Detecting Sparse Cointegration[*]

JESÚS GONZALO† and JEAN-YVES PITARAKIS‡

†*Department of Economics, Universidad Carlos III de Madrid, C/Madrid 126, 28903 Getafe (Madrid), Spain*
(e-mail: jesus.gonzalo@uc3m.es)

‡*Department of Economics, University of Southampton, UK*
(e-mail: j.pitarakis@soton.ac.uk)

January 22, 2025

Abstract

We propose a two-step procedure to detect cointegration in high-dimensional settings, focusing on sparse relationships. First, we use the adaptive LASSO to identify the small subset of integrated covariates driving the equilibrium relationship with a target series, ensuring model-selection consistency. Second, we adopt an information-theoretic model choice criterion to distinguish between stationarity and nonstationarity in the resulting residuals, avoiding dependence on asymptotic distributional assumptions. Monte Carlo experiments confirm robust finite-sample performance, even under endogeneity and serial correlation.

Keywords: Cointegration, High Dimensional Data, Adaptive LASSO, Unit Roots.

JEL: C32, C52.

[*] Jesus Gonzalo gratefully acknowledges financial support from the *Spanish Ministerio de Ciencia e Innovacion*: PID2019-104960GB-IOO; TED2021-129784B-IOO; CEX2021-001181-M and *Spanish Ministerio de Ciencia, Innovacion y Universidades*: PID2023-147593NB-IOO, financed by MICIU/AEI /10.13039/501100011033. Jean-Yves Pitarakis gratefully acknowledges financial support from the UK *Economic and Social Research Council* through grant ES/W000989/1.

1. Introduction

Understanding long-run equilibrium relationships linking economic or financial variables is central to economic modelling. In practice, such relationships result in time-series moving together due to the presence of common trends in their dynamics. This phenomenon is referred to as cointegration. Cointegration analysis, which determines whether a group of nonstationary variables share common stochastic trends has traditionally been conducted in low-dimensional settings. Classical methods for detecting cointegration like the Engle-Granger two-step procedure (Engle and Granger (1987)) or Johansen's maximum likelihood approach (Johansen (1991)) are well-suited for environments with few variables. However, with the advent of high-dimensional data where the number of series of interest to an investigation may be large traditional methods become inadequate for identifying cointegrating relationships or testing for their presence. In high-dimensional settings, standard estimation techniques like least squares often become unstable, even when computationally feasible. Moreover, the lack of oracle knowledge makes it challenging to identify a small subset of cointegration-inducing variables from a large pool of candidates.

The goal of this paper is to propose a simple method to test for the presence of cointegration among a large pool of I(1) candidate series, particularly in cases where cointegration, if present, is assumed to be sparse and one wishes to uncover these cointegration-inducing covariates. The environment is that of a single equation cointegration setting where a target series of interest potentially cointegrates with a small number of series from a large pool of candidates, and which the investigator wishes to detect. This covariate selection stage is particularly important as the resulting residuals will be used to assess the presence or absence of cointegration. Consider the problem of determining whether the stock price of a specific constituent of a stock market index, such as the FTSE100 or S&P500, cointegrates with other stocks in the index. This question holds important implications for portfolio diversification, as cointegrated stocks exhibit stable long-run relationships that mitigate risks. Similarly, it may be of interest to determine whether stocks within similar industries share a common stochastic trend. Such analyses require cointegration methods capable of handling a large number of covariate candidates to detect meaningful relationships. The sparse nature of cointegration that we operate under reflects the idea that most variables in high-dimensional economic and financial data will be irrelevant to any given cointegrating relationship. For example, while a stock's price may cointegrate with a handful of other stocks in the market (e.g., stocks in related industries), it is unlikely to share such a relationship with all 99 other constituents in an index like the FTSE100. Identifying this sparse active subset of predictors is therefore essential for meaningful subsequent analysis.

A key challenge in this context therefore, is accurately identifying the small subset of integrated, I(1), series that are linked through a cointegrating relationship, particularly when selecting from a large pool of candidate series. We address this by proposing an estimator based on the adaptive LASSO. This method offers both computational efficiency and model-selection consistency, asymptotically identifying the true subset of series driving cointegration. Furthermore, this estimator is also shown to deliver slope parameter estimates that are super-consistent.

A second challenge is to establish whether the residuals obtained from this adaptive LASSO based procedure are I(1) or I(0). Although one may be inclined to invoke existing techniques such as ADF type unit root tests (Engle and Granger (1987), Engle and Yoo (1987), Phillips and Ouliaris (1990)) or KPSS type stationarity tests (Kwiatowski et al. (1992), Shin (1994)), both of these face limitations that tend to amplify in high dimensional contexts (e.g., non-standard limiting distributions that depend on the number of fitted covariates requiring model specific tabulations). In this paper we depart from these testing based methods and propose an information-theoretic approach that avoids reliance on asymptotic distribution-based inferences. An important additional advantage of such an approach is its robustness to phenomena such as endogeneity and serial correlation while also being immune to the the number of fitted covariates. The viewing of inferences about stationarity and non-stationarity as a model selection problem has been explored in the context of unit-root detection in Phillips and Ploberger (1996), Phillips (2008) among others and in the context of vector error correction models in Gonzalo and Pitarakis (1998).

This work adds to the growing research on high-dimensional estimation and inference in nonstationary settings. Incorporating modern high-dimensional statistical methods into time series analysis is particularly important for understanding economic data. However, the combination of high dimensionality and nonstationarity presents unique challenges. High dimensionality can lead to misleading results in nonstationary environments, as highlighted by studies like Gonzalo and Pitarakis (1999, 2021) and Onatski and Wang (2018). Recent research has developed theoretical tools to address these challenges in high-dimensional contexts involving I(1) processes and cointegration. For example, Kock (2016) studied the properties of the adaptive LASSO in autoregressive models with unit roots. Koo et al. (2020) and Lee, Shi, and Gao (2022) established precise limiting distributions for LASSO-based estimators in high-dimensional predictive regressions with unit-root covariates. Closer to our approach, Smeekes and Wijler (2021) introduced a penalized error correction model and proposed a LASSO-based method for estimating its parameters. Building on this literature, our paper presents a simple, practical method for identifying cointegrating relationships in settings where the pool of candidate variables may be large.

The remainder of the paper is structured as follows. Section 2 introduces the theoretical framework. Section 3 introduces the two-steps approach to detecting cointegration and obtains its theoretical properties. Section 3 evaluates the performance of the proposed procedures using Monte Carlo simulations. Section 4 discusses their practical implementation and Section 5 concludes.

2. Theoretical Framework

We consider a single-equation cointegration setting where the target variable y_t may cointegrate with a subset of predictors drawn from a large pool of p I(1) series. The operating model is given by:

$$y_t = \beta_0 + \boldsymbol{x}_t'\boldsymbol{\beta} + z_t, \quad t = 1, \dots, n, \tag{1}$$

where $y_t \in \mathbb{R}$ is the target variable, $\boldsymbol{x}_t = (x_{1t}, x_{2t}, \dots, x_{pt})' \in \mathbb{R}^p$ is a p-dimensional vector of covariates, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)' \in \mathbb{R}^p$ is the vector of unknown slope coefficients, $\beta_0 \in \mathbb{R}$ is the intercept term, and $z_t \in \mathbb{R}$ represents the deviation from the equilibrium relationship.

The covariates $\boldsymbol{x}_t$ are modeled as I(1) processes, possibly correlated across dimensions and with serially correlated disturbances:

$$x_{jt} = x_{j,t-1} + v_{jt}, \quad j = 1, \dots, p, \quad t = 1, \dots, n. \tag{2}$$

The deviation term z_t is modelled as

$$z_t = \rho z_{t-1} + u_t, \quad t = 1, \dots, n. \tag{3}$$

Letting $\boldsymbol{\eta}_t = (u_t, v_{1t}, \dots, v_{pt})'$, we model these $p+1$ disturbance series as

$$\boldsymbol{\eta}_t = \boldsymbol{C}(L)\ \boldsymbol{e}_t \tag{4}$$

where $\boldsymbol{C}(L) = \sum_{i=0}^{\infty} C_i e_{t-i}$ for $\sum_{i=0}^{\infty} i|\boldsymbol{C}_i|< \infty$, $\boldsymbol{C}_0 = I$ and $\boldsymbol{e}_t \sim i.i.d.(0, \boldsymbol{\Sigma}_e)$ with $E\|\boldsymbol{e}\|^{2+\delta}< \infty$ for some positive δ. These assumptions are standard in this literature and essentially ensure that an FCLT holds for the $\boldsymbol{\eta}_t$ sequence. Under cointegration, the existence of a long-run equilibrium relationship implies that z_t is stationary with $|\rho|< 1$, even though both y_t and the components of $\boldsymbol{x}_t$ are individually I(1) processes. If $\rho = 1$ instead, (1) is viewed as a spurious regression.

