et al. (2010), 'High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints', *The Annals of Statistics* **38**(6), 3487–3566.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Fan, J. & Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**(1), 101.

Fryzlewicz, P., Rao, S. S. et al. (2011), 'Mixing properties of arch and time-varying arch processes', *Bernoulli* **17**(1), 320–346.

Gallier, J. H. (2019), 'Fundamentals of linear algebra and optimization'.
URL: *http://www.cis.upenn.edu/ cis515/*

Huijskens, T. (2018), 'A scikit-learn compatible implementation of stability selection'. https://github.com/scikit-learn-contrib/stability-selection.
URL: *https://github.com/scikit-learn-contrib/stability-selection*

Kapetanios, G., Labhard, V. & Price, S. (2008), 'Forecast combination and the bank of england's suite of statistical forecasting models', *Economic Modelling* **25**(4), 772–792.

Karoui, N. E. & Purdom, E. (2016), 'Can we trust the bootstrap in high-dimension?', *arXiv preprint arXiv:1608.00696* .

Koreisha, S. G. & Fang, Y. (2001), 'Generalized least squares with misspecified serial correlation structures', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(3), 515–531.

Kunsch, H. R. (1989), 'The jackknife and the bootstrap for general stationary observations', *The annals of Statistics* pp. 1217–1241.

Meinshausen, N. & Bühlmann, P. (2010), 'Stability selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4), 417–473.

Mokkadem, A. (1988), 'Mixing properties of arma processes', *Stochastic processes and their applications* **29**(2), 309–315.

Politis, D. N. & Romano, J. P. (1994), 'The stationary bootstrap', *Journal of the American Statistical association* **89**(428), 1303–1313.

Politis, D. N., Romano, J. P. & Wolf, M. (1999), *Subsampling*, Springer Science & Business.

Rosenblatt, M. (1956), 'A central limit theorem and a strong mixing condition', *Proceedings of the National Academy of Sciences of the United States of America* **42**(1), 43.

Schmidt, P. (1976), *Econometrics*, M. Dekker.

Åse, D. (2013), 'Big data, for better or worse', *Science Daily* .

Shao, J. & Tu, D. (2012), *The jackknife and bootstrap*, Springer Science & Business Media.

Song, S. & Bickel, P. J. (2011), 'Large vector auto regressions', *arXiv preprint arXiv:1106.3915* .

Stock, J. H. & Watson, M. W. (2002), 'Macroeconomic forecasting using diffusion indexes', *Journal of Business & Economic Statistics* **20**(2), 147–162.

Székely, G. J., Rizzo, M. L., Bakirov, N. K. et al. (2007), 'Measuring and testing dependence by correlation of distances', *The annals of statistics* **35**(6), 2769–2794.

Székely, G. J., Rizzo, M. L. et al. (2014), 'Partial distance correlation with methods for dissimilarities', *The Annals of Statistics* **42**(6), 2382–2412.

Teräsvirta, T., Tjøstheim, D., Granger, C. W. J. et al. (2010), *Modelling nonlinear economic time series*, Oxford University Press Oxford.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L. & Canales-Rodríguez, E. (2005), 'Estimating brain functional connectivity with sparse multivariate autoregression', *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**(1457), 969–981.

Wainwright, M. J. (2019), *High-dimensional statistics: A non-asymptotic viewpoint*, Vol. 48, Cambridge University Press.

Wickens, M. (1969), 'The consistency and efficiency of generalized least squares in simultaneous equation systems with autocorrelated errors', *Econometrica: Journal of the Econometric Society* pp. 651–659.

Wu, W. B. (2005), 'Nonlinear system theory: Another look at dependence', *Proceedings of the National Academy of Sciences* **102**(40), 14150–14154.

Wu, W. B. & Pourahmadi, M. (2009), 'Banding sample autocovariance matrices of stationary processes', *Statistica Sinica* pp. 1755–1768.

Yousuf, K. & Feng, Y. (2018), 'Partial distance correlation screening for high dimensional time series', *arXiv preprint arXiv:1802.09116* .

Yousuf, K. et al. (2018), 'Variable screening for high dimensional time series', *Electronic Journal of Statistics* **12**(1), 667–702.

Yuen, C. (2019), 'Exploiting disagreements between high-dimensional variable selection for uncertainty visualization', *LSE Department of Statistics, PhD presentation event* .