An automated approach towards sparse single-equation cointegration modelling

Journal of Econometrics 221 (2021) 247–276

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

An automated approach towards sparse single-equation cointegration modelling☆

Stephan Smeekes, Etienne Wijler*

Maastricht University, Department of Quantitative Economics, The Netherlands

ARTICLE INFO

Article history:
Received 25 September 2018
Received in revised form 16 January 2020
Accepted 1 July 2020
Available online 6 August 2020

JEL classification:
C32
C52
C55

Keywords:
SPECS
Penalized regression
Single-equation error-correction model
Cointegration
High-dimensional data

ABSTRACT

In this paper we propose the Single-equation Penalized Error Correction Selector (SPECS) as an automated estimation procedure for dynamic single-equation models with a large number of potentially (co)integrated variables. By extending the classical single-equation error correction model, SPECS enables the researcher to model large cointegrated datasets without necessitating any form of pre-testing for the order of integration or cointegrating rank. Under an asymptotic regime in which both the number of parameters and time series observations jointly diverge to infinity, we show that SPECS is able to consistently estimate an appropriate linear combination of the cointegrating vectors that may occur in the underlying DGP. In addition, SPECS is shown to enable the correct recovery of sparsity patterns in the parameter space and to possess the same limiting distribution as the OLS oracle procedure. A simulation study shows strong selective capabilities, as well as superior predictive performance in the context of nowcasting compared to high-dimensional models that ignore cointegration. An empirical application to nowcasting Dutch unemployment rates using Google Trends confirms the strong practical performance of our procedure.

1. Introduction

In this paper we propose the Single-equation Penalized Error Correction Selector (SPECS) as a tool to perform automated modelling of a potentially large number of time series of unknown order of integration. In many economic applications, datasets will contain possibly (co)integrated time series, which has to be taken into account in the statistical analysis. Traditional approaches include modelling the full system of time series as a vector error correction model (VECM), estimated by methods such as maximum likelihood estimation (Johansen, 1995), or transforming all variables to stationarity before performing further analysis. However, both methods have considerable drawbacks when the dimension of the dataset increases.

While the VECM approach allows for flexible modelling of potentially cointegrated series, these estimators suffer from the curse of dimensionality due to the large number of parameters to estimate. In practice they therefore quickly become difficult to interpret and computationally intractable on even moderately sized datasets. As such, to reliably apply such

☆ The first author was financially supported by the Netherlands Organization for Scientific Research (NWO) under grant number 452-17-010. Previous versions of this paper were presented at CFE-CM Statistics 2017, NESG 2018 and $(EC)^2$ 2018. We gratefully acknowledge the comments by participants at these conferences. In addition, we thank the editor and two anonymous referees as well as Robert Adámek, Alain Hecq, Luca Margaritella, Alexei Onatski, Hanno Reuvers, Sean Telg, Ines Wilms and Qiwei Yao for valuable comments and feedback, and Caterina Schiavoni for help with the data collection. All remaining errors are our own.

* Correspondence to: Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
E-mail address: e.wijler@maastrichtuniversity.nl (E. Wijler).

https://doi.org/10.1016/j.jeconom.2020.07.021
0304-4076/

full-system estimators requires non-trivial a priori choices on the relevance of specific variables to keep the dimension manageable. Moreover, often one only has a single variable of interest, and estimating the parameter-heavy full system is not necessary.

On the other hand, the alternative strategy of prior transformations to stationarity is more easily compatible with single variables of interest and larger dimensions, but requires either a priori knowledge of the order of integration of individual variables, or pre-testing for unit roots, which is prone to errors in particular if the number of variables is large (cf. Smeekes and Wijler, 2020). Additionally, this approach ignores the presence of cointegration among the variables, which may have detrimental effects on the subsequent analysis.

SPECS is a form of penalized regression designed to sparsely estimate a conditional error correction model (CECM). We demonstrate that SPECS possesses the oracle property as defined in Fan and Li (2001); in particular, SPECS simultaneously allows for consistent estimation of the non-zero coefficients and the correct recovery of sparsity patterns in the single-equation model. It therefore provides a fully data-driven way of selecting the relevant variables from a potentially large dataset of (co)integrated time series. Moreover, due to the flexible specification of the single-equation model, SPECS is able to take into account cointegration in the dataset without requiring any form of pre-testing for unit roots or testing for the cointegrating rank, and can thus be applied "as is" to any dataset containing an (unknown) mix of stationary and integrated time series. As a companion to this paper, an *R* package is made available that implements a fast and easy-to-interpret algorithm for SPECS estimation, and provides immediate access to the dataset used in the empirical application.[1]

Single-equation error correction models are frequently employed in tests for cointegration (e.g. Engle and Granger, 1987; Phillips and Ouliaris, 1990; Boswijk, 1994; Banerjee et al., 1998) as well as in forecasting applications (e.g. Engle and Yoo, 1987; Chou et al., 1996), but require a weak exogeneity assumption for asymptotically efficient inference (Johansen, 1992). Weak exogeneity entails the existence of a single cointegrating vector that only appears in the marginal equation for the variable of interest. If this assumption holds, our procedure can be interpreted as an alternative to cointegration testing in the ECM framework (Boswijk, 1994; Palm et al., 2010). However, weak exogeneity may not be realistic in large datasets and we provide detailed illustrations of the implications of failure of this assumption and demonstrate that absent of weak exogeneity our procedure consistently estimates a linear combination of the true cointegrating vectors. While this impedes inference on the cointegrating relations, when the main aim of the model is nowcasting or forecasting, our procedure remains theoretically justifiable and provides empirical researchers with a simple and powerful tool for automated analysis of high-dimensional non-stationary datasets. In addition, for modelling a single variable of interest using a large set of potential regressors, SPECS provides a variable selection mechanism, allowing the researcher to discard variables that are irrelevant for this particular analysis. Our simulation results demonstrate strong selective capabilities in both low and high dimensions. Furthermore, a simulated nowcasting application highlights the importance of incorporating cointegration in the data as our proposed estimators obtain higher nowcast accuracies in comparison to a penalized autoregressive distributed lag (ADL) model. This finding is confirmed in an empirical application, where SPECS is employed to nowcast Dutch unemployment rates with the use of a dataset containing Google Trends series.

The use of penalized regression in time series analysis has gained in popularity, with a wide range of variants showing promising performance in applications (see Smeekes and Wijler, 2018b, for a recent overview). Recent literature has also seen the development of methods for analysing high-dimensional (co)integrated time series.

Kock (2016) proposes the adaptive lasso to estimate an augmented Dickey–Fuller regression. While this univariate model is inherently different from ours, it provides an insightful demonstration of how the lasso may be used as an alternative to testing for non-stationarity, parallelling our suggestion to consider SPECS as an alternative for cointegration testing under the assumption of weak exogeneity.

For VECM systems, Wilms and Croux (2016) propose a penalized maximum likelihood approach, with shrinkage performed on the cointegrating vectors, the coefficients regulating the short-run dynamics and the covariance matrix. While their method is shown to obtain forecast gains relative to the traditional Johansen method, no theoretical results are provided. Liao and Phillips (2015) provide an automated method of joint rank selection and parameter estimation with the use of an adaptive penalty and derive oracle properties in a fixed-dimensional framework. Next to this theoretical limitation on its applicability to large datasets, practical implementation is further complicated due to reliance on the eigenvalue decomposition of an asymmetric matrix, which introduces complex values into the corresponding objective function. As noted by Liang and Schienle (2019, p. 424), this results in a non-standard harmonic function optimization problem. Liang and Schienle (2019) propose joint parameter estimation and rank determination by employing a penalty that makes use of the *QR*-decomposition of the long-run coefficient matrix. This method possesses oracle-like properties under a high-dimensional asymptotic regime, but it requires the availability of an initial OLS estimator, thereby preventing applications on datasets in which the number of variables exceeds, or is close to, the number of available time series observations. Additionally, estimation of the long-run and short-run dynamics is performed sequentially rather than simultaneously, necessitating a two-step procedure.

In a single-equation setting, Lee et al. (2018) derive fixed-dimensional oracle properties for the adaptive lasso applied to predictive regressions where the regressors are allowed to be of mixed orders of integration. However, as a consequence

[1] https://cran.r-project.org/web/packages/specs/index.html.

of their model formulation in which all variables enter in levels, their estimator appears to be susceptible to spurious regression when the regressors are not cointegrated.

Finally, outside the penalized regression framework, Zhang et al. (2019) propose an eigenvalue decomposition to estimate the cointegrating space in the presence of any integer and fractional order of integration of the variables. However, the estimation procedure proposed by Zhang et al. does not perform variable selection, nor does it provide explicit estimates of the transient dynamics in a VECM. Onatski and Wang (2019) develop a novel inference procedure for the cointegrating rank in high dimensions. Similar to the Johansen procedure, their test is based on the squared canonical correlations, for which they derive the limit spectral distribution under joint asymptotics with the use of arguments from random matrix theory.

Our proposed method provides several contributions to this existing literature. First, our theoretical results are derived in a high-dimensional framework where the number of parameters is allowed to grow with the sample size. This requires non-standard theoretical results on bounds of the smallest eigenvalue of a matrix of (co)integrated regressors, similar to those in Zhang et al. (2019), which are further developed in this paper. Second, unlike many of the penalized regression methods surveyed above, the practical implementation of SPECS is straightforward for large datasets, including cases where the number of parameters is larger than the time dimension. Third, our method completely removes the need for pre-testing for the order of integration or cointegrating rank, and is not sensitive to spurious regression. Fourth, to the best of our knowledge, our paper is the first to explicitly allow for the presence of deterministic components in the theory, a crucial feature for many applications.

The paper is structured as follows. In Section 2 we discuss the data generating process. Section 3 describes the SPECS estimator. The main theoretical results of the paper are presented in Section 4. Section 5 contains several simulation studies, followed by an empirical application in Section 6. We conclude in Section 7. The main proofs and preliminary lemmas needed for them are contained in Appendix A, while contains results on minimum eigenvalue bounds. Finally, Appendix C contains supplementary material on proofs of preliminary lemmas and additional theorems, as well as further details on the empirical application.