To formalise the notion of sparse cointegration, let

$$S = \{j \colon \beta_j \neq 0\}, \quad j = 1, \dots, p \tag{5}$$

denote the set of active covariates inducing cointegration, with $|S|= s \ll p$. The remaining covariates with $\beta_j = 0$ are irrelevant for the cointegrating relationship. The notion of sparse cointegration posits that the target variable y_t is cointegrated with only a small number of covariates among the large pool of p candidates. Formally, this is expressed by assuming that the cardinality of the active set S is much smaller than p, i.e., $s = |S| \ll p$. This assumption reflects the realistic scenario where most covariates are irrelevant for the equilibrium relationship, as often encountered in applications with high-dimensional economic and financial data. Sparse cointegration implies that $\boldsymbol{\beta}$ is a sparse vector, where:

$$\beta_j \neq 0 \quad \text{if and only if } j \in S, \quad \beta_j = 0 \quad \text{for } j \notin S. \tag{6}$$

In line with the above description of the notion of sparse cointegration, we let the $s \times 1$ vector $\boldsymbol{\beta}_S$ collect the parameters whose covariates induce cointegration, when the latter is present. Similarly we let $\boldsymbol{x}_{S,t}$ denote the $s - vector$ of active covariates associated with $\boldsymbol{\beta}_S$. The parameters associated with the variables that do not actively enter (1) are in turn collected in the $(p - s)$ vector $\boldsymbol{\beta}_{S^c}$ while $\boldsymbol{x}_{S^c,t}$ collects the $(p - s)$ inactive series.

Our goal is to assess whether (1) is truly a cointegrating relationship by analyzing the stationarity of z_t when p is large and only a small unknown number of these series induce cointegration, if the latter is truly present. If sparse cointegration is present for instance we expect the residuals from the oracle regression $y_t = \beta_0 + \boldsymbol{x}'_{S,t}\boldsymbol{\beta}_S + z_t$ to behave like an I(0) process which our approach will be designed to detect.

Throughout this paper our theoretical analyses will operate under a fixed p setting with $n \to \infty$. This has been the norm in the context of adaptive LASSO estimation and other developments in this literature (e.g., Zou (2016), Kock (2012)) and does not preclude the treatment of high dimensionality while ruling out ultra-high dimensional settings under which p may exceed the sample size n. When dealing with p candidate predictors, even moderately large values of p result in an impractically large number of 2^p possible models, making conventional model selection methods impractical. It is also well known that OLS becomes highly instable when p/n is large even under $p < n$. Finally, and perhaps most importantly, the residuals obtained from a large number of estimated slope parameters via least-squares for instance, will accumulate large estimation errors and are unlikely to result in reliable inferences.

3. Estimation and Testing

We initially motivate and describe the use of the adaptive LASSO as a method to perform model-selection and estimation in a single shot within the cointegrating regression model in (1).The adaptive LASSO estimator of the model parameters is given by

$$(\hat{\beta}_0^{AL}, \hat{\boldsymbol{\beta}}^{AL}) = \arg\min_{\beta_0,\boldsymbol{\beta}} \left\{ \sum_{t=1}^{n} (y_t - \beta_0 - \boldsymbol{x}_t'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} w_j |\beta_j| \right\}. \tag{7}$$

where λ_n is a regularization parameter (penalty term) and the w_j's are the adaptive weights associated with β_j. Specifically, given an initial estimator $\widetilde{\beta}_j$ (e.g., OLS) these weights are set as $w_j = 1/|\widetilde{\beta}_j|^{\gamma}$ with $\gamma > 0$.

The adaptive LASSO is particularly well-suited to the cointegration setting considered in this paper, as it addresses several limitations of the standard Lasso. In high-dimensional regression problems, the standard LASSO suffers from shortcomings that are particularly detrimental to our goal of consistently estimating the residuals of a sparse cointegrating regression. Unless strong assumptions are imposed (e.g., irrepresentable condition) it does not satisfy the oracle property, which guarantees asymptotic consistency and correct model selection. The adaptive Lasso resolves these issues by introducing data-dependent weights that penalize small coefficients more heavily, allowing for consistent estimation of large coefficients while effectively shrinking irrelevant predictors to zero. Note for instance that if $\widetilde{\beta}_j$ is near zero, this results in very large associated $w_j's$, effectively penalizing parameters associated with variables that are unlikely to be active. Unlike the standard Lasso, which imposes the same penalty on all coefficients regardless of their magnitude, the adaptive Lasso adjusts the penalty weights based on an initial estimate of the coefficients, typically obtained via an OLS or ridge regression. By applying smaller penalties to coefficients with larger initial estimates, the adaptive Lasso allows these coefficients to converge more accurately to their true values, while continuing to shrink the irrelevant coefficients to zero. This feature makes the adaptive LASSO ideal for constructing consistent residuals $\hat{z}_t$, which are crucial for the subsequent test that we develop.

Locating the optima of a function such as (7) is a convex optimisation problem which guarantees that any local minimum is also a global minimum. The structure of the program however is challenging due to the inclusion of the ℓ_1 norm penalty in the objective function. Note for instance that although the ℓ_1 norm is convex, it is non-differentiable at points where $\beta_j = 0$. This creates difficulties for standard optimisation techniques such as gradient descent which relies on smoothness. Instead, coordinate descent type of algorithms which avoid the need to compute a full gradient at non-differentiable points are typically considered.

Given $\hat{\boldsymbol{\beta}}^{AL}$ estimated from (7), we write

$$\hat{S} = \{j \colon \hat{\beta}_j^{AL} \neq 0\} \tag{8}$$

for the active set of predictors selected by the adaptive LASSO and in the sequel refer to the covariates associated with this estimated active set as $\boldsymbol{x}_{\hat{S}t}$. We next use this estimated active set $\hat{S}$ to form the residuals of interest. These can be obtained using the adaptive lasso estimates directly:

$$\hat{z}_t = y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}^{AL} \tag{9}$$

noting that the vast majority of the components of $\hat{\boldsymbol{\beta}}^{AL}$ are set to zero, by the effect of the ℓ_1 penalization. Inferences about the presence or absence of cointegration can now be implemented using the residual sequence in (9). Alternatively, we may also consider running a post adaptive lasso OLS by regressing y_t on the selected set $\boldsymbol{x}_{\hat{S},t}$.

Given the residual sequence $\hat{z}_t$ constructed as above, our next goal is to assess whether or not it contains a unit root in its autoregressive representation. For this purpose consider the following two competing models:

$$\Delta\hat{z}_t = \mu + \sum_{j=1}^{k} \phi_j \Delta\hat{z}_{t-j} + \epsilon_t \quad \text{model} \ \ \mathcal{M}_0 \tag{10}$$

$$\Delta\hat{z}_t = \mu + \phi \ \hat{z}_{t-1} + \sum_{j=1}^{k} \phi_j \Delta\hat{z}_{t-j} + \epsilon_t \quad \text{model} \ \ \mathcal{M}_1. \tag{11}$$

Model $\mathcal{M}_0$ imposes I(1) behaviour by fitting an AR(k) process in first differences to the residuals. In contrast, model $\mathcal{M}_1$ allows the $\hat{z}_t' s$ to be stationary. Because the $\hat{z}_t' s$ are centered by construction, the inclusion of an intercept term is not crucial. The order k of the autoregressions is directly linked to the behaviour of the error process u_t driving the equilibrium errors as formulated in (3).

We now view the objective of testing for cointegration as a model selection problem between models $\mathcal{M}_0$ and $\mathcal{M}_1$. Selecting model $\mathcal{M}_0$ indicates absence of cointegration (i.e., (1) is a spurious regression). In contrast, support for model $\mathcal{M}_1$ implies that the $\hat{z}_t' s$ behave like an I(0) process so that y_t and the set of covariates selected by the adaptive lasso are cointegrated.

Letting $\overline{\sigma}_0^2$ and $\overline{\sigma}_1^2$ denote the residuals from (10) and (11), selection between the two models is made using the criteria

$$IC_0 = \ln \overline{\sigma}_0^2 + \frac{c_n}{n} \ (k+1) \tag{12}$$

$$IC_1 = \ln \overline{\sigma}_1^2 + \frac{c_n}{n} \ (k+2) \tag{13}$$

where c_n is a deterministic penalty term and $k \in \{0, 1, \ldots, k_{max}\}$. The proposed model-selection based approach is based on comparing IC_0 with IC_1 and leads to choosing $\mathcal{M}_0$ if $IC_0 \leq IC_1$ and to choosing $\mathcal{M}_1$ if $IC_0 > IC_1$. Note that this model selection based approach bears strong resemblance with a likelihood ratio type test since the requirement $IC_0 > IC_1$ for *rejecting* $\mathcal{M}_0$ reduces to $n \ln \overline{\sigma}_0^2 / \overline{\sigma}_1^2 > c_n$. Here the deterministic penalty term c_n plays a similar role to the critical values used to form the rejection region of likelihood ratio type statistics.

REMARK 1. In (10) and (11) the order k of the fitted autoregressions is meant to capture the potential presence of serial correlation in the $u_t's$ driving the equilibrium errors. This augmentation is fundamental for achieving nuisance parameter free asymptotics when implementing test based inferences but is much less important in our model selection context. Indeed, in large enough samples the ability of our proposed model selection based approach for distinguishing between $\mathcal{M}_0$ and $\mathcal{M}_1$ is unaffected by the presence of serial correlation in the $u_t's$. As pointed out earlier for instance, the model selection based approach leads to *rejecting* $\mathcal{M}_0$ when $n \ln \overline{\sigma}_0^2/\overline{\sigma}_1^2 > c_n$. Heuristically, if $c_n \to \infty$, correct decisions will typically be ensured provided that $n \ln \overline{\sigma}_0^2/\overline{\sigma}_1^2$ is $O_p(1)$ regardless of whether its limit is characterized by nuisance parameters induced by serial correlation and/or endogeneity. Nevertheless, the inclusion of the right number of lags will favourably influence finite sample properties.