A word on notation. For any an N-dimensional vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^N x_i^p\right)^{1/p}$ denotes the ℓ_p-norm, while for any matrix $\boldsymbol{D}$ with N columns, $\|\boldsymbol{D}\|_p = \max_{\boldsymbol{x}\in\mathbb{R}^N} \frac{\|\boldsymbol{D}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}$ is the corresponding induced norm and $\|\boldsymbol{D}\|_F$ denotes the Frobenius norm. For an index set $S \subset \{1, \dots, N\}$, let $\boldsymbol{x}_S$ be the vector containing the elements of $\boldsymbol{x}$ corresponding to S. Similarly, for a matrix $\boldsymbol{D}$ with N rows, $\boldsymbol{D}_S$ is the sub-matrix containing the rows of $\boldsymbol{D}$ indexed by S. The orthogonal complement of $\boldsymbol{D}$ is denoted by $\boldsymbol{D}_\perp$, such that $\boldsymbol{D}_\perp'\boldsymbol{D} = \boldsymbol{0}$. When $\boldsymbol{D}$ is a square matrix, we denote its N ordered eigenvalues by $\lambda_1(\boldsymbol{D}) \geq \cdots \geq \lambda_N(\boldsymbol{D})$ and $\lambda_{\min}(\boldsymbol{D})$ and $\lambda_{\max}(\boldsymbol{D})$ denote the minimum and maximum eigenvalue, respectively. Furthermore, we use $\boldsymbol{D} \succ 0$ to denote that the matrix is positive definite. A vector of ones of length N is denoted by $\boldsymbol{\iota}_N$ and the N-dimensional identity matrix by $\boldsymbol{I}_N$. We use $\xrightarrow{p}$ ($\xrightarrow{d}$) to denote convergence in probability (distribution) and $\overset{d}{=}$ denotes equivalence in distribution. Finally, we frequently make use of an arbitrary positive and finite constant K whose value may change throughout the paper, but is always independent of the time and cross-sectional dimensions.

2. The high-dimensional error correction model

In this section we first discuss the data generating process for the vector time series along with the assumptions made. Next we transform the multivariate model to a single equation describing our variable of interest.

2.1. Data generating process

Assume one is interested in modelling a single variable of interest, say y_t, based on an N-dimensional time series $\boldsymbol{z}_t = (y_t, \boldsymbol{x}_t')$ observed at $t = 1, \dots, T$. Let $\boldsymbol{z}_t$ be described by

$$\boldsymbol{z}_t = \boldsymbol{\mu} + \boldsymbol{\tau} t + \boldsymbol{\zeta}_t, \tag{1}$$

with the stochastic component given by

$$\Delta\boldsymbol{\zeta}_t = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{\zeta}_{t-1} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta\boldsymbol{\zeta}_{t-j} + \boldsymbol{\epsilon}_t, \tag{2}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $(N \times r)$-dimensional matrices containing the adjustment rates and cointegrating vectors, respectively. The innovations $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \boldsymbol{\epsilon}_{2,t}')'$ satisfy the following assumptions:

Assumption 1. The sequence of innovations $\{\boldsymbol{\epsilon}_t\}_{t\geq 1}$ is an N-dimensional martingale difference sequence (m.d.s.) with $\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}_\epsilon$. Furthermore, we assume that

(1) There exists an $m > 2$, such that $\max_{1\leq i\leq N, 1\leq t\leq T} \mathbb{E}\left|\epsilon_{i,t}\right|^{2m} \leq K_m$, and
(2) There exist constants $\phi_{\min}, \phi_{\max} > 0$, such that $\phi_{\min} \leq \lambda_{\min}(\boldsymbol{\Sigma}_\epsilon) < \lambda_{\max}(\boldsymbol{\Sigma}_\epsilon) \leq \phi_{\max}$.

This assumption implies that $\boldsymbol{\epsilon}_t$ is a martingale difference sequence with at least (a bit more than) four moments existing. The eigenvalue bounds in the second part place some restrictions on the dependence among the elements of $\boldsymbol{\epsilon}_t$, ruling out for instance a strong common factor affecting all errors. However, a wide range of contemporaneous dependence structures, such as spatial dependence, is still allowed.

The model can be rewritten into a VECM form by substituting (1) into (2) to obtain

$$\Delta \boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{B}' \left(\boldsymbol{z}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\tau}(t-1)\right) + \tau^* + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta \boldsymbol{z}_{t-j} + \boldsymbol{\epsilon}_t, \tag{3}$$

where $\tau^* = (I - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)\boldsymbol{\tau}$. From this representation, it can directly be observed that the presence of a constant in (1) results in a constant within the cointegrating relationship if $\boldsymbol{B}'\boldsymbol{\mu} \neq \boldsymbol{0}$. Furthermore, the linear trend in (1) appears as a constant in the differenced series and may additionally appear as a trend within the cointegrating vector if $\boldsymbol{B}'\boldsymbol{\tau} \neq \boldsymbol{0}$, the latter implying that the equilibrium error $\boldsymbol{B}'\boldsymbol{z}_t$ is a trend stationary process.

The following assumption asserts that the process is (at most) I(1), and the Granger Representation Theorem (e.g. Johansen, 1995, p. 49) can be applied.

Assumption 2. Define $\boldsymbol{A}(z) := (1-z)\boldsymbol{I}_N - \boldsymbol{A}\boldsymbol{B}'z - \sum_{j=1}^{p} \boldsymbol{\Phi}_j (1-z)z^j$.

(1) The determinantal equation $|\boldsymbol{A}(z)|$ has all roots on or outside the unit circle.
(2) $\boldsymbol{A}$ and $\boldsymbol{B}$ are $N \times r$ matrices with $1 \leq r \leq N$ and $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{B}) = r$.
(3) The $((N-r) \times (N-r))$ matrix $\boldsymbol{A}_\perp' \left(I_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j\right) \boldsymbol{B}_\perp$ is invertible.

Assumption 2 enables (3) to be written as a vector moving average (VMA) process

$$\boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{C}(L)\boldsymbol{\epsilon}_t + \boldsymbol{C}\boldsymbol{z}_0, \tag{4}$$

where $\boldsymbol{C} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I}_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j\right) \boldsymbol{B}_\perp\right)^{-1} \boldsymbol{A}_\perp'$, $\boldsymbol{s}_t = \sum_{s=1}^{t} \boldsymbol{\epsilon}_s$, $\boldsymbol{C}(L)\boldsymbol{\epsilon}_t$ is a stationary linear process and $\boldsymbol{z}_0$ are initial values. Without loss of generality, we assume henceforth that $\boldsymbol{z}_0 = \boldsymbol{0}$.

We need a further restriction on the dependence in the VMA representation in the form of the following assumption, which ensures norm-summability of the coefficients in the Beveridge–Nelson decomposition.

Assumption 3. There exists a $K < \infty$ such that $\boldsymbol{C}$ in (4) satisfies $\|\boldsymbol{C}\|_\infty \leq K$. In addition, the matrix lag polynomial $\boldsymbol{C}(L)$ is given by $\boldsymbol{C}(z) = \sum_{l=0}^{\infty} \boldsymbol{C}_l z^l$ and satisfies $\sum_{l=0}^{\infty} l \|\boldsymbol{C}_l\|_\infty \leq K$.

2.2. Single-equation representation

The number of parameters to estimate in (3) is at least $2Nr + N^2p$, such that the system quickly grows too large to accurately estimate based on traditional methods. As we assume a single variable y_t is of interest, we therefore instead consider the lighter parameterized single-equation model for y_t. To ensure that the variables modelling the variation in y_t remain exogenous, we orthogonalize the errors driving the single-equation model, say $\epsilon_{y,t}$, from the errors driving the marginal equations of the endogenous variables $\boldsymbol{x}_t$. This is achieved by decomposing $\epsilon_{1,t}$ into its best linear prediction based on $\boldsymbol{\epsilon}_{2,t}$ and the corresponding orthogonal prediction error. To this end, partition the covariance matrix of $\boldsymbol{\epsilon}_t$ as

$$\boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} \mathbb{E}(\epsilon_{1,t})^2 & \mathbb{E}(\epsilon_{1,t}\boldsymbol{\epsilon}_{2,t}') \\ \mathbb{E}(\epsilon_{1,t}\boldsymbol{\epsilon}_{2,t}) & \mathbb{E}(\boldsymbol{\epsilon}_{2,t}\boldsymbol{\epsilon}_{2,t}') \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{21}' \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \tag{5}$$

such that we obtain

$$\epsilon_{1,t} = (0, \boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1})\boldsymbol{\epsilon}_t + \left(1, -\boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1}\right)\boldsymbol{\epsilon}_t = \hat{\epsilon}_{1,t} + \epsilon_{y,t}. \tag{6}$$

Define $\pi_0 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$. Then, writing out (6) in terms of the observable time series results in the single-equation model

$$\begin{aligned} \Delta \boldsymbol{y}_t &= \left(1, -\boldsymbol{\pi}_0'\right) \left(\boldsymbol{A}\boldsymbol{B}'(\boldsymbol{z}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\tau}(t-1)) + \boldsymbol{\tau}^* + \sum_{j=1}^{p} \boldsymbol{\Phi}_j' \Delta \boldsymbol{z}_{t-j} \right) + \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \epsilon_{y,t} \\ &= \boldsymbol{\delta}'\boldsymbol{z}_{t-1} + \boldsymbol{\pi}'\boldsymbol{w}_t + \mu_0 + \tau_0(t-1) + \epsilon_{y,t}, \end{aligned} \tag{7}$$

where $\boldsymbol{\delta}' = \left(1, -\boldsymbol{\pi}_0'\right)\boldsymbol{A}\boldsymbol{B}'$, $\boldsymbol{\pi} = (\boldsymbol{\pi}_0', \ldots, \boldsymbol{\pi}_p')'$ with $\boldsymbol{\pi}_j' = (1, -\boldsymbol{\pi}_0')\boldsymbol{\Phi}_j$ for $j = 1, \ldots, p$, $\mu_0 = (1, -\boldsymbol{\pi}_0')\left(-\boldsymbol{A}\boldsymbol{B}'\boldsymbol{\mu} + \boldsymbol{\tau}^*\right)$ and $\tau_0 = (1, -\boldsymbol{\pi}_0')\boldsymbol{\tau}^*$. Note that $\boldsymbol{\delta}$ is a vector of length N, whereas $\boldsymbol{\pi}$ is a vector of length $M = N(p+1) - 1$. Additionally, $\boldsymbol{w}_t = (\Delta \boldsymbol{x}_t', \Delta \boldsymbol{z}_{t-1}', \ldots, \Delta \boldsymbol{z}_{t-p}')'$ and $\epsilon_{y,t} = (1 - \boldsymbol{\pi}_0')\boldsymbol{\epsilon}_t$. Finally, we write the single-equation model in matrix notation as

$$\Delta \boldsymbol{y} = \boldsymbol{Z}_{-1}\boldsymbol{\delta} + \boldsymbol{W}\pi + \boldsymbol{\iota}_T\mu_0 + \boldsymbol{t}\tau_0 + \boldsymbol{\epsilon}_y = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\epsilon}_y, \tag{8}$$

where $\boldsymbol{Z}_{-1} = (\boldsymbol{z}_0, \ldots, \boldsymbol{z}_{T-1})'$, $\boldsymbol{W} = (\boldsymbol{w}_t, \ldots, \boldsymbol{w}_T)'$, $\boldsymbol{t} = (0, \ldots, T-1)'$, $\boldsymbol{V} = (\boldsymbol{Z}_{-1}, \boldsymbol{W})$, $\boldsymbol{D} = (\boldsymbol{\iota}_T, \boldsymbol{t})$, $\boldsymbol{\gamma} = (\boldsymbol{\delta}', \boldsymbol{\pi}')'$ and $\boldsymbol{\theta} = (\mu_0, \tau_0)'$.