3.1. Theoretical Properties

In a first instance we aim to obtain the model selection consistency of the adaptive LASSO in the cointegrated regression setting. The ability of the adaptive lasso to uncover the true sparse structure of cointegration is key to obtaining residuals that accurately mimic the true underlying equilibrium errors when cointegration is present. Note that within our context, we will solely be concerned with obtaining consistent estimates of $\boldsymbol{\beta}$ rather than the explicit limiting distributions of $\hat{\boldsymbol{\beta}}^{AL}$.

Proposition 1 below summarizes the adaptive LASSO's ability to identify the true sparse cointegration structure. It is here implicitly assumed that our results hold under the theoretical framework introduced in Section 2 and with $|\rho|< 1$ so that (1) is truly a cointegrating regression.

Proposition 1 Under the conditions $\lambda_n/n \to 0$ and $\lambda_n/n^{1-\gamma} \to \infty$ on the penalty term of the adaptive lasso estimator and $|\rho|< 1$ (cointegrated regression), we have

$$\lim_{n\to\infty} P(\hat{S} = S) = 1. \tag{14}$$

Proposition 1 thus establishes that, under cointegration, the adaptive LASSO asymptotically selects the correct subset of cointegration-inducing covariates. As part of the proof of Proposition 1, it is also shown that the resulting adaptive lasso parameter estimates are n-consistent in the sense that $n(\hat{\boldsymbol{\beta}}_S^{AL} - \boldsymbol{\beta}_S^0) = O_p(1)$ and $n(\hat{\boldsymbol{\beta}}_{S^c}^{AL} - \mathbf{0}) = O_p(1)$. An immediate implication of these results is that, when $|\rho|< 1$ and the model is estimated via the adaptive lasso, we should expect $\hat{z}_t \sim I(0)$.

It is now also instructive to examine the model selection properties of the adaptive lasso in a spurious regression context where the $z_t's$ are no longer equilibrium errors, behaving like an I(1) process instead.

Proposition 2 Under the conditions $\lambda_n/n \to 0$ and $\lambda_n/n^{1-\gamma} \to \infty$ on the penalty term of the adaptive lasso estimator and $\rho = 1$ (spurious regression), we have

$$\lim_{n\to\infty} P(\hat{S} = \{1, 2, \ldots, p\}) = 1. \tag{15}$$

The result in Proposition 2 indicates that the adaptive LASSO tends to include all predictors in the pool when the true model is spurious, leaving none of the slope parameters at zero asymptotically The intuition follows from the construction of the adaptive lasso weights, which are formed using initial least-squares estimates $\widetilde{\beta}_j$. In a spurious regression setting these estimates converge in distribution and satisfy $\widetilde{\beta}_j = O_p(1)$. Thus the adaptive lasso weights $w_j = 1/|\widetilde{\beta}_j|^\gamma$ are bounded, regardless of the underlying magnitude of the $\beta_j's$. As a result (see the proof of Proposition 2), the adaptive lasso effectively treats β_j as unpenalized, so no shrinkage is applied to any parameters. By analogy with Proposition 1, this also implies that under a spurious regression, the adaptive lasso residuals follow an I(1) process, so that $\Delta\hat{z}_t \sim I(0)$.

We next focus on detecting whether the estimated $\hat{z}_t's$ are consistent with an I(0) or an I(1) process. Recall from (12)-(13) that our model selection based approach points to $\mathcal{M}_0$ if $IC_0 < IC_1$ and to $\mathcal{M}_1$ otherwise. Proposition 3 establishes that such a selection process is model selection consistent asymptotically provided that the deterministic penalty term c_n satisfies suitable requirements.

Proposition 3 As $n \to \infty$, *(i)* $P(IC_0 < IC_1 \mid \mathcal{M}_0) \to 1$ *if* $c_n \to \infty$ *and (ii)* $P(IC_1 < IC_0 \mid \mathcal{M}_1) \to 1$ *if* $c_n/n \to 0$.

Part (i) of Proposition 3 establishes that in sufficiently large samples, model $\mathcal{M}_0$ will be correctly selected provided that the penalty term satisfies $c_n \to \infty$. Similarly, part (ii) shows that if the $\hat{z}_t's$ are an I(0) process (i.e., there is cointegration) the model selection based approach will correctly point to $\mathcal{M}_1$ provided that $c_n/n \to 0$. Evidently, a criterion such as the BIC with $c_n = \ln n$ satisfies both of these requirements and is therefore model selection consistent.

4. Implementation

Penalty Term in the adaptive lasso

The implementation of the adaptive LASSO requires choosing the penalty parameter λ_n and setting the adaptive weights w_j. As it is common in the model selection literature, the theoretical requirements on the penalty term λ_n are such that a multitude of valid choices can be considered in practice. In the context of a unit-root setting, Kock (2016) proposed to select λ_n via the BIC criterion. This entails estimating the models via the adaptive LASSO across a grid of λ_n values and picking the one that minimizes the BIC criterion, say $\hat{\lambda}_n^{bic}$. Here we adopt a similar approach. Letting $RSS(\lambda_n) = \sum_{t=1}^n (y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}^{AL})^2$ denote the residual sum of squares associated with the use of a penalty parameter λ_n we introduce

$$\text{BIC}(\lambda_n) = n \ln \left(\frac{\text{RSS}(\lambda_n)}{n} \right) + k(\lambda_n) \; \ln n \tag{16}$$

where $k(\lambda_n)$ is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}^{AL}(\lambda_n)$. The data-based penalty we use in our adaptive lasso implementation is then

$$\hat{\lambda}_n^{bic} = \arg\min_{\lambda \in \Lambda} \{ n \ln \left(\frac{\mathrm{RSS}(\lambda_n)}{n} \right) + k(\lambda_n) \ \ln n. \} \tag{17}$$

The practical implementation of (17) requires choosing a suitable range for λ for which we create a grid of values ranging from a magnitude near zero up to ten (e.g., $\Lambda = \mathit{logspace}(-1.5, 0.5, 10)$ in a matlab based implementation).

Adaptive Lasso weights

Regarding the adaptive weights w_j, we follow the literature by basing them on the standard least squares estimates of the $\beta_j' s$ in the original cointegrating regression. Such estimates are known to be n-consistent so that for $j \in S$ we will have $w_j \xrightarrow{p} 1/|\beta_j|^\gamma$ while for $j \notin S$ we will have $w_j \xrightarrow{p} \infty$. This highlights the fact that parameters whose least squares estimates are near zero will receive a strong penalization shrinking them exactly to zero.

Finally the γ exponent is most commonly set at $\gamma = 1$ or $\gamma = 2$ so that the conditions on the adaptive lasso penalty become $\lambda_n/n \to 0$ and $\lambda_n \to \infty$ or $\lambda_n/n \to 0$ and $n\lambda_n \to \infty$ for $\gamma = 1$ and $\gamma = 2$ respectively. In the sequel our experiments consider both of these scenarios. Recall that γ controls the strength of the adaptation in the adaptive lasso. A larger γ induces a more aggressive penalization of coefficients with small initial estimates. This allows for a slower growth rate of λ_n for achieving model selection consistency. A smaller γ implies a less aggressive penalization, requiring a faster growth rate of λ_n.

Optimization algorithm

The convex optimization programme in (7) is implemented via matlab's lasso function modified to account for the adaptive weights. This is achieved by scaling the regressors as $\widetilde{x}_j = x_j w_j^{-1}$ and modifying the design matrix as $\widetilde{X} = X \ \mathrm{diag}(w_j^{-1})$. The resulting plain lasso estimates, say $\hat{\boldsymbol{\beta}}^{plain}$ can then be mapped back to obtain the adaptive lasso counterparts

$\hat{\boldsymbol{\beta}}^{AL} = \hat{\boldsymbol{\beta}}^{plain} \ \text{diag}(w_j)$.

Choosing the lag order k in (10)-(11)

In line with most of the literature on lag length selection in ADF type regressions, we estimate k via a BIC criterion implemented on the fitted model in (11) using a given upper bound k_{max}. With $\hat{k}_{bic}$ denoting such an estimate, the ensuing model selection based approach for selecting between $\mathcal{M}_0$ and $\mathcal{M}_1$ is implemented using this $\hat{k}_{bic}$ when fitting (10)-(11) via least squares. Note that in the present context the accuracy of $\hat{k}_{bic}$ is not fundamental for achieving model selection consistency when comparing models $\mathcal{M}_0$ and $\mathcal{M}_1$. This is because the selection consistency result of Proposition 3 does not require serially uncorrelated $\epsilon_t' s$ in (10)-(11).

5. Finite Sample Experiments

Our simulations are structured as follows. In a first instance we document the model selection ability of the adaptive LASSO by generating samples from a sparsely cointegrated DGP and reporting the averaged false positive rates (FPR), false negative rates (FNR) and false discovery rates (FDR) across replications.

The FPR quantifies the proportion of irrelevant covariates (true negatives) that are incorrectly selected as relevant (false positives) by the adaptive LASSO algorithm. A high FPR implies that the adaptive LASSO is selecting too many inactive covariates (overfitting). Similarly, the FNR quantifies the proportion of truly relevant covariates (true positives) that are incorrectly excluded by the adaptive LASSO algorithm (underfitting). Model selection consistency implies that these metrics should be small and converge to zero for large n, indicating that the adaptive LASSO applied to the cointegrating regression selects the correct covariates and discards the inactive covariates. A third metric we also report is the false discovery rate (FDR) which assesses the proportion of irrelevant regressors that have been selected relative to the total number of selected covariates. It aims to assess how the algorithm balances between discovering relevant covariates and limiting the selection of irrelevant ones.