Remark 1. The single-equation model may similarly be derived under the assumption of normal errors. In this framework, $\epsilon_{y,t}$ has the conditional normal distribution from which (7) can be obtained (cf. Boswijk, 1994). A benefit of assuming normality is that, under the additional assumption of weak exogeneity, the OLS estimates based on (7) are optimal in the mean-squared sense. However, the assumption of normality is unnecessarily restrictive when the, perhaps overly, ambitious goal of complete and correct specification is abandoned.

Remark 2. An additional benefit of the conditional error-correction model, as opposed to the predictive regressions specified in levels considered in Lee et al. (2018), is that the former avoids spurious regression. In the case where all variables in $\boldsymbol{z}_t$ are integrated of order one and independent of one another, the left-hand side of (8) would remain stationary. Intuitively, any "best fitting" linear combination between the stationary component Δy_t and $(\boldsymbol{z}_{t-1}', \boldsymbol{w}_t')'$ would seek to minimize the contribution of the variables in $\boldsymbol{z}_t$, as their stochastically trending nature substantially inflates the fitting error. This behaviour is well-documented for the fixed-dimensional OLS estimator – cf. Boswijk (1994, A.9) in which $\hat{\boldsymbol{\delta}}_{OLS}$ turns out to be superconsistent – and carries over to SPECS in high-dimensions.

In general, the implied cointegrating vector $\boldsymbol{\delta}$ in the single-equation model for y_t contains a linear combination of the cointegrating vectors in $\boldsymbol{B}$ with their weights being given by $\left(1, -\boldsymbol{\pi}_0'\right)\boldsymbol{A}$. Since the marginal equations of $\boldsymbol{x}_t$ contain information about the cointegrating relationship, efficient estimation within the single-equation model is only attained under an assumption of weak exogeneity. Johansen (1992) shows that sufficient conditions for weak exogeneity to hold are (i) normality of $\boldsymbol{\epsilon}_t$, (ii) $\text{rank}(\boldsymbol{A}\boldsymbol{B}') = 1$, i.e. there is a single cointegrating N-dimensional cointegrating vector $\boldsymbol{\beta}$, and (iii) the vector of adjustment rates takes on the form $\boldsymbol{\alpha} = (\alpha_1, \boldsymbol{0}')'$. However, these conditions are rather restrictive when considering high-dimensional economic datasets that are likely to possess multiple cointegrating relationships and complex covariance structures across the errors. Therefore, we opt to derive our results without assuming weak exogeneity, while acknowledging that direct interpretation of the estimated cointegrating vector will only be valid in the presence of weak exogeneity. Furthermore, we believe that whether the potential loss of asymptotic efficiency in our more parsimonious single-equation model translates to inferior performance in finite samples ultimately remains an empirical question.

As we consider sparse estimation of this single-equation model, let us briefly touch upon the required sparsity. For measuring the sparsity, we work directly in the single-equation representation.[2] Let $S_\delta = \{i|\delta_i \neq 0\}$ denote the index set of the non-zero elements in $\boldsymbol{\delta}$, with its cardinality denoted by $|S_\delta|$, and let S_π be defined accordingly for $\boldsymbol{\pi}$. In addition, let r^* denote the dimension of the cointegration space of $\boldsymbol{z}_{S_\delta,t}$, i.e. the number of independent linear stationary combinations of $\boldsymbol{z}_{S_\delta,t}$ (cf. Remark 3), and define $s_\delta = |S_\delta| - r^*$ and $s_\pi = |S_\pi| + r^*$ as the number of "effective" relevant non-stationary and stationary variables, respectively. Our estimation goal will then be to obtain estimates of S_δ and S_π, as well as estimate $\boldsymbol{\delta}_{S_\delta}$ and $\boldsymbol{\pi}_{S_\pi}$. To obtain consistency, we need the following assumptions on the amount of sparsity.

Assumption 4. Assume that (1) $s_\delta = o(T^{1/4})$; (2) $s_\pi = o(\sqrt{T})$ and (3) $\max\{s_\delta, \sqrt{s_\pi}\} = o(\gamma_{\min}\sqrt{T})$, where $\gamma_{\min} = \min\{|\gamma_i| : \gamma_i \neq 0\}$.

Parts (1) and (2) put restrictions on how fast the number of relevant parameters is allowed to grow. The "effective" number of relevant stationary variables (s_π) is allowed to grow faster than the "effective" number of integrated variables (s_δ), as a result of the collinearity induced by the stochastic trends (cf. Remark 4). Part (3) puts an additional restriction on the number of relevant coefficients as a function of the smallest non-zero coefficient. Clearly, if all coefficients are assumed to be fixed, (3) is not binding. In fact, one can allow $\gamma_{\min}$ to shrink at a rate up to $T^{-1/4}$ before it becomes binding. This assumption may therefore be interpreted as determining the fastest rate at which the population coefficients are allowed to decrease, as a function of T, s_δ and s_π, to still ensure it can be consistently picked up by our estimation method.

2.3. Rotations and bounds on eigenvalues

Bounds on eigenvalues play a crucial role in establishing consistency properties of lasso-type penalized regression methods. However, due the mixed integrated nature of our data, where parts of the regressors are stationary, and other parts are only stationary after rotation, the object of our assumptions is not the sample covariance matrix directly, but instead a carefully transformed version. Under Assumptions 1–3, it is then possible to ensure eigenvalue conditions on the sample covariance matrices. Before we can state the assumption, we must therefore establish some further notation and rotations to be used later.

Let $\boldsymbol{\gamma} = (\boldsymbol{\delta}', \boldsymbol{\pi}')'$ and S_γ its active set. Without loss of generality, we partition the data matrix as $\boldsymbol{V} = (\boldsymbol{V}_{S_\gamma}, \boldsymbol{V}_{S_\gamma^c})$, with $\boldsymbol{V}_{S_\gamma} = (\boldsymbol{Z}_{-1,S_\delta}, \boldsymbol{W}_{S_\pi})$ representing the time series carrying non-zero coefficients in the population single-equation model,

[2] In absence of weak exogeneity, it may not be directly obvious how we obtain a sparse single-equation model from the VECM. We therefore provide a more detailed discussion of the interpretation of sparsity absent of weak exogeneity in Section 4.3.1. In this section we just take the single-equation model directly as starting point.

henceforth referred to as the set of relevant variables. In the presence of cointegration, it follows from (4) that the relevant lagged levels can be written as

$$\boldsymbol{z}_{S_\delta,t} = \boldsymbol{C}_{S_\delta}\boldsymbol{s}_t + \boldsymbol{\mu}_{S_\delta} + \boldsymbol{\tau}_{S_\delta} t + \boldsymbol{u}_{S_\delta,t}, \quad \boldsymbol{C}_{S_\delta} = \boldsymbol{B}_{\perp,S_\delta}\left(\boldsymbol{A}_\perp'\left(\boldsymbol{I}_N - \sum_{j=1}^{p}\boldsymbol{\Phi}_j\right)\boldsymbol{B}_\perp\right)^{-1}\boldsymbol{A}_\perp' \tag{9}$$

where $\boldsymbol{B}_{\perp,S_\delta}$ is an $(|S_\delta| \times (N-r))$-dimensional matrix containing the rows of $\boldsymbol{B}_\perp$ indexed by S_δ and $\boldsymbol{u}_{S_\delta,t} = \boldsymbol{C}_{S_\delta}(L)\boldsymbol{\epsilon}_t$. The left null space of $\boldsymbol{B}_{\perp,S_\delta}$, defined as $\boldsymbol{B}^* = \left\{\boldsymbol{x} \in \mathbb{R}^{|S_\delta|} |\ \boldsymbol{B}_{\perp,S_\delta}'\boldsymbol{x} = \boldsymbol{0}\right\}$, contains the linear combinations that convert $\boldsymbol{z}_{S_\delta,t}$ to a stationary process. Accordingly, we also refer to this null space as the cointegrating space of $\boldsymbol{z}_{S_\delta,t}$. By construction, $\boldsymbol{\delta}_{S_\delta} \in \boldsymbol{B}^*$, such that this cointegrating space is non-empty whenever $\boldsymbol{\delta} \neq \boldsymbol{0}$. In this case, we define $\boldsymbol{B}_{S_\delta}$ as a $(|S_\delta| \times r^*)$-dimensional basis matrix of $\boldsymbol{B}^*$, with $r^* \leq |S_\delta|$ representing the dimension of the cointegrating space.[3]

Similarly, we define $\boldsymbol{B}_{S_\delta,\perp}$ as a basis matrix of the left null-space of $\boldsymbol{B}_{S_\delta}$, i.e. a $(|S_\delta| \times (|S_\delta| - r^*))$-dimensional matrix of full column rank with the property that $\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{B}_{S_\delta} = \boldsymbol{0}$. Then, we are able to define a $\boldsymbol{Q}$-transformation that decomposes the reduced system into a stationary and non-stationary contribution as