REMARK 2: Although metrics such as the FPR, FNR or FDR are commonly used when assessing covariate selection methods caution should be exercised when interpreting these rates in the context of highly imbalanced datasets. In the context of our sparse cointegration setting for instance, positives are rare compared to negatives (e.g., under $p = 100$ with $|s|= 5$ for instance, 5 out of 100 covariates are true positives while the remainder 95 are true negatives) so that a low FPR may coincide with a particularly large FDR. This would be a common occurrence in datasets where negatives vastly outnumber positives. This imbalance means that even a small number of false positives can lead to a high proportion of incorrect positive predictions (high FDR). Intuitively, the fact that positives are rare, if a method is associated with a low FNR it is likely that its FDR will be substantial. To illustrate suppose that $p = 100$ and that there are $s = 5$ active covariates. Furthermore, suppose that the adaptive lasso points to 7 active covariates (matching the correct 5 and in addition, 2 unnecessary ones). The FDR would be $\text{FDR} = \text{FP}/(\text{TP} + \text{FP}) = 2/(2 + 5) \approx 29\%$ i.e., about 29% of the covariates identifed as active are actually false positives. The FPR for this example is $\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) = 2/(2 + 93) \approx 2\%$ i.e., approximately 2% of the actual negatives were incorrectly identified as positives.

Our points under Remark 2 suggest that the weight given to these different metrics must be context dependent. In the setting considered in this paper, obtaining residuals $\hat{z}_t$ that mimic their true counterparts is particularly important and for this purpose, avoiding false negatives so that one does not face omitted relevant variables is perhaps more important than the inefficiencies caused by the inclusion of irrelevant covariates, provided that the latter are not large in number.

Once we have documented the model selection consistency of the adaptive LASSO we subsequently focus on the finite sample properties of our model-selection based approach designed to distinguish between I(0) and I(1) residuals. For this purpose we report correct decision frequencies associated with both a cointegrated DGP (i.e., $|\rho|< 1$) and a non-cointegrated one (i.e., $\rho = 1$).

5.1. DGP Parameterizations

We consider the following DGP:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1t} + \ldots + \beta_p x_{pt} + z_t \\ x_{j,t} &= x_{j,t-1} + v_{j,t} \\ z_t &= \rho z_{t-1} + e_t \end{aligned} \tag{18}$$

where we set the degree of sparsity to 5 active covariates ($\beta_j \neq 0$ in (18)), i.e., $|S|= 5$ and $|S^c|= p - 5$ and consider $p \in \{10, 50, 100\}$. The five active covariates are taken as the first five of the $x'_{jt}s$ (i.e, $j = 1, \ldots, 5$). We experiment with two alternative parameterizations of $\boldsymbol{\beta}_S$ corresponding to strong and weaker signals. These are given by $\boldsymbol{\beta}_S = (1, 0.5, 1.5, 0.8, 1)'$ and $\boldsymbol{\beta}_S = (0.25, 0.25, 0.25, 0.25, 0.25)'$ respectively. For the random disturbances we take $\boldsymbol{\eta}_t = (e_t, v_{1,t}, \ldots, v_{p,t}) \sim \text{NID}(0, \boldsymbol{\Omega})$ where

$$\boldsymbol{\Omega}_{p+1\times p+1} = \begin{pmatrix} \sigma_e^2 & \boldsymbol{\omega}'_{ev} \\ \boldsymbol{\omega}_{ev} & \boldsymbol{\Omega}_{vv} \end{pmatrix} \tag{19}$$

with $\boldsymbol{\omega}'_{ev} = (\sigma_{ev_1}, \ldots, \sigma_{ev_p})$ and $\boldsymbol{\Omega}_{vv} = E[\boldsymbol{v}_t \boldsymbol{v}'_t]$. We set $\sigma_e^2 = 4$, $[\Omega_{vv}]_{i,j} = 0.5^{|i-j|}$, and $\boldsymbol{\omega}'_{ev} = \tau \ \min(\text{diag}(\boldsymbol{\Omega}_{vv}))\mathbf{1}_{1\times p}$. This non-zero nature of $\boldsymbol{\omega}_{ev}$ captures the presence of endogeneity provided that $\tau \neq 0$, and our experiments set $\tau = 1$ throughout.

Despite its simplicity this DGP captures all the key phenomena that have been the focus of attention in this literature. Namely, serial correlation and endogeneity.

5.2. Model selection properties of the Adaptive LASSO

In a first instance, we focus our attention on the model selection ability of the adaptive LASSO by documenting its ability to pick the correct $s = 5$ active series from the pool of p candidate series. Results from this set of experiments are summarized in Tables 1-2 below which present the FPR, FNR and FDR across different parameter configurations, and obtained as averages across replications. Table 1 focuses on the scenario with strong signals as described above and Table 2 focuses on the DGP with weaker signals. Note that the bottom panels of both of these Tables correspond to the spurious regression scenario ($\rho = 1$) which we include in order to illustrate our result in Proposition 2.

Table 1: Adaptive LASSO Model Selection - Strong signals

	$p = 10$			$p = 50$			$p = 100$		
	FPR	FNR	FDR	FPR	FNR	FDR	FPR	FNR	FDR
$\boldsymbol{\rho = 0.00}$									
$\gamma = 1$	0.031	0.000	0.025	0.049	0.003	0.217	0.037	0.015	0.289
$\gamma = 2$	0.018	0.000	0.014	0.043	0.003	0.199	0.041	0.022	0.302
$\boldsymbol{\rho = 0.50}$									
$\gamma = 1$	0.205	0.001	0.147	0.383	0.014	0.768	0.357	0.025	0.870
$\gamma = 2$	0.189	0.001	0.138	0.373	0.021	0.766	0.345	0.029	0.868
$\boldsymbol{\rho = 1.00}$									
$\gamma = 1$	0.799	0.114	0.470	0.996	0.000	0.900	0.899	0.000	0.944
$\gamma = 2$	0.775	0.128	0.466	0.941	0.000	0.894	0.498	0.000	0.904

Table 2: Adaptive LASSO Model Selection - Weak signals

	$p = 10$			$p = 50$			$p = 100$		
	FPR	FNR	FDR	FPR	FNR	FDR	FPR	FNR	FDR
$\boldsymbol{\rho = 0.00}$									
$\gamma = 1$	0.044	0.031	0.038	0.115	0.170	0.487	0.096	0.283	0.648
$\gamma = 2$	0.035	0.040	0.031	0.136	0.230	0.560	0.122	0.361	0.738
$\boldsymbol{\rho = 0.50}$									
$\gamma = 1$	0.236	0.133	0.191	0.394	0.227	0.815	0.366	0.297	0.906
$\gamma = 2$	0.248	0.163	0.208	0.388	0.281	0.825	0.351	0.344	0.909
$\boldsymbol{\rho = 1.00}$									
$\gamma = 1$	0.805	0.172	0.492	0.996	0.000	0.900	0.901	0.000	0.944
$\gamma = 2$	0.774	0.203	0.492	0.941	0.000	0.894	0.501	0.020	0.906

Looking first at Table 1, which focuses on strong signals, we see clear distinctions across different parameter settings. When $\rho = 0$ (i.e., cointegration residuals behave like noise), the adaptive LASSO often achieves very low FPR and FNR for smaller p. For instance, at $p = 10$ the FPR is frequently near or below 0.03, and the FNR is close to 0.00 in many instances.

This indicates that the adaptive LASSO is quite effective at identifying relevant features that induce cointegration, without including too many spurious ones.

As p grows from 10 to 100, we see that the FDR tends to rise. This makes sense: with more variables in the model, even modest false positives can lead to higher false discovery rates. Raising the adaptive LASSO tuning parameter γ from 1 to 2 generally leads to smaller FPR and FDR, as the penalty more agressively shrinks coefficients of non-informative variables.

Under $\rho = 0.5$ (more memory in the cointegrated residuals), we observe elevated FPRs - sometimes as high as 0.37 or 0.38 for $p = 50$ or 100. Thus a higher ρ which corresponds to a relatively weaker cointegration strength makes it harder to distinctly isolate the true signals, thus driving up the probability of including false positives. Nevertheless, the overall picture that comes across from the top two panels of Table 1 is that the adaptive LASSO is quite effective in model selection in cointegrated regressions. The outcomes associated with $\rho = 1$ also clearly support our result in Proposition 2. The FPR's jump up significantly (e.g., $FPR = 0.996$ under $p = 50$ and for $\gamma = 1$) illustrating the fact that in a spurious regression context the adaptive LASSO is essentially similar to performing least squares on the entire set of predictors.

Turning to Table 2, we now examine the scenario with weak signals. In this more challenging setting, the adaptive LASSO experiences higher FNRs, as weak true signals are more likely to be missed. When $\rho = 0$ and $p = 10$, performance is relatively stable ($FPR < 0.05$, $FNR < 0.04$), but as soon as p grows to 50 or 100, or ρ increases to 0.5, the selection process becomes more error-prone. Overall however we view these outcomes as very favorable. They suggest that the proposed adaptive LASSO works particularly well in cointegrated context where predictors are $I(1)$ time series. Moreover, the method also appears to be reliable under a fairly strong degree of endogeneity used in our DGPs. This is in line with standard least squares based estimation of cointegrated regressions where endogeneity is known not to influence estimate consistency.