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{B}_{S_\delta}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{|S_\pi|} \\ \boldsymbol{B}_{S_\delta,\perp}' & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{Q}^{-1} = \begin{bmatrix} \boldsymbol{B}_{S_\delta}\left(\boldsymbol{B}_{S_\delta}'\boldsymbol{B}_{S_\delta}\right)^{-1} & \boldsymbol{0} & \boldsymbol{B}_{S_\delta,\perp}\left(\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{B}_{S_\delta,\perp}\right)^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_{|S_\pi|} & 0 \end{bmatrix}. \tag{10}$$

For the case $\boldsymbol{\delta} = \boldsymbol{0}$, we define $\boldsymbol{Q} = \boldsymbol{I}_{|S_\pi|}$. Post-multiplication of the data matrix by $\boldsymbol{Q}'$ gives

$$\boldsymbol{V}_{S_\gamma}\boldsymbol{Q}' = \begin{bmatrix} \boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta} & \boldsymbol{W}_{S_\pi} & \boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta,\perp} \end{bmatrix} \tag{11}$$

which we refer to as the $\boldsymbol{Q}$-transformed version of $\boldsymbol{V}_{S_\gamma}$. The first $s_\pi = |S_\pi| + r^*$ columns of (11), corresponding to $(\boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta}, \boldsymbol{W}_{S_\pi})$, contain independent stationary linear combinations of the variables that are relevant to Δy_t in the single-equation model. The remaining $s_\delta = |S_\delta| - r^*$ columns, given by $\boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta,\perp}$, contain all linearly independent combinations that are integrated of order one.

Remark 3. We may interpret r^* as the "effective" cointegration rank, where "effective" relates to variable of interest y_t. Essentially, we remove all variables not relevant to y_t in the long-run (those indexed by S_δ^c) and then reconstruct a VECM from the remaining variables, which now has rank r^*.

Finally, we construct a transformed version of the sample covariance matrix based on $\boldsymbol{V}_{S_\gamma}$, which plays a crucial role in the development of our theory. First, to regress out the deterministic components of the observed time series in (3), we define the matrix $\boldsymbol{M} = \boldsymbol{I}_T - \boldsymbol{D}\left(\boldsymbol{D}'\boldsymbol{D}\right)^{-1}\boldsymbol{D}'$.[4] Then, after rotating by $\boldsymbol{Q}$ and regressing out the deterministic components by $\boldsymbol{M}$, the stationary and non-stationary components are scaled via the matrix $\boldsymbol{S}_T = \text{diag}(\sqrt{T}\boldsymbol{I}_{s_\pi}, \frac{T}{\sqrt{s_\delta}}\boldsymbol{I}_{s_\delta})$. Hence, our transformed sample covariance matrix is defined as

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_{S_\gamma}'\boldsymbol{M}\boldsymbol{V}_{S_\gamma}\boldsymbol{Q}'\boldsymbol{S}_T^{-1} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{bmatrix}, \tag{12}$$

$$\text{with} \quad \hat{\boldsymbol{\Sigma}}_{11} = \frac{1}{T}\begin{bmatrix} \boldsymbol{B}_{S_\delta}'\boldsymbol{Z}_{-1,S_\delta}'\boldsymbol{M}\boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta} & \boldsymbol{B}_{S_\delta}'\boldsymbol{Z}_{-1,S_\delta}'\boldsymbol{M}\boldsymbol{W}_{S_\pi} \\ \boldsymbol{W}_{S_\pi}'\boldsymbol{M}\boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta} & \boldsymbol{W}_{S_\pi}'\boldsymbol{M}\boldsymbol{W}_{S_\pi} \end{bmatrix} \tag{13}$$

and $\hat{\boldsymbol{\Sigma}}_{22} = \frac{s_\delta}{T^2}\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{Z}_{-1,S_\delta}'\boldsymbol{M}\boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta,\perp}$. We can now state the eigenvalue assumptions.

Assumption 5. Assume that, on a set with probability converging to 1 as $T, N, p \to \infty$, there exists a constant $\phi > 0$, such that $\inf_{\boldsymbol{x}\in\mathbb{R}^{s_\pi}} \frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{11}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi$ and $\inf_{\boldsymbol{x}\in\mathbb{R}^{s_\delta}} \frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{22}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi$.

The first part of Assumption 5 applies to stationary data and is known to hold when the minimum eigenvalue of the corresponding population covariance matrix is bounded away from zero (e.g. Medeiros and Mendes, 2016, Section B.2). The second part, however, applies to integrated variables and requires arguments that are unique to the non-stationary setting. In particular, we note the necessity of applying a scaling by $\frac{s_\delta}{T^2}$, rather than the usual $\frac{1}{T^2}$ one may expect from the fixed-dimensional literature, cf. Remark 4. In , we show several cases under which Assumption 5 is satisfied.

Remark 4. As an illustration of the problems with adopting the usual scaling by T^{-2}, consider the simple example of an s-dimensional white noise sequence $\boldsymbol{u}_t \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_s)$ and define $\boldsymbol{h}_t = \sum_{j=1}^{t}\boldsymbol{u}_j$. Then, in Lemma B.2 in we show that

[3] The matrix $\boldsymbol{B}_{S_\delta}$ is not uniquely defined. However, in most instances, including those contained in the current work, identification of the span of $\boldsymbol{B}_{S_\delta}$ is sufficient.

[4] Note that $\boldsymbol{D}$ may vary depending on the deterministic specification of the model; setting $\boldsymbol{D} = (\boldsymbol{\iota}_T, \boldsymbol{t})$ allows for both a non-zero constant and linear trend, while simply setting $\boldsymbol{M} = \boldsymbol{I}_T$ may be desired (although not required) when it is believed that $\boldsymbol{\mu} = \boldsymbol{\tau} = \boldsymbol{0}$.

$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T}\boldsymbol{h}_t\boldsymbol{h}_t'\right) > \phi\right) \to 0$, as $s, T \to \infty$, regardless of their relative rates. Hence, even in this simple case we cannot assume that the minimum eigenvalue is bounded away from zero if we stick to the T^{-2} scaling.

Remark 5. There are several noteworthy instances in which $\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right)$ is bounded away from zero with arbitrarily high probability without the need for Assumption 5. Assume that the dimension of the orthogonal complement of the cointegrating space in the subset of relevant non-stationary variables converges to a finite constant, i.e. $s_\delta \to K$. Then, based on a standard functional central limit theorem,

$$\hat{\boldsymbol{\Sigma}}_{22} \xrightarrow{d} K\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\int_0^1 \tilde{\boldsymbol{B}}(r)\tilde{\boldsymbol{B}}'(r)dr\right)\boldsymbol{C}_{S_\delta}'\boldsymbol{B}_{S_\delta,\perp} \overset{d}{=} \int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr,$$

where $\tilde{\boldsymbol{B}}(r)$ is an s_δ-dimensional Gaussian process, described in the proof of Lemma A.2 in Phillips and Hansen (1990), and $\boldsymbol{B}^*(r)$ is simply the linearly transformed version. By the same lemma, it follows that $\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr$ is positive-definite almost surely. Then, for any $\epsilon > 0$, we may choose $\phi(\epsilon) > 0$ such that

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) \le \phi(\epsilon)\right) \to \mathbb{P}\left(\lambda_{\min}\left(\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr\right) \le \phi(\epsilon)\right) \le \epsilon.$$

A straightforward case in which s_δ remains finite is to simply assume that the number of relevant integrated variables stays finite, i.e. $|S_\delta| \le K$. However, a more general example occurs when the dimension of the cointegrating space of $\boldsymbol{z}_{S_\delta,t}$ diverges at the rate $|S_\delta|$. This occurs in the case of a non-stationary factor model with stationary idiosyncratic components, as proposed by Banerjee et al. (2014). Further illustrations are provided in Remark 10.

3. The single-equation penalized error correction selector

Despite the dimension reduction obtained from moving towards a single-equation representation, regularization remains a necessity in high dimensions. The single-equation model (7) contains a total of $N(p+2)+1$ parameters, compared to the $2N(r+1)+N^2p$ parameters in the full-system VECM in (3), resulting in a substantial reduction in dimensionality. However, the dimension may still grow large when either: (1) the number of potentially relevant variables is large or (ii) when the number of lagged differences required to appropriately model the short-run dynamics is large. Therefore, we consider the use of ℓ_1-regularization to enable estimation in high dimensions.

The resulting estimator, henceforth referred to as the Single-equation Penalized Error Correction Selector (SPECS), is defined as the minimizer of the following objective function:

$$G_T(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \|\Delta\boldsymbol{y} - \boldsymbol{V}\boldsymbol{\gamma} - \boldsymbol{D}\boldsymbol{\theta}\|_2^2 + \lambda_I \sum_{i=1}^{N+M} \omega_i |\gamma_i| + \lambda_G \|\boldsymbol{\delta}\|_2, \tag{14}$$

where $M = (N+1)p-1$ refers to the number of transformed variables in $\boldsymbol{w}_t$, i.e. the length of $\boldsymbol{\pi}$. We denote the minimizers of (14) by $\left(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}\right)$. The group penalty, regulated by λ_G, serves to promote exclusion of the lagged levels as a group when there is no cointegration present in the data. In this case, the model is effectively estimated in differences and corresponds to a conditional model derived from a vector autoregressive model specified in differences. The individual ℓ_1-penalties, regulated by λ_I, serve to enforce sparsity in the coefficient vectors $\boldsymbol{\delta}$ and $\boldsymbol{\pi}$ respectively.

The penalty of each coefficient γ_i is weighted by ω_i to enable simultaneous estimation and selection consistency of the coefficients. Therefore, SPECS resembles a sparse group lasso (e.g. Simon et al., 2013) with adaptive weighting, applied to the conditional error correction model. The weights ω_i in (14) are typically derived from an initial estimation procedure such as OLS (if the number of variables is small enough), ridge, or lasso. In particular, let $\hat{\boldsymbol{\gamma}}_I$ denote initial estimates obtained for $\boldsymbol{\gamma}$ using one of the aforementioned methods. The weights can then be constructed as $\omega_i = \left|\hat{\gamma}_{I,i}\right|^{-k}$ for some $k > 0$. As the coefficients of the irrelevant variables tend to zero, this will "blow up" the weights for these coefficients, making them unlikely to be selected in the final estimation. On the other hand, the weights of the relevant coefficients converge to a positive constant leaving them unaffected. This wedge between the weights of relevant and irrelevant coefficients is exactly needed to achieve selection consistency. As demonstrated by Zou (2006), under such assumptions on the weights, the adaptive lasso attains simultaneous selection and estimation consistency, without the necessity for the rather stringent irrepresentable condition in Zhao and Yu (2006).[5] To maintain generality we work with general weights without specifying how they are obtained, and therefore define appropriate assumptions directly on these weights. In Section 4.2 we then return to weight construction and propose a feasible way to construct weights that are theoretically shown to satisfy our assumptions.

[5] In fact, as the adaptive lasso can be written as a regular lasso on a transformed design matrix, the irrepresentable condition, while still needed, operates on this transformed design matrix and becomes a weighted irrepresentable condition. This condition is then in turn implied by appropriate assumptions on the weights. In this paper we directly take this route rather than going via an irrepresentable condition. Section 7.5 of Bühlmann and Van De Geer (2011) provides details on the links between these assumptions.

Assumption 6. Assume that the weights and regularization penalties satisfy:

1. $\omega_{S_\gamma,\max} = o_p(T^\xi)$ for some $\xi > 0$, where $\omega_{S,\max} = \max\{\omega_i : i \in S\}$.
2. $\lambda_I = o\left(\frac{(s_\delta+\sqrt{s_\pi})T^{1/2-\xi}}{\sqrt{s_\delta+s_\pi}}\right)$ and $\lambda_G = o(\sqrt{T})$.
3. Let $\omega_{S,\min} = \min\{\omega_i : i \in S\}$. Then
$$\omega^{-1}_{S^c_\delta,\min} = o_p\left(\min\left\{(s_\delta+s_\pi)^{-1/2}T^{-1/2-\xi}N^{-1/2}, \lambda_I(s_\delta+\sqrt{s_\pi})^{-1}T^{-1}N^{-1/2}\right\}\right),$$
$$\omega^{-1}_{S^c_\pi,\min} = o_p\left(\min\left\{(s_\delta+s_\pi)^{-1/2}T^{-\xi}(Np)^{-1/2}, \lambda_I(s_\delta+\sqrt{s_\pi})^{-1}(TNp)^{-1/2}\right\}\right).$$

Part (1) puts an upper bound on the rate at which the weights corresponding to the relevant variables diverge. Part (2) restricts the maximum admissible growth rate of the penalty. Exceeding this rate would in an excess of shrinkage bias that impedes estimation consistency. Finally, part (3) states that the weights of the irrelevant variables – interacting with the penalty parameter λ_I – grow sufficiently fast in order to guarantee that irrelevant variables are removed from the model with probability converging to one. The required minimum growth rate of the penalty parameter is inversely related to the growth rate of the weights of the irrelevant variables; faster diverging weights require less penalization to identify irrelevant variables.

Remark 6. The only restriction that Assumption 6 imposes on the growth rate of the group penalty is that $\frac{\lambda_G}{\sqrt{T}} \to 0$, which is necessary for preventing shrinkage bias induced by the group penalty from impeding estimation consistency. Since $\lambda_G = 0$ is an admissible value, it follows that the theoretical results presented in the following section apply to the minimizer of $G^*_T(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \|\Delta\boldsymbol{y} - \boldsymbol{V}\boldsymbol{\gamma} - \boldsymbol{D}\boldsymbol{\theta}\|_2^2 + \lambda_I \sum_{i=1}^{N+M} \omega_i |\gamma_i|$ as well, as long as the remaining conditions are satisfied.

Remark 7. Note that the deterministic components $\boldsymbol{\theta}$ are left unpenalized in (14), as their inclusion in the model is desirable to enable identification of the limiting distribution of the estimators. Similar to the classical Frisch–Wraugh–Lovell Theorem, Yamada (2017) show that the inclusion of unpenalized components is equivalent to performing the estimation after regressing out those components. In other words, we may define $\boldsymbol{M} = \boldsymbol{I}_T - \boldsymbol{D}\left(\boldsymbol{D}'\boldsymbol{D}\right)^{-1}\boldsymbol{D}'$ such that $\hat{\boldsymbol{\gamma}}$ may equivalently be defined as

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \|\boldsymbol{M}(\Delta\boldsymbol{y} - \boldsymbol{V}\boldsymbol{\gamma})\|_2^2 + \lambda_I \sum_{i=1}^{N+M} \omega_i |\gamma_i| + \lambda_G \|\boldsymbol{\delta}\|_2 .$$

If one believes that the trend or constant is zero, one may reflect this knowledge in the construction of $\boldsymbol{M}$, with the convention that $\boldsymbol{M} = \boldsymbol{I}_T$ when $\boldsymbol{\mu} = \boldsymbol{\tau} = \boldsymbol{0}$.

Two common data-driven ways to select the tuning parameters λ_I and λ_G are using cross-validation and information criteria. As standard K-fold cross-validation does not respect the time order of the data, we instead consider a time series cross-validation (TSCV) scheme as proposed by e.g. Hyndman and Athanasopoulos (2018) and Wilms et al. (2017), where for different values of $\boldsymbol{\lambda} = (\lambda_I, \lambda_G)'$ the model is estimated on the first part of the sample, and its prediction for the next observation is recorded. The sample is then recursively moved forward towards the end, and the $\boldsymbol{\lambda}$ with the lowest mean squared prediction error is selected. We refer to Smeekes and Wijler (2018b) for details on the implementation and a comparison with traditional K-fold cross-validation.

While cross-validation works well for prediction (Chetverikov et al., 2016), it tends to generally select fairly low penalty levels and therefore includes many variables. An alternative way to select $\boldsymbol{\lambda}$ is using information criteria, where we find the value of $\boldsymbol{\lambda}$ as

$$\hat{\boldsymbol{\lambda}}_{IC} = \arg\min_{\boldsymbol{\lambda}} \ln\left(\frac{1}{T}\left\|\Delta\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) - \boldsymbol{D}\hat{\theta}\right\|_2^2\right) + \frac{C_T \widehat{df}(\boldsymbol{\lambda})}{T},$$

where $\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})$ and $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ denote the minimizers of $G_T(\boldsymbol{\gamma}, \boldsymbol{\theta})$ in (14) for a particular value of $\boldsymbol{\lambda}$. In addition, $\widehat{df}(\boldsymbol{\lambda})$ is an estimate of the degrees of freedom and C_T is the criterion-specific penalty; for the latter we use the Bayesian Information Criterion (Schwarz, 1978, BIC) with $C_T = \ln(T)$.

Zou et al. (2007) show that for the (adaptive) lasso the number of non-zero coefficients is an appropriate estimate for the degrees of freedom for model selection using information criteria. For group lasso penalties, estimating the degrees of freedom is more complicated. Yuan and Lin (2006) propose a heuristic rule, but this requires the least squares estimator which is not available for large N. Alternative rules are provided by Breheny and Huang (2009) and Vaiter et al. (2012) among others, but none are theoretically valid in our setting. For this reason we propose a simple, heuristic rule where we set $\widehat{df}(\boldsymbol{\lambda})$ equal to the number of non-zero coefficients. Essentially this means we ignore the strength of the group penalty on the complexity of the model as long as the group is selected, thereby overestimating $df(\boldsymbol{\lambda})$. As a consequence, we will only choose non-zero values of λ_G if they either improve the fit directly or result in setting the whole group to zero without affecting the fit too much. This is an intentional choice, consistent with our theoretical treatment of the group penalty. As discussed in Remark 6, the group penalty is not necessary and consistency can be achieved even with $\lambda_G = 0$, and can therefore be seen as an optional add-on penalty.

Finally, we note that in practice both methods require the respective objective function to be minimized for a two-dimensional grid of values for $\boldsymbol{\lambda}$. By choosing the lower and upper bounds of the grid carefully, one can ensure that the selected tuning parameters satisfy the assumptions listed in the next subsection. Of course, even though this ensures the theoretical validity of the selection method, its practical performance can still vary considerably. Therefore we investigate the practical performance of BIC and TSCV in the simulations and empirical application respectively.

4. Theoretical results

In this section we derive the asymptotic properties of SPECS, describe the construction of the weights and discuss implications for particular model specifications.

4.1. Asymptotic properties

The first result that we pursue is that of selection consistency, i.e. the ability of an estimation procedure to select the correct set of relevant variables with probability converging to one. In fact, Zhao and Yu (2006) define a stronger property referred to as sign consistency, which additionally requires the procedure to identify the correct signs of the non-zero coefficients with probability converging to one. In the following theorem, we derive sign consistency of SPECS.

Theorem 1. *Under Assumptions 1–6, as* $T, N, p, \to \infty$ *it holds that* $\mathbb{P}\left(sign\left(\hat{\boldsymbol{\gamma}}\right) = sign\left(\boldsymbol{\gamma}\right)\right) \to 1$.

Theorem 1 provides an asymptotic justification for implementing SPECS as a high-dimensional variable selection device. Furthermore, selection consistency is a crucial property when one aims to obtain interpretable solutions or even utilize the estimator as an alternative to classical tests for cointegration. An example of a traditional test for cointegration is the ECM-test by Banerjee et al. (1998) which looks at the t-ratio of the ordinary least squares coefficient of the lagged dependent variable. Alternatively, Boswijk (1994) proposes to test for the joint significance of the least squares coefficients of all lagged variables with a Wald-type test. In our case, one could interpret exclusion of the lagged levels of the dependent variable, or the lagged levels of all variables, as evidence against the presence of cointegration. However, as discussed, an assumption of weak exogeneity is necessary when the aim is a direct interpretation of the estimated cointegration vector. Notwithstanding this caveat, selection consistency offers valuable insights when viewed as a screening mechanism that excludes irrelevant variables even in the absence of weak exogeneity. Moreover, since the set of variables included is strictly smaller than the time series dimension, it is possible to apply a traditional consistent estimator to the selected set of variables (e.g. Belloni and Chernozhukov, 2013). However, ideally SPECS would contain desirable properties that omit the need of a second estimation procedure. For this reason, we establish the simultaneous consistency of the estimated coefficients in the following theorem.