Table 3 presents replication averages of adaptive LASSO coefficient estimates. Overall, these outcomes reveal that these estimates show low bias for large or moderate true coefficients. Smaller coefficients (weak signals) are more prone to bias, especially in scenarios with more features.

Table 3: Adaptive LASSO Estimator Properties

	Strong signals				Weak signals		
	$p = 10$	$p = 50$	$p = 100$		$p = 10$	$p = 50$	$p = 100$
$\boldsymbol{\beta}_0$	$E[\hat{\boldsymbol{\beta}}^{AL}]$	$E[\hat{\boldsymbol{\beta}}^{AL}]$	$E[\hat{\boldsymbol{\beta}}^{AL}]$	$\boldsymbol{\beta}_0$	$E[\hat{\boldsymbol{\beta}}^{AL}]$	$E[\hat{\boldsymbol{\beta}}^{AL}]$	$E[\hat{\boldsymbol{\beta}}^{AL}]$
				$\gamma = 1$			
$\boldsymbol{\rho = 0.0}$				$\boldsymbol{\rho = 0.0}$			
0.50	0.494	0.427	0.368	0.25	0.253	0.235	0.180
1.50	1.516	1.556	1.587	0.25	0.247	0.223	0.215
0.80	0.802	0.779	0.751	0.25	0.249	0.240	0.228
1.00	1.012	1.006	0.998	0.25	0.247	0.233	0.206
0.00	0.002	0.002	0.002	0.25	0.246	0.203	0.178
$\boldsymbol{\rho = 0.5}$				$\boldsymbol{\rho = 0.5}$			
0.50	0.498	0.488	0.472	0.25	0.264	0.328	0.276
1.50	1.537	1.592	1.604	0.25	0.260	0.280	0.283
0.80	0.809	0.835	0.811	0.25	0.256	0.280	0.270
1.00	1.020	1.067	1.064	0.25	0.255	0.292	0.270
0.00	0.014	0.043	0.045	0.25	0.240	0.266	0.257
				$\gamma = 2$			
$\boldsymbol{\rho = 0.0}$				$\boldsymbol{\rho = 0.0}$			
0.50	0.499	0.428	0.349	0.25	0.256	0.248	0.191
1.50	1.514	1.556	1.604	0.25	0.250	0.230	0.230
0.80	0.802	0.788	0.759	0.25	0.250	0.238	0.209
1.00	1.016	1.016	1.014	0.25	0.246	0.243	0.233
0.00	0.001	0.003	0.004	0.25	0.249	0.203	0.181
$\boldsymbol{\rho = 0.5}$				$\boldsymbol{\rho = 0.5}$			
0.50	0.500	0.500	0.477	0.25	0.272	0.342	0.280
1.50	1.531	1.610	1.634	0.25	0.255	0.293	0.278
0.80	0.822	0.835	0.835	0.25	0.263	0.298	0.305
1.00	1.019	1.087	1.080	0.25	0.271	0.285	0.291
0.00	0.012	0.050	0.048	0.25	0.236	0.285	0.262

5.3. Two-Step cointegration testing

In this section we evaluate the correct detection ability of our proposed model selection based approach for distinguishing between cointegration (i.e., $\hat{z}_t$ behaves like a stationary process) and no-cointegration (i.e., $\hat{z}_t$ behaves like an I(1) process). For the cointegration cases we experimented with $\rho \in \{0.00, 0.50\}$ while for the no-cointegration scenarios we set $\rho = 1$. Results for these experiments are summarized in Table 4 which displays the frequencies of choosing the stationary specification and unit-root specifications respectively.

It is useful to recall that the residuals have been formed via a post-adaptive LASSO least squares estimation on the predictors selected by the adaptive LASSO. For this reason the scenarios corresponding to strong and weak signals are not expected to deliver significantly different results as the parameter of importance is ρ rather than the $\beta' s$ in the cointegrating regression.

From the outcomes presented in Table 4 we note that when the true model is a cointegrated regression our proposed model selection approach points to the I(0) outcomes 100% of the times. When the true model is a spurious regression, correct decision frequencies are near 90% under $n = 500$, suggesting that the model selection based approach for distinguishing between the I(0) and I(1)'ness of the residuals is highly effective.

Table 4: Cointegration detection frequencies ($\gamma = 2$)

Strong signals	$p = 10$		$p = 50$		$p = 100$	
			True: I(0) ($\rho = 0$)			
Selected	$I(1)$	$I(0)$	$I(1)$	$I(0)$	$I(1)$	$I(0)$
$n = 250$	0.0	100.0	0.0	100.0	0.0	100.0
$n = 500$	0.0	100.0	0.0	100.0	0.0	100.0
			True: I(0) ($\rho = 0.5$)			
$n = 250$	0.0	100.0	0.0	100.0	0.0	100.0
$n = 500$	0.0	100.0	0.0	100.0	0.0	100.0
			True: I(1) ($\rho = 1.0$)			
$n = 250$	82.8	17.2	81.6	18.4	85.2	14.8
$n = 500$	87.2	12.8	88.0	12.0	86.8	13.2
Weak signals	$p = 10$		$p = 50$		$p = 100$	
			True: I(0) ($\rho = 0$)			
Selected	$I(1)$	$I(0)$	$I(1)$	$I(0)$	$I(1)$	$I(0)$
$n = 250$	0.0	100.0	0.0	100.0	0.0	100.0
$n = 500$	0.0	100.0	0.0	100.0	0.0	100.0
			True: I(0) ($\rho = 0.5$)			
$n = 250$	0.0	100.0	0.0	100.0	0.0	100.0
$n = 500$	0.0	100.0	0.0	100.0	0.0	100.0
			True: I(1) ($\rho = 1.0$)			
$n = 250$	81.0	16.0	81.0	17.9	85.1	13.9
$n = 500$	87.6	12.4	87.2	12.8	88.8	11.2

6. Summary and Conclusions

In this paper, we introduced a novel two-step procedure for detecting cointegration in high-dimensional settings. Our approach addresses the dual challenges of model selection and residual testing under the assumptions of sparse cointegration. By employing an adaptive LASSO-based methodology, we achieved model selection consistency, enabling accurate identification of active covariates that induce cointegration. This ensures that the residuals used for subsequent testing are well-behaved and closely approximate the true equilibrium errors when cointegration is present.

The second step of our methodology departs from traditional residual-based testing techniques, such as ADF or KPSS tests, which are often complicated by the presence of estimated residuals and high dimensionality. Instead, we adopted an information-theoretic model selection approach to distinguish between stationary and non-stationary residuals. This method offers significant advantages, including robustness to serial correlation, endogeneity, and the number of covariates. Additionally, our approach circumvents the need for asymptotic distribution-based inferences, which are sensitive to nuisance parameters and computationally challenging in high-dimensional contexts.

Future research could extend our framework to develop an inference theory for the parameter estimates obtained via the adaptive LASSO, providing a deeper understanding of their statistical properties. Additionally, extending the framework to accommodate multiple cointegrating relationships would open new avenues for analyzing more complex environments with several long-run equilibrium relationships.

Appendix

Notation: (i) For a vector $\boldsymbol{v} \in \mathbb{R}^p$, the sign vector is given by $sgn(\boldsymbol{v}) = (sgn(v_1), \dots, sgn(v_p))'$ where the $sgn(.)$ function is such that for any $c \in \mathbb{R}$, $sgn(c) = 1$ if $c > 0$, $sgn(c) = 0$ if $c = 0$, and $sgn(c) = -1$ if $c < 0$. (ii) The subdifferential set of $\|.\|_1$ at $\boldsymbol{v}$ is denoted $\partial\|\boldsymbol{v}\|_1 := (\partial|v_1|\dots,\partial|v_p|)'$ where $\partial|v_i| = \{1\}$ if $v_i > 0$, $\partial|v_i| = [-1,1]$ if $v_i = 0$, and $\partial|v_i| = \{-1\}$ if $v_i < 0$. Thus the subdifferential $\partial|v_i|$ is given by $sgn(v_i)$ if $v_i \neq 0$ and by $[-1,1]$ if $v_i = 0$.

Before proceeding with the proof of Proposition 1 we establish the n-consistency of the adaptive lasso estimator in Lemma A1 below. Specifically, that $n(\hat{\beta}_j^{AL} - \beta_j^0) = O_p(1)$ for $j \in S$ and $n(\hat{\beta}_j^{AL}) = O_p(1)$ for $j \in S^c$. For notational simplicity we omit the fitted intercept from this analysis. Some steps in our derivations invoke well known results from the unit-root and cointegration literature which we take as given. Specifically, under our operating assumptions stated in Section 2, the following large sample results are known to hold (see e.g., Phillips and Durlauf (1986)):

$$\frac{1}{n^2}\sum_{t=1}^{n} x_t x_t' = O_p(1) \tag{20}$$

$$\frac{1}{n\sqrt{n}}\sum_{t=1}^{n} x_t = O_p(1) \tag{21}$$

$$\frac{1}{n}\sum x_t z_t = O_p(1) \tag{22}$$

from which we also infer that

$$\begin{aligned} n(\widetilde{\beta}_j^{ols} - \beta_j) &= O_p(1) \\ \sqrt{n}(\widetilde{\beta}_0^{ols} - \beta_0) &= O_p(1). \end{aligned} \tag{23}$$

In what follows, we also refer to the limits in (20)-(22) as $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{B}$ respectively.