Theorem 2. *Let* $\boldsymbol{S}_T = diag\left(\sqrt{T}\boldsymbol{I}_{s_\pi}, \frac{T}{\sqrt{s_\delta}}\boldsymbol{I}_{s_\delta}\right)$ *and* $\boldsymbol{Q}$ *as defined in (10). Under the same assumptions as in Theorem 1, it holds that* $\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 = O_p\left(s_\delta + \sqrt{s_\pi}\right)$.

The estimation consistency derived in Theorem 2 does not place any restrictions on the relative growth rates of T, N, p, because it relies solely on high-level assumptions stated in the preceding section. However, when we derive sufficient conditions for the eigenvalue assumptions in Assumption 5 in and provide a feasible method to construct weights that satisfy Assumption 6 in Section 4.2, these restrictions do appear. We refer to Section 4.3.3 for an explicit discussion.

Remark 8. As an immediate consequence of Theorem 2, we have $\left\|\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right\|_2 = O_p\left(\frac{s_\delta + \sqrt{s_\pi}}{\sqrt{T}}\right)$, such that SPECS attains $\sqrt{T}$-consistency when s_δ and s_π remain finite. To see this, note that by the assumption on s_δ, it holds that $\frac{T}{\sqrt{s_\delta}} \geq \sqrt{T}$ for sufficiently large T. Then,

$$\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 \geq \sqrt{T}\left\|\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2.$$

Moreover, since the basis matrices $\boldsymbol{B}_{S_\delta}$ and $\boldsymbol{B}_{S_\delta,\perp}$ are not uniquely defined, we may impose a normalization such that $\|\boldsymbol{Q}\|_2 \leq 1$. Then,

$$\left\|\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right\|_2 = \left\|\boldsymbol{Q}'\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 \leq \|\boldsymbol{Q}\|_2\left\|\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 \leq \left\|\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2,$$

such that $\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 \geq \sqrt{T}\left\|\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right\|_2$.

As a corollary to Theorem 2, it is possible to establish a relationship between the limit distribution of SPECS and the OLS estimator based on the subset of relevant variables.

Corollary 1. *Define the OLS oracle estimator as* $\hat{\boldsymbol{\gamma}}_{OLS,S_\gamma} = \arg\min_{\boldsymbol{\gamma}}\left\|\boldsymbol{M}(\Delta\boldsymbol{y} - \boldsymbol{V}_{S_\gamma}\boldsymbol{\gamma})\right\|_2^2$. *Then, with* $\xi > 0$ *as in Assumption 6, under the same assumptions as Theorem 1 it holds that*

$$\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \hat{\boldsymbol{\gamma}}_{OLS,S_\gamma}\right)\right\|_2 = o_p\left(\frac{\lambda_I(\sqrt{s_\delta} + \sqrt{s_\pi})}{T^{1/2-\xi}}\right). \tag{15}$$

The oracle results in Corollary 1, combined with the sign consistency from Theorem 1, are suggestive of a post-selection inferential procedure. In particular, one may implement a two-step estimation procedure in which SPECS is used to perform variable selection in the first step and a regular OLS regression is performed on the selected variables in the second step. Then, after strengthening part 2 of Assumption 6 to $\lambda_I = o\left(\frac{T^{1/2-\xi}}{\sqrt{s_\delta}+\sqrt{s_\pi}}\right)$, Corollary 1 seems to validate the use of the regular OLS distribution for this two-step estimator, essentially ignoring the variable selection from the first stage. For example, in the case where $\left|S_\gamma\right|$ remains finite, one could use the standard fixed-dimensional results (e.g. Boswijk, 1994) to perform inference. However, such a post-selection inferential procedure should be treated with caution, as it is well known that the selection step impacts the sampling properties of the estimator (see Leeb and Pötscher, 2005). The convergence results of many selection procedures, SPECS included, hold pointwise only, i.e. the finite-sample distributions do not converge uniformly over the parameter space to their asymptotic distribution. The practical implication is that for certain values in the parameter space, relying on the oracle properties for post-selection test statistics may provide strongly misleading results. While developing a valid post-selection inference procedure to, for example, test for cointegration is certainly of interest, the field of valid post-selection inference is, despite its rapid development, still in its infancy. None of the currently existing methods, such as those considered in Berk et al. (2013), Van de Geer et al. (2014), Lee et al. (2016) or Chernozhukov et al. (2018), can easily be adapted to – let alone validated in – our setting. Developing such a method therefore requires a full new theory which is outside the scope of the current paper.

4.2. Initial estimates

In this section, we provide the reader with a directly implementable method to construct weights that satisfy Assumption 6. As discussed in Section 3, we construct the weights as $\omega_i = \left|\hat{\gamma}_{I,i}\right|^{-k}$. For our initial estimator we focus here on the ridge estimator, from which we can derive results for OLS as a special case, and comment on the lasso later on in the section.

Note that the power k gives one the flexibility to adjust how big the wedge between relevant and irrelevant variables is. To illustrate, assume that $\hat{\gamma}_{I,i} = \gamma_i + O_p\left(T^{-a}\right)$ for all i. Then, it is clear that $\omega_i = O_p(1)$ when $\gamma_i \neq 0$ and $\omega_i = O_p\left(T^{ka}\right)$ when $\gamma_i = 0$. Therefore, larger values of k will increase the rate at which the weights corresponding to the irrelevant variables diverge. Based on this principle, the availability of a consistent initial estimator allows us to construct weights that satisfy the conditions in Assumption 6. However, while the idea of adjusting divergence rates through imposing varying values of k seems theoretically attractive, large values of k result in substantial amplification of finite-sample estimation error. As a result, the finite-sample performance of the lasso becomes unstable for large k, such that in practice one may want to set the value for k as low as theoretically admissible.

Regardless of the choice of k, the basic ingredient for good adaptive weights is a consistent initial estimator. Therefore, we derive the consistency of the ridge estimator. Recall that the ridge estimator is defined as the minimizer of the following objective function:

$$G_R(\boldsymbol{\gamma}, \boldsymbol{\theta}) := \|\Delta \boldsymbol{y} - \boldsymbol{V}\boldsymbol{\gamma} - \boldsymbol{D}\boldsymbol{\theta}\|_2^2 + \lambda_R \|\boldsymbol{\gamma}\|_2^2 . \tag{16}$$

The properties of the ridge estimator are well-studied in the stationary setting (e.g Hastie et al., 2008, Section 3.4.1). However, to the best of our knowledge, no explicit results are available in the high-dimensional non-stationary case considered here.

In order to derive consistency of the ridge estimator, we redefine the transformed sample covariance matrix from Section 2.3 and the corresponding bound on its minimum eigenvalue. Let $N_\delta = N - r$, $M_\pi = M + r$ and define the new scaling and rotation matrices as $\boldsymbol{S}_R = \operatorname{diag}\left(\sqrt{T}\boldsymbol{I}_{M_\pi}, \frac{T}{\sqrt{N_\delta}}\boldsymbol{I}_{N_\delta}\right)$ and

$$\boldsymbol{Q}_R = \begin{bmatrix} \left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1/2}\boldsymbol{B}' & 0 \\ 0 & \boldsymbol{I}_M \\ \left(\boldsymbol{B}_\perp'\boldsymbol{B}_\perp\right)^{-1/2}\boldsymbol{B}_\perp' & 0 \end{bmatrix},$$

respectively. The new transformed sample covariance matrix, based on the full dataset, is given by

$$\hat{\boldsymbol{\Sigma}}_R = \boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{M}\boldsymbol{V}\boldsymbol{Q}_R\boldsymbol{S}_R^{-1} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{R,11} & \hat{\boldsymbol{\Sigma}}_{R,12} \\ \hat{\boldsymbol{\Sigma}}_{R,21} & \hat{\boldsymbol{\Sigma}}_{R,22} \end{bmatrix}, \tag{17}$$

with $\hat{\boldsymbol{\Sigma}}_{R,11} = \frac{1}{T}\begin{bmatrix} \boldsymbol{B}'\boldsymbol{Z}_{-1}'\boldsymbol{M}\boldsymbol{Z}_{-1}\boldsymbol{B} & \boldsymbol{B}'\boldsymbol{Z}_{-1}'\boldsymbol{M}\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{M}\boldsymbol{Z}_{-1}\boldsymbol{B} & \boldsymbol{W}'\boldsymbol{M}\boldsymbol{W} \end{bmatrix}$, and $\hat{\boldsymbol{\Sigma}}_{R,22} = \frac{N_\delta}{T^2}\boldsymbol{B}_\perp'\boldsymbol{Z}_{-1}'\boldsymbol{M}\boldsymbol{Z}_{-1}\boldsymbol{B}_\perp$. Then, we extend the minimum eigenvalue bound in Assumption 5 to (17) as follows.

Assumption 7. Assume that, on a set with probability converging to 1 as $T, N, p \to \infty$, there exists a constant $\phi_R > 0$, such that $\inf_{\boldsymbol{x}\in\mathbb{R}^{M_\pi}} \frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{R,11}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi_R$ and $\inf_{\boldsymbol{x}\in\mathbb{R}^{N_\delta}} \frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{R,22}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi_R$.

We now derive the convergence rate of the ridge estimator under a further restriction on the growth rates of N, M. The consistency of the ridge estimator is given in the following theorem.

Theorem 3. *Assume that* $\frac{N_\delta}{T^{1/4}} \to 0$, $\frac{M_\pi}{\sqrt{T}} \to 0$, *and* $\lambda_R = O\left(\frac{(N_\delta+\sqrt{M_\pi})\sqrt{T}}{\sqrt{|S_\delta|+|S_\pi|}}\right)$. *Then, under Assumptions 1–3 and 7, it holds that* $\left\|\boldsymbol{S}_R\boldsymbol{Q}_R'^{-1}\left(\hat{\boldsymbol{\gamma}}_R-\boldsymbol{\gamma}\right)\right\|_2 = O_p\left(N_\delta+\sqrt{M_\pi}\right)$.