Lemma A1 As $n \to \infty$, under the conditions $\lambda_n/n \to 0$ and $\lambda_n/n^{1-\gamma} \to \infty$, it holds that $n(\hat{\boldsymbol{\beta}}_S^{AL} - \boldsymbol{\beta}_S^0) = O_p(1)$ and $n(\hat{\boldsymbol{\beta}}_{S^c}^{AL} - \boldsymbol{0}) = O_p(1)$.

Proof of Lemma A1. Consider the adaptive lasso objective function under the local parameterization $\beta_j = \beta_j^0 + u_j/n$ and $\boldsymbol{u} = (u_1, \ldots, u_p)'$

$$\begin{aligned}\mathcal{L}_n(u) &= \sum_{t=1}^{n}(y_t - \boldsymbol{x}_t'(\boldsymbol{\beta}^0 + \frac{\boldsymbol{u}}{n}))^2 + \lambda_n \sum_{j=1}^{p} w_j|\beta_j^0 + \frac{u_j}{n}| \\ &= \sum_{t=1}^{n}(z_t - \frac{1}{n}\boldsymbol{x}_t'\boldsymbol{u})^2 + \lambda_n \sum_{j=1}^{p} w_j|\beta_j^0 + \frac{u_j}{n}| \qquad (24)\end{aligned}$$

Our goal is to show that $\hat{\boldsymbol{u}}^{AL} = n(\hat{\boldsymbol{\beta}}^{AL} - \boldsymbol{\beta}^0) = O_p(1)$ which implies the stated consistency results. Expanding the square in the first term, we get:

$$\begin{aligned}\mathcal{L}_n(u) &= \sum_{t=1}^{n}\left(z_t^2 - 2z_t\frac{1}{n}\boldsymbol{x}_t'\boldsymbol{u} + \frac{1}{n^2}(\boldsymbol{x}_t'\boldsymbol{u})^2\right) + \lambda_n \sum_{j=1}^{p} w_j|\beta_j^0 + \frac{u_j}{n}| \\ &= \sum_{t=1}^{n} z_t^2 - \frac{2}{n}\sum_{t=1}^{n} z_t\boldsymbol{x}_t'\boldsymbol{u} + \frac{1}{n^2}\sum_{t=1}^{n}\boldsymbol{u}'\boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{u} + \lambda_n \sum_{j=1}^{p} w_j|\beta_j^0 + \frac{u_j}{n}| \\ &= \sum_{t=1}^{n} z_t^2 - \frac{2}{n}\sum_{t=1}^{n} z_t\boldsymbol{x}_t'\boldsymbol{u} + \frac{1}{n^2}\boldsymbol{u}'\left(\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t'\right)\boldsymbol{u} + \lambda_n \sum_{j=1}^{p} w_j|\beta_j^0 + \frac{u_j}{n}| \qquad (25)\end{aligned}$$

Let $\hat{\boldsymbol{u}}^{AL} = n(\hat{\boldsymbol{\beta}}^{AL} - \boldsymbol{\beta}^0)$ be the minimizer of $\mathcal{L}_n(u)$. We can write:

$$\mathcal{L}_n(u) - \mathcal{L}_n(0) = -\frac{2}{n}\sum_{t=1}^{n} z_t\boldsymbol{x}_t'\boldsymbol{u} + \frac{1}{n^2}\boldsymbol{u}'\left(\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t'\right)u + \lambda_n \sum_{j=1}^{p} w_j\left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|\right) \qquad (26)$$

The first two components in the RHS of (26) are both bounded by invoking (20) and (22). For the third term we proceed by separating active and inactive covariates, splitting it into two parts, one for the active predictors ($j \in S$) and one for the inactive predictors ($j \in S^c$):

$$\begin{aligned}\lambda_n \sum_{j=1}^{p} w_j\left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|\right) &= \lambda_n \sum_{j \in S} w_j\left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|\right) \\ &\quad + \lambda_n \sum_{j \in S^c} w_j\left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|\right) \qquad (27)\end{aligned}$$

Case $j \in S$ ($\beta_j^0 \neq 0$): we have $w_j = 1/|\widetilde{\beta}_j^{ols}|^\gamma$ and since $\widetilde{\beta}_j^{ols} \xrightarrow{p} \beta_j^0$ by the consistency of OLS it follows that under this scenario $w_j \xrightarrow{p} 1/|\beta_j^0|^\gamma$, hence $w_j = O_p(1)$. Next, we invoke the mean value theorem, being cautious with the fact that the absolute value function is non-differentiable at zero. Indeed, if the interval between β_j^0 and $\beta_j^0 + u_j/n$ includes zero, the standard MVT wouldn't apply. We therefore consider two cases:

(a) β_j^0 and $\beta_j^0 + \frac{u_j}{n}$ have the same sign (i.e. $\beta_j^0(\beta_j^0 + u_j/n) > 0$): in this case the absolute value function is differentiable on the open interval between β_j^0 and $\beta_j^0 + u_j/n$. From the MVT we therefore obtain

$$|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0| = \frac{u_j}{n}\text{sign}(\beta_j^0 + \frac{u_j^*}{n}) = \frac{u_j}{n}\text{sign}(\beta_j^0) \tag{28}$$

where u_j^* lies between 0 and u_j, and the last equality follows because β_j^0 and $\beta_j^0 + u_j/n$ have the same sign.

(b) β_j^0 and $\beta_j^0 + u_j/n$ have opposite signs (i.e., $\beta_j^0(\beta_j^0 + u_j/n) < 0$). In this case, zero lies between β_j^0 and $\beta_j^0 + u_j/n$. However, using the triangle inequality we have

$$||\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|| \leq |\frac{u_j}{n}| \tag{29}$$

and this bound is sufficient for our needs since it establishes that the difference in absolute values is bounded by a term of order $1/n$, allowing us to establish the limiting behaviour of (25).

For this $j \in S$ case we therefore have

$$\begin{aligned}\lambda_n \sum_{j \in S} w_j \; n \left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|\right) &= \frac{\lambda_n}{n} \sum_{j \in S} w_j \; n \left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|\right) \\ &= \frac{\lambda_n}{n} \; O_p(1) \; \Big(n \; O_p(1/n)\Big) && (30) \\ &= \frac{\lambda_n}{n} \; O_p(1) = o_p(1) && (31)\end{aligned}$$

since $\lambda_n/n \to 0$.

We write (for any j):

$$\mathcal{L}_n(u) - \mathcal{L}_n(0) = \boldsymbol{u'Au} - 2\boldsymbol{u'B} + \lambda_n \sum_{j \in S^c} w_j \left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0|\right) + o_p(1) \tag{32}$$

Next, we consider the case $j \in S^c$.

<u>Case $j \in S^c$</u> ($\beta_j^0 = 0$): We now have $\beta_j^0 = 0$ so that from $w_j = 1/|\widetilde{\beta}_j^{ols}|^\gamma$ we can write

$$\frac{w_j}{n^\gamma} = \frac{1}{|n\ \widetilde{\beta_j}^{ols}|^\gamma} \tag{33}$$

and from the properties of least squares in cointegrated regressions it immediately follows that $w_j/n^\gamma = O_p(1)$. We can now evaluate the relevant component in the RHS of (25). We have

$$\lambda_n \sum_{j \in S^c} w_j \left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0| \right) = \frac{\lambda_n}{n^{1-\gamma}} \sum_{j \in S^c} \frac{w_j}{n^\gamma}\ |u_j| \tag{34}$$

giving

$$\lambda_n \sum_{j \in S^c} w_j \left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0| \right) \xrightarrow{p} 0 \quad if \quad u_j = 0 \quad \forall j \in S^c \tag{35}$$

and

$$\lambda_n \sum_{j \in S^c} w_j \left(|\beta_j^0 + \frac{u_j}{n}| - |\beta_j^0| \right) \to \infty \quad if \quad u_j \neq 0 \quad for \quad some \quad j \in S^c \tag{36}$$

since $\lambda_n/n^{1-\gamma} \to \infty$ and u_j is fixed. In this latter case, the cost becomes unbounded, forcing $u_j = 0$ for $j \in S^c$. The penalty structure ensures that $u_j = 0$ for all $j \in S^c$ at the minimizer.

We can now conclude up to $o_p(1)$ terms that

$$\mathcal{L}_n(u) - \mathcal{L}_n(0) = \boldsymbol{u'Au} - 2\boldsymbol{u'B} + \begin{cases} o_p(1) \ \text{ if } u_j = 0 \ \ \forall j \in S^c \\ \infty \ \text{ if } u_j \neq 0 \text{ for some } j \in S^c \end{cases} \tag{37}$$

Remark: Since $w_j/n^\gamma = O_p(1)$ and $\lambda_n/n^{1-\gamma} \to \infty$, if any $u_j \neq 0$ for $j \in S^c$, the expression will diverge to infinity. This forces $u_j = 0$ for all $j \in S^c$ at the minimizer.

Here $\boldsymbol{A}$ and $\boldsymbol{B}$ represent the limits of appropriately scaled sums involving $\boldsymbol{x}_t$ and z_t. By the known results on cointegrated regressions (and invoking continuous mapping arguments), these limits exist and are finite.