Similar to Remark 8, it follows from Theorem 3 that $\left\|\hat{\boldsymbol{\gamma}}_R-\boldsymbol{\gamma}\right\|_2 = O_p\left(\frac{N_\delta+\sqrt{M_\pi}}{\sqrt{T}}\right)$. Based on the assumption that $\frac{N_\delta}{T^{1/4}} \to 0$ and $\frac{M_\pi}{\sqrt{T}} \to 0$ in Theorem 3, it follows directly that $\left\|\hat{\boldsymbol{\gamma}}_R-\boldsymbol{\gamma}\right\|_2 = o_p(1)$, and therefore ridge can be used to construct weights that satisfy our Assumption 6. The exact values of k that are needed theoretically vary depending on the number of (total and relevant) variables in the dataset; we return to this issue in Section 4.3.3.

The attentive reader may note that the admissible growth rates of N_δ, M_π in Theorem 3 are the same as those initially assumed on the subsets of relevant variables, i.e. s_δ, s_π, in Theorem 1. The restriction imposed on the number of stochastic trends, $\frac{N_\delta}{T^{1/4}} \to 0$, corresponds closely to that of Corollary 2.1 in Liang and Schienle (2019), who consider (co)integrated processes as well and roughly require that $\frac{N}{T^{1/4-\nu}} \to 0$ for some $\nu > 0$. The growth rate of the total number of (implied) stationary variables is restricted to $\frac{M_\pi}{\sqrt{T}} \to 0$. While this may seem limited in comparison to the admissible (near) exponential growth in the stationary setting with i.i.d. Gaussian errors (e.g. Kock and Callot, 2015, Thm 3), we stress that our time series framework is more general, allowing not only for integrated processes, but also substantial dependence in the stationary component. Regarding the latter, our assumptions closely match those in the second row of Table 6 of Medeiros and Mendes (2016) with $\zeta = 1$, where our allowed growth rates are only slightly slower.

Ideally, we would like to allow for faster rates of divergence for the set of the irrelevant variables. A prospective strategy to attain this, would be to implement the lasso as an initial estimator, the consistency of which may be derived with the use of a compatibility condition (see for example Bühlmann and Van De Geer, 2011, Ch. 6). While desirable, deriving the validity of an appropriate compatibility condition is a considerable task. In addition to the difficulty of showing the theoretical validity of a compatibility condition in the non-stationary setting considered here, the use of a compatibility condition is further complicated by the fact that the stochastic trends asymptotically dominate the variation. More specifically, in order to attain a non-singular limit matrix, a rotation similar to $\boldsymbol{Q}$ is required that separates the stationary and non-stationary components in the full dataset. The standard compatibility condition would have to be adjusted in a non-trivial manner to account for such a rotation. Consequently, we leave the development of a suitable compatibility condition to future research, and instead focus on the ridge estimator under the more stringent growth rates on the number of variables. In the simulations we explore settings beyond these restrictive assumptions, and our adaptive weights continue to function in this case as well. We therefore conjecture that the suitability of the ridge estimator can be extended to a more general setting.

Remark 9. Theorem 3 imposes no minimum growth rate of the penalty term λ_R in (16). Therefore, in the case where $M + N < T$, the choice $\lambda_R = 0$ is both theoretically admissible and computationally feasible, such that consistency of the OLS estimator follows as a by-product of our result. Similarly, under the conditions imposed in Theorem 3, the lasso can also be shown to be a consistent initial estimator. In particular, Assumption 7 allows for the derivation of a minimum eigenvalue bound for the sample covariance matrix of the full data set, which enables application of standard proofs of consistency that are familiar from the fixed-dimensional setting. Due to space consideration, we refrain from providing a full proof on this conjecture, but refer the interested reader to Theorem 3.1 in Liao and Phillips (2015), the proof of which may be adjusted to fit the current setting.

4.3. Implications for particular model specifications

To fully appreciate the theoretical results in the preceding section, a detailed understanding of the generality provided by the set of imposed assumptions is helpful. For example, as the results are derived without requiring weak exogeneity, our set of assumptions allows for the presence of stationary variables in the data. However, in the absence of weak exogeneity, model interpretation becomes non-standard and the notion of sparsity carries non-trivial annotations. Therefore, in this section we elaborate on several relevant model specifications to demonstrate the flexibility of the single-equation model and highlight the practical implications of variable selection in such a general framework.

4.3.1. Sparsity and weak exogeneity

The benefit of ℓ_1-regularized estimation stems from its ability to identify sparse parameter structures. However, the concept of sparsity in the conditional models here considered merits additional clarification, as the potential absence of weak exogeneity obscures standard interpretability. Accordingly, in this section we comment on the interplay between

weak exogeneity and sparsity and provide several illustrative examples of sparse DGPs. For simplicity of illustration, we assume in this and the following section that $\boldsymbol{\mu} = \boldsymbol{\tau} = \mathbf{0}$.

In Section 2 we argue that the coefficients regulating the long-run dynamics in the conditional model are generally derived from linear combinations of the cointegrating vectors in the VECM representation (3). By decomposing the matrix with adjustment rates as $\boldsymbol{A} = (\boldsymbol{\alpha}_1, \boldsymbol{A}_2')'$, we obtain the explicit construction $\boldsymbol{\delta} = \boldsymbol{B}(\boldsymbol{\alpha}_1 - \boldsymbol{A}_2' \boldsymbol{\Sigma}_{\epsilon,22}^{-1} \boldsymbol{\sigma}_{\epsilon,21})$. Hence, it follows that $\delta_i = 0$ if the sparsity condition $\boldsymbol{\beta}_i' \left(\boldsymbol{\alpha}_1 - \boldsymbol{A}_2' \boldsymbol{\Sigma}_{\epsilon,22}^{-1} \boldsymbol{\sigma}_{\epsilon,21}\right) = 0$ is satisfied, where $\boldsymbol{\beta}_i$ is the ith row of $\boldsymbol{B}$. While this condition may hold in a variety of non-trivial ways, specific cases of interest that lead to sparsity in $\boldsymbol{\delta}$ can be derived. For example, an integrated variable $x_{i,t}$ that does not cointegrate with any of the variables in the system ($\boldsymbol{\beta}_i = \mathbf{0}$), will carry a zero coefficient in the derived single-equation long-run equilibrium.

As a more general example, assume that the researcher observes the N-dimensional time series $\boldsymbol{z}_t = (\boldsymbol{z}_{1,t}', \boldsymbol{z}_{2,t}')' = (y_t, \boldsymbol{x}_t')'$, from time $t = 1, \ldots, T$, where $\boldsymbol{z}_{1,t} = (y_t, \boldsymbol{x}_{1,t}')'$ is an N_1-dimensional time series and $\boldsymbol{z}_{2,t}$ is an N_2-dimensional time series. Moreover,

$$\begin{aligned} \begin{bmatrix} \Delta \boldsymbol{z}_{1,t} \\ \Delta \boldsymbol{z}_{2,t} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ \boldsymbol{\Pi}_{21} & \boldsymbol{\Pi}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_{1,t-1} \\ \boldsymbol{z}_{2,t-1} \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} \boldsymbol{\Phi}_{j,11} & \boldsymbol{\Phi}_{j,12} \\ \boldsymbol{\Phi}_{j,21} & \boldsymbol{\Phi}_{j,22} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{z}_{1,t-j} \\ \Delta \boldsymbol{z}_{2,t-j} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{1,t} \\ \boldsymbol{\epsilon}_{2,t} \end{bmatrix} \\ &= \boldsymbol{\Pi} \boldsymbol{z}_{t-1} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta \boldsymbol{z}_{t-j} + \boldsymbol{\epsilon}_t. \end{aligned} \tag{18}$$

In addition, assume that $\boldsymbol{\Sigma}_\epsilon = \mathbb{E}\left(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t'\right)$ satisfies Assumption 1 and can be decomposed as

$$\boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} \boldsymbol{\Sigma}_{\epsilon,11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\epsilon,22} \end{bmatrix}, \text{ with } \boldsymbol{\Sigma}_{\epsilon,11} = \begin{bmatrix} \sigma_{1,11} & \boldsymbol{\sigma}_{1,21}' \\ \boldsymbol{\sigma}_{1,21} & \boldsymbol{\Sigma}_{1,22} \end{bmatrix}. \tag{19}$$

Then, the quantities appearing in the single-equation model in (7) take on the form

$$\begin{aligned} \boldsymbol{\pi}_0 &= \begin{bmatrix} \boldsymbol{\Sigma}_{1,22}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\epsilon,22}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\sigma}_{1,21} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_{0,1} \\ \mathbf{0} \end{bmatrix}, \\ \boldsymbol{\delta} &= \begin{bmatrix} \boldsymbol{\Pi}_{11}' & \boldsymbol{\Pi}_{21}' \\ \boldsymbol{\Pi}_{12}' & \boldsymbol{\Pi}_{22}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Pi}_{11}' \\ \boldsymbol{\Pi}_{12}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_{0,1} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \boldsymbol{\delta}_2 \end{bmatrix}, \\ \boldsymbol{\pi}_j &= \begin{bmatrix} \boldsymbol{\Phi}_{j,11}' & \boldsymbol{\Phi}_{j,21}' \\ \boldsymbol{\Phi}_{j,12}' & \boldsymbol{\Phi}_{j,22}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_{j,11}' \\ \boldsymbol{\Phi}_{j,12}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_{0,1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_{j,1} \\ \boldsymbol{\pi}_{j,2} \end{bmatrix}. \end{aligned} \tag{20}$$

The definitions in (20) demonstrate that, under the restriction that the errors driving $\boldsymbol{z}_{1,t}$ and $\boldsymbol{z}_{2,t}$ are uncorrelated, sparsity in the single-equation model arises when (a subset of) $\boldsymbol{z}_{2,t}$ does not Granger-Cause $\boldsymbol{z}_{1,t}$. For example, in the extreme case where $\boldsymbol{\Pi}_{12} = \mathbf{0}$ and $\boldsymbol{\Phi}_{12} = \mathbf{0}$, we have $\boldsymbol{\delta}_2 = \mathbf{0}$ and $\boldsymbol{\pi}_{j,2} = 0$, respectively. Consequently, then the single-equation model reads as

$$\begin{aligned} \Delta y_t &= \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \sum_{j=1}^{p} \boldsymbol{\pi}_j' \Delta \boldsymbol{z}_{t-j} + \epsilon_{y,t} \\ &= \boldsymbol{\delta}_1' \boldsymbol{z}_{1,t-1} + \boldsymbol{\pi}_{0,1}' \Delta \boldsymbol{x}_{1,t} + \sum_{j=1}^{p} \boldsymbol{\pi}_{1,j}' \Delta \boldsymbol{z}_{1,t-j} + \epsilon_{y,t}. \end{aligned} \tag{21}$$

As an interesting special case, consider the decomposition in (18) in which $z_{2,t} = \epsilon_{2,t}$ is scalar-valued with $\mathbb{E}(\epsilon_{2,t} \boldsymbol{\epsilon}_{1,t}) = \mathbf{0}$. Then, it is straightforward to see that $\boldsymbol{\pi}_{12} = \boldsymbol{\pi}_{21} = \mathbf{0}$, $\pi_{22} = -1$ and, consequently, $\delta_N = 0$. This finding highlights that stationary variables result in sparsity in $\boldsymbol{\delta}$ only when they are fully exogenous, as said variables may enter the implied cointegrating vector through their correlation structure with the other variables in the system. This further demonstrates the difficulty of direct interpretation of $\boldsymbol{\delta}$ without imposing additional restrictions on the DGP. From a prediction perspective, however, the model's ability to include stationary variables through their correlation structure is clearly a desirable feature.

Finally, we consider a DGP in which $\boldsymbol{\Sigma}_\epsilon$ follows a Toeplitz structure with $\sigma_{\epsilon,ij} = \rho^{|i-j|}$. After partitioning $\boldsymbol{\Sigma}_\epsilon$ as in (5), we can rewrite

$$\boldsymbol{\sigma}_{\epsilon,21} = \begin{bmatrix} \rho^1 \\ \vdots \\ \rho^{N-1} \end{bmatrix} = \begin{bmatrix} \rho^0 & \cdots & \rho^{N-2} \\ \vdots & \ddots & \vdots \\ \rho^{N-2} & \cdots & \rho^0 \end{bmatrix} \begin{bmatrix} \rho^1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{\Sigma}_{\epsilon,22} \boldsymbol{\pi}_0, \tag{22}$$

thus showing that $\boldsymbol{\pi}_0 = \boldsymbol{\Sigma}_{\epsilon,22}^{-1}\boldsymbol{\sigma}_{\epsilon,21} = (\rho, 0, \ldots, 0)'$.[6] As $\boldsymbol{\delta}' = (1, -\boldsymbol{\pi}_0')\boldsymbol{A}\boldsymbol{B}'$, this implies that only the long-run equilibria that occur in the equations for Δy_t or its cross-sectionally neighbouring variable will be part of the linear combination in the derived the single-equation model. Consequently, any variables in the dataset that are not contained in the equilibria occurring in these equations will induce sparsity in $\boldsymbol{\delta}$.

4.3.2. Mixed orders of integration

One of the most prominent benefits of SPECS is the ability to model potentially non-stationary and cointegrated data without the need to adopt a pre-testing procedure with the aim of checking, and potentially correcting, for the order of integration or to decide on the appropriate cointegrating rank of the system. The assumptions under which our theory is developed are compatible with a wide variety of DGPs, including settings where the dataset contains an arbitrary mix of $I(1)$ and $I(0)$ variables. The researcher simply transforms the dataset according to (7) and SPECS provides consistent estimation of the parameters and identification of the correct implied sparsity pattern. The purpose of this section is to demonstrate this feature by means of some illustrative examples.

The central idea underlying the above feature is that a single-equation model can be derived from any system admitting a finite order VECM representation. In a VECM system containing variables with mixed orders of integration, however, each stationary variable adds an additional trivial cointegrating vector. Such a vector corresponds to a unit vector that equals 1 on the index of the stationary variable. For illustrative purposes, we consider the following general example. Define $\boldsymbol{z}_t = (\boldsymbol{z}_{1,t}', \boldsymbol{z}_{2,t}')'$, where $\boldsymbol{z}_{1,t} \sim I(0)$ and $\boldsymbol{z}_{2,t} \sim I(1)$ and possibly cointegrated. Let the dimensions of $\boldsymbol{z}_{1,t}$ and $\boldsymbol{z}_{2,t}$ be N_1 and N_2 respectively. Then, $\boldsymbol{z}_t$ admits the representation

$$\begin{bmatrix}\Delta\boldsymbol{z}_{1,t}\\ \Delta\boldsymbol{z}_{2,t}\end{bmatrix} = \begin{bmatrix}-\boldsymbol{I}_{N_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Pi}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{z}_{1,t-1}\\ \boldsymbol{z}_{2,t-2}\end{bmatrix} + \boldsymbol{\Phi}(L)\Delta\boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{z}_{t-1} + \boldsymbol{\Phi}(L)\Delta\boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t, \tag{23}$$

where $\boldsymbol{\Phi}(L)$ corresponds to a p-dimensional matrix lag polynomial by Assumption 2 and $\boldsymbol{\epsilon}_t$ satisfies the conditions in Assumption 1. As long as the design of (23) conforms to Assumptions 2 and 3, our main results apply to this setting and both selection and estimation consistency is maintained. For the extreme case in which all variables are integrated of order one, but none are cointegrated, we define $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{0}$. Clearly, it follows that $\boldsymbol{\delta} = \boldsymbol{0}$, such that the single-equation model can be seen as a conditional model obtained from a VAR specified in differences. In the other extreme case, when the levels of all variables in the VECM are weakly stationary, decomposition (23) would simply lead to a VECM in which $-\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{I}_N$, thereby enabling the results in Section 4.1 to carry through.[7]

4.3.3. Rates of convergence

We conclude our theoretical analysis with a detailed illustration of the attainable rates of convergence in different asymptotic frameworks. The rates of convergence of $\hat{\boldsymbol{\gamma}}_R$ and $\hat{\boldsymbol{\gamma}}$, as well as the specific construction of the initial weights, are dependent on the growth rates of $N, p, r, |S_\delta|$ and $|S_\pi|$. Because of the trade-off between the admissible dimension and the rate of convergence, the choice of the desired asymptotic framework is likely dependent on the specific application. For example, typical macro-economic applications are characterized by short panel datasets which would require a framework in which the cross-sectional dimension grows as fast as theoretically admissible. On the other hand, in applications with a large number of time series observations, such as forecasting based on high-frequency data, the assumption that the number of (potentially) relevant variables grows slow relative to the available time periods seems reasonable. Therefore, to aid interpretation of our results, we provide an overview with different asymptotic frameworks and the corresponding penalty parameters, weight constructions and convergence rates of the initial estimator in Table 1. The weights for δ_i and π_j are constructed as $\omega_i = \left|\hat{\delta}_{R,i}\right|^{-k_\delta}$ and $\omega_{N+j} = \left|\hat{\pi}_{R,j}\right|^{-k_\pi}$.

The first row of Table 1 corresponds to the classic fixed-dimensional case. It is reassuring that, similar to the OLS estimator, SPECS obtains $\sqrt{T}$-convergence, with the additional benefit of allowing for consistent recovery of the sparsity pattern. In fact the next three rows highlight that when N, p or r diverge, while the number of relevant variables remains fixed, SPECS maintains its $\sqrt{T}$-convergence as long as the penalty weights k_δ and k_π are adjusted appropriately. In the fifth row, we allow the number of relevant stationary variables, i.e. $|S_\pi|$ to diverge as well. This setting may be preferred when the integrated time series remain persistent after being transformed to stationarity by differencing. We observe that consistency is maintained, although even sharper weights are required and the rate of convergence has reduced to $T^{3/8}$. In the sixth row we additionally allow the number of relevant non-stationary variables, i.e. $|S_\delta|$, to increase, whereas the number of cointegrating vectors remains fixed. The increased number of non-zero coefficients corresponding to non-stationary variables reduces the rate of convergence to $T^{1/4}$. Interestingly, in the last row we let the dimension of the cointegrating subspace r grow at the same rate. As illustrated in Remark 10, this setting naturally occurs when the data is modelled by a non-stationary factor model with idiosyncratic components. In this framework, the number of stochastic trends driving the subset of relevant variables, i.e. s_δ, remains fixed, which positively affects the convergence rate of SPECS.

[6] It is straightforward to show that this property carries over to covariance matrices with a block-diagonal Toeplitz structure, with each block $\boldsymbol{\Sigma}_\epsilon^{(k)}$ having the form $\sigma_{i,j}^{(k)} = \rho_{(k)}^{|i-j|}$. The number of non-zero elements in the resulting vector $\boldsymbol{\pi}_0$ will equal the number of blocks in the covariance matrix.

[7] When all variables are stationary, SPECS can also be shown to consistently estimate the parameters based on the well-documented properties of the adaptive lasso in stationary time series settings, such as those considered in Medeiros and Mendes (2016) and Masini et al. (2019).

Table 1
Dimensions, Penalties, Weights and convergence rates.

N	p	r	$\lvert S_\delta \rvert$	$\lvert S_\pi \rvert$	k_δ	k_π	λ_R, λ_I	$\left\lVert \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \right\rVert_2$
Fixed	Fixed	Fixed	Fixed	Fixed	2	1	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	Fixed	Fixed	Fixed	Fixed	3	1	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	$T^{1/4}$	Fixed	Fixed	Fixed	3	2	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	Fixed	Fixed	3	2	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	Fixed	$T^{1/4}$	4	2	$KT^{2/5}$	$O_p\left(T^{-3/8}\right)$
$T^{1/4}$	$T^{1/4}$	Fixed	$T^{1/4}$	$T^{1/4}$	4	2	$KT^{2/5}$	$O_p\left(T^{-1/4}\right)$
$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	4	2	$KT^{2/5}$	$O_p\left(T^{-3/8}\right)$

This table displays possible settings for the weights (k_δ, k_π) and penalty parameters (λ_I, λ_R) that satisfy Assumption 6 under a variety of asymptotic frameworks $(N, r, p, |S_\delta|, |S_\pi|)$. The convergence rate of SPECS is displayed in the last column.

We consider the theoretical results presented in this section to be of a double nature. On the one hand, it is reassuring that consistent estimation remains feasible in growing dimensions and that suitable weights are available. On the other hand, we acknowledge that the required restrictions on the growth rate of the number of variables seem to caution against application of penalized regression in very high-dimensional settings. However, it is worth noting that the restrictions on N and p largely result from the use of ridge regression as an initial estimator. Indeed, the availability of a novel compatibility condition could justify the use of the lasso as an initial estimator and will allow for generalization of our theoretical