Next, $(\mathcal{L}_n(u) - \mathcal{L}_n(0))$ is convex in $\boldsymbol{u}$ and the family of convex functions $\{\mathcal{L}_n(u) - \mathcal{L}_n(0)\}$ converges pointwise to $\mathcal{L}_\infty(u)$ which is also convex since it is quadratic with an added infinite penalty outside a certain subspace. This convergence of $(\mathcal{L}_n(u) - \mathcal{L}_n(0))$ to $\mathcal{L}_\infty(u)$ ensures that for large n, the minimizer of $(\mathcal{L}_n(u) - \mathcal{L}_n(0))$ lies near the minimizer of $\mathcal{L}_\infty(u)$. Specifically, focusing on $\mathcal{L}_\infty(u)$ we note that if $u_j \neq 0$ for some inactive covariates $j \in S^c$, the objective is infinite. Thus for minimization we must have $u_j = 0$ for all $j \in S^c$. This effectively constrains the minimization problem to the subspace where $u_j = 0$ for $j \in S^c$. On that subspace $\mathcal{L}_\infty(u) = \boldsymbol{u}'_S \boldsymbol{A} \boldsymbol{u}_S - 2\boldsymbol{u}'_S \boldsymbol{B}$, where $\boldsymbol{u}_S$ is the subvector of $\boldsymbol{u}$ corresponding to the active set S. This is a strictly convex quadratic form, ensuring a unique minimizer. The unique minimizer on that subspace is given by $\boldsymbol{u}_S = \boldsymbol{A}^{-1}\boldsymbol{B}$ and $\boldsymbol{u}_{S^c} = \boldsymbol{0}$. Thus $\mathcal{L}_\infty(u)$ has a unique global minimizer at $(\boldsymbol{A}^{-1}\boldsymbol{B}, \boldsymbol{0})$.

In summary, since $(\mathcal{L}_n(u) - \mathcal{L}_n(0))$ is convex and $\mathcal{L}_\infty(u)$ has a unique minimum of $(\boldsymbol{A}^{-1}\boldsymbol{B}, \boldsymbol{0})'$, following Knight (1999) it holds that $\hat{\boldsymbol{u}}_n^{AL} := \arg\min_u (\mathcal{L}_n(u) - \mathcal{L}_n(0)) \to \arg\min_u \mathcal{L}_\infty(u) = (\boldsymbol{A}^{-1}\boldsymbol{B}, \boldsymbol{0})' = O_p(1)$ as required. □

Proof of Proposition 1. Let $\mathcal{L}_n(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ denote the adaptive lasso objective function given by

$$\mathcal{L}_n(\beta_0, \boldsymbol{\beta}) = \sum_{t=1}^{n} (y_t - \beta_0 - \boldsymbol{x}'_t \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} w_j |\beta_j|. \tag{38}$$

where we have now also included a fitted intercept. The objective function is not differentiable with respect to β_j at $\beta_j = 0$ due to the absolute value function. Therefore we need to consider its subgradient $\partial\boldsymbol{\mathcal{L}}(\beta_0, \boldsymbol{\beta})$ given by

$$\partial\boldsymbol{\mathcal{L}}(\beta_0, \boldsymbol{\beta}) = \begin{pmatrix} -2\sum_{t=1}^{n}(y_t - \beta_0 - \boldsymbol{x}_t'\beta) \\ -2\sum_{t=1}^{n} x_{1t}(y_t - \beta_0 - \boldsymbol{x}_t'\boldsymbol{\beta}) + \lambda_n w_1 \partial|\beta_1| \\ \vdots \\ -2\sum_{t=1}^{n} x_{pt}(y_t - \beta_0 - \boldsymbol{x}_t'\boldsymbol{\beta}) + \lambda_n w_p \partial|\beta_p| \end{pmatrix} \tag{39}$$

where $\partial|\beta_j|$ is the subdifferential of the absolute value function at β_j:

$$\partial|\beta_j| = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0, \\ [-1, 1] & \text{if } \beta_j = 0. \end{cases} \tag{40}$$

At the optimum, the KKT conditions for the adaptive lasso estimator are:

$$0 \in \partial\boldsymbol{\mathcal{L}}(\hat{\beta}_0^{AL}, \hat{\boldsymbol{\beta}}^{AL}) \tag{41}$$

which translate to the following two conditions:

$$\begin{gathered} \sum_{t=1}^{n}(y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}^{AL}) = 0 \\ -2\sum_{t=1}^{n} x_{jt}(y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}^{AL}) + \lambda_n w_j \partial|\hat{\beta}_j^{AL}| = 0 \quad j = 1, \ldots, p \end{gathered} \tag{42}$$

where

$$\partial|\hat{\beta}_j^{AL}| = \begin{cases} \text{sign}(\hat{\beta}_j^{AL}) & \text{if } \hat{\beta}_j^{AL} \neq 0, \\ s_j & \text{if } \hat{\beta}_j^{AL} = 0, \text{ where } s_j \in [-1, 1]. \end{cases} \tag{43}$$

We can now obtain the optimality conditions for the adaptive lasso estimator by analyzing these KKT conditions. For the intercept condition we have

$$\hat{\beta}_0^{AL} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}^{AL}) \tag{44}$$

and for the slopes, we distinguish between the two cases $\hat{\beta}_j^{AL} \neq 0$ and $\hat{\beta}_j^{AL} = 0$. Specifically:

Case $\hat{\beta}_j^{AL} \neq 0$:

If $\hat{\beta}_j^{AL} \neq 0$ then $\partial|\hat{\beta}_j^{AL}|= sign(\hat{\beta}_j^{AL})$ and the earlier KKT condition becomes

$$-2\sum_{t=1}^{n} x_{jt}(y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}^{AL}) + \lambda_n w_j sign(\hat{\beta}_j^{AL}) = 0 \quad j = 1, \ldots, p \tag{45}$$

implying

$$\sum_{t=1}^{n} x_{jt}(y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}^{AL}) = \frac{\lambda_n w_j}{2} sign(\hat{\beta}_j^{AL}) \quad j = 1, \ldots, p \tag{46}$$

Case $\hat{\beta}_j^{AL} = 0$:

If $\hat{\beta}_j^{AL} = 0$ then $\partial|\hat{\beta}_j^{AL}|= s_j \in [-1, 1]$. The KKT conditions for β_j then become

$$-2\sum_{t=1}^{n} x_{jt}(y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}^{AL}) + \lambda_n w_j s_j = 0 \tag{47}$$

implying

$$\left|\sum_{t=1}^{n} x_{jt}(y_t - \hat{\beta}_0^{AL} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}^{AL})\right| \leq \frac{\lambda_n w_j}{2}. \tag{48}$$

Given these optimality conditions, we next aim to establish that under the stated conditions on λ_n, the adaptive lasso is model selection consistent.

We recall that $w_j = 1/|\widetilde{\beta}_j^{ols}|^\gamma$ with $\gamma > 0$. It therefore follows that for $j \in S$, $\widetilde{\beta}_j^{ols}$ converges in probability to a nonzero value and therefore w_j is bounded, with $w_j = 1/|\widetilde{\beta}_j^{ols}|^\gamma= O_p(1)$. For $j \in S^c$, we have $\widetilde{\beta}_j^{ols} = O_p(1/n)$ by the consistency of the OLS estimator. Hence $w_j = 1/|\widetilde{\beta}_j^{ols}|^\gamma= O_p(n^\gamma)$.

Using these properties of the adaptive weights we next concentrate on the conditions on λ_n

that ensure the validity of the KKT conditions. Formally, we need to establish that for $j \in S^c$ and for sufficiently large n, the KKT conditions can only be satisfied if $\hat{\beta}_j^{AL} = 0$ (equivalently that $\hat{\beta}_j^{AL} \neq 0$ cannot occur). Similarly, for $j \in S$, the KKT conditions can only be satisfied for $\hat{\beta}_j^{AL} \neq 0$ (equivalently, $\hat{\beta}_j^{AL} = 0$ cannot occur).

Before proceeding further it is useful to reformulate the KKT conditions by embedding within them the optimality condition associated with the intercept. This can simply be achieved by suitably demeaning the variables appearing in the KKT conditions.

$$\sum_{t=1}^{n} (\widetilde{y}_t - \widetilde{\boldsymbol{x}}_t' \hat{\boldsymbol{\beta}}^{AL}) = \frac{1}{2} \lambda_n \ w_j \ \text{sign}(\hat{\beta}_j^{AL}) \quad \text{if} \quad \hat{\beta}_j^{AL} \neq 0, \quad j = 1, \ldots, p \tag{49}$$

$$\left| \sum_{t=1}^{n} x_{jt} (\widetilde{y}_t - \widetilde{\boldsymbol{x}}_t' \hat{\boldsymbol{\beta}}^{AL}) \right| \leq \frac{1}{2} \lambda_n w_j \quad \text{if} \quad \hat{\beta}_j^{AL} = 0, \quad j = 1, \ldots, p. \tag{50}$$

Using (49)-(50) we next aim to show that $P(\hat{\beta}_j^{AL} = 0 \ \ \forall j \in S^c) \to 1$ and $P(\hat{\beta}_j^{AL} \neq 0 \ \ \forall j \in S) \to 1$ so that model selection consistency of the adaptive lasso is established.

We proceed by contradiction:

Case $j \in S^c$: suppose that $\hat{\beta}_j^{AL} \neq 0$ (i.e., contradiction) so that (49) would be expected to hold. Normalizing both its sides by n gives

$$\frac{1}{n} \sum_{t=1}^{n} (\widetilde{y}_t - \widetilde{\boldsymbol{x}}_t' \hat{\boldsymbol{\beta}}^{AL}) = \frac{1}{2} \frac{\lambda_n}{n} \ w_j \ \text{sign}(\hat{\beta}_j^{AL}). \tag{51}$$

Since $j \in S^c$, $w_j = O_p(n^{\gamma})$, implying

$$\frac{\lambda_n}{n} \ w_j = \frac{\lambda_n}{n^{1-\gamma}} \ O_p(1). \tag{52}$$

Focusing on the LHS we note that for irrelevant variables in a correctly specified model (i.e. after including all $j \in S$), the residual $(\widetilde{y}_t - \widetilde{\boldsymbol{x}}_t' \hat{\beta}^{AL})$ is $(z_t + o_p(1))$, a stationary sequence. It therefore follows that the LHS in (51) is $O_p(1)$ while its RHS diverges to infinity by $\lambda_n / n^{1-\gamma} \to \infty$, a contradiction implying that $\hat{\beta}_j^{AL}$ cannot be nonzero for $j \in S^c$. We conclude that $P(\hat{\beta}_j^{AL} = 0 \ \ \forall j \in S^c) \to 1$ as $n \to \infty$ and as stated.

<u>Case $j \in S$</u>: Supppose that $\hat{\beta}_j^{AL} = 0$ (i.e., a contradiction) so that (50) would be expected to hold. Normalizing both of its sides by n we rewrite it as

$$\left| -\frac{1}{n} \sum_{t=1}^{n} x_{jt}(\widetilde{y}_t - \widetilde{\boldsymbol{x}}_t' \hat{\boldsymbol{\beta}}^{AL,-j}) \right| \leq \frac{\lambda_n}{n} w_j \tag{53}$$

where $\hat{\boldsymbol{\beta}}^{AL,-j}$ refers to $\hat{\boldsymbol{\beta}}^{AL}$ with its j^{th} component replaced by zero when evaluating the residual. Here since $\widetilde{\beta}_j^{ols} \xrightarrow{p} \beta_j$ we have $w_j \xrightarrow{p} |\beta_j|^{-\gamma}$ a bounded quantity. If $\hat{\beta}_j^{AL} = 0$, omitting a true active predictor leads to a residual sequence containing an I(1) component. For large n therefore

$$\frac{1}{n} \sum_{t=1}^{n} x_{jt}(\widetilde{y}_t - \widetilde{\boldsymbol{x}}_t' \hat{\boldsymbol{\beta}}^{AL,-j}) = O_p(n) \tag{54}$$

so that with $\lambda_n/n \to 0$, $\exists N' > 0$ such that $\forall n > N'$

$$P\left(\left| -\frac{1}{n} \sum_{t=1}^{n} x_{jt}(\widetilde{y}_t - \widetilde{\boldsymbol{x}}_t' \hat{\boldsymbol{\beta}}^{AL,-j}) \right| > \frac{\lambda_n}{n} w_j \right) \to 1 \tag{55}$$

hence a contradiction to $\hat{\beta}_j^{AL} = 0$. Thus $P(\hat{\beta}_j^{AL} \neq 0 \ \ \forall j \in S) \to 1$, as stated. □

Proof of Proposition 2. Here we show that if z_t is truly I(1) with no cointegration, and if λ_n is such that $\lambda_n/n \to 0$ then the adaptive LASSO penalty effectively shrinks nothing to zero, i.e., it selects (nearly) all predictors with high probability. Recall that in a spurious regression setting the least squares estimates $\widetilde{\beta}_j^{ols}$ are $O_p(1)$. As a result, the adaptive LASSO weights also satisfy $w_j = O_p(1)$, that is, all adaptive weights remain bounded away from ∞ and from 0. From the KKT condition in (50) we note that the LHS is not shrinking, being an $O_p(1)$ random variable, while the right hand side tends to 0 since $\lambda_n w_j = o(n)$. Therefore with high probability, the inequality fails for each j, so $\hat{\beta}_j^{AL} \neq 0$ and all predictors remain active. □

Proof of Proposition 3. For simplicity we set $\mu = 0$ and $k = 0$ in the auxiliary regressions. Part (i): here the true model is a spurious regression with $\hat{z}_t \sim (1)$. From $n(\overline{\sigma}_0^2 - \overline{\sigma}_1^2) = -\hat{\phi}^2 \sum \hat{z}_{t-1}^2 + 2\hat{\phi} \sum \hat{z}_{t-1} \Delta \hat{z}_t$ and the properties of I(1) processes it immediately follows that $n(\overline{\sigma}_0^2 - \overline{\sigma}_1^2) = O_p(1)$ and $\overline{\sigma}_1^2$ converges in probability to a finite limit. Next, $P(IC_0 < IC_1) = P(n \ln \overline{\sigma}_0^2 / \overline{\sigma}_1^2 < c_n)$. Since $n \ln \overline{\sigma}_0^2 / \overline{\sigma}_1^2 = O_p(1)$ and $c_n \to \infty$ it immediately follows that $P(IC_0 < IC_1 | \mathcal{M}_0) \to 1$ as required. Part (ii) Here the true model is a cointegrated regression with $\hat{z}_t \sim I(0)$. We have $P(IC_1 < IC_0) = P(\ln \overline{\sigma}_0^2 / \overline{\sigma}_1^2 > \frac{c_n}{n})$. Since now $\ln \overline{\sigma}_0^2 / \overline{\sigma}_1^2$ is strictly positive with high probability and $c_n / n \to 1$, it follows that $P(IC_1 < IC_0 | \mathcal{M}_1) \to 1$ as required. □

References

Engle, R. F. and Granger, C. W. J. (1987). 'Co-Integration and Error Correction: Representation, Estimation, and Testing', *Econometrica*, Vol. 55, pp. 251–276.

Engle, R. F. and Yoo, B. S. (1987). 'Forecasting and Testing in Cointegrated Systems', *Journal of Econometrics*, Vol. 35, pp. 143-159.

Gonzalo, J. and Pitarakis, J. (1998). 'Specification via model selection in error correction models', *Economics Letters*, Vol. 60, pp. 321-328.

Gonzalo, J. and Pitarakis, J. (1999). Dimensionality effect in cointegration analysis. In: Engle, R. F., White, H. (Eds.) Cointegration, Causality and Forecasting: Festschrift in Honour of Clive Granger. Oxford University Press, Oxford, pp. 212–229.

Gonzalo, J. and Pitarakis, J. (2021). 'Spurious relationships in high-dimensional systems with strong or mild persistence', *International Journal of Forecasting*, Vol. 37, pp. 1480-1497.

Johansen, S. (1991). 'Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models', *Econometrica*, Vol. 59, pp. 1551-1580.

Knight, K. and Fu, W. (2000). 'Asymptotics for LASSO type Estimators', *Annals of Statistics*, Vol. 28, pp. 1356-1378.

Knight, K. (2008). ‘Shrinkage estimation for nearly singular designs’, *Econometric Theory*, Vol. 24, 323-357.

Kock, A. B. (2016). ‘Consistent and Conservative Model Selection with the Adaptive LASSO in Stationary and Nonstationary Autoregressions’, *Econometric Theory*, Vol. 32, pp. 243-259.

Koo, B., Anderson, H. M., Seo, M. H., and Yao, W. (2020). ‘High Dimensional Predictive Regression in the presence of cointegration’, *Journal of Econometrics*, Vol. 219, pp. 456-477.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992). ‘Testing the Null of Stationarity Against the Alternative of a Unit Root’, *Journal of Econometrics*, Vol. 54, pp. 159–178.

Lee, J. J., Shi, Z., and Gao, Z. (2022). ‘On LASSO for predictive regression’, *Journal of Econometrics*, Vol. 229, pp. 322-349.

MacKinnon, J. G. (1991). ‘Critical Values for Cointegration Tests’, *Oxford Bulletin of Economics and Statistics*, Vol. 53, pp. 25–43.

Onatski, A. and Wang, C. (2018). ‘Alternative asymptotics for cointegration tests in large VARs’, *Econometrica* 86, pp. 1465–1478.

Phillips, P. C. B. (2008). ‘Unit Root Model Selection’, *Journal of the Japan Statistical Society*, Vol. 38, pp. 65-74.

Phillips, P. C. B. and Durlauf, S. N. (1986). ‘Multiple Time Series Regression with Integrated Processes’, *Review of Economic Studies*, Vol. 53, pp. 473-495.

Phillips, P. C. B. and Ouliaris, S. (1990). ‘Asymptotic Properties of Residual-Based Tests for Cointegration’, *Econometrica*, Vol. 58, pp. 165–193.

Phillips, P. C. B. and Ploberger, W. (1996). ‘ An asymptotic theory of Bayesian inference for time series’, *Econometrica*, Vol. 64, pp. 381–413.

Shin, Y. (1994). ‘A Residual-Based Test of the Null of Cointegration against the Alternative of No Cointegration’, *Econometric Theory*, Vol. 10, pp. 91-115.

Smeekes, S. and Wijler, E. (2021). ‘An automated approach towards sparse single-equation cointegration modelling’, *Journal of Econometrics*, Vol. 221, pp. 247-276.

Xiao, Z. and Phillips, P. C. B. (2002). ‘A CUSUM test for cointegration using regression residuals’, *Journal of Econometrics*, Vol. 108, pp. 43-61.

Zou, H. (2016). ‘The Adaptive LASSO and its Oracle Properties’, *Journal of the American Statistical Association*, Vol. 476, pp. 1418-1429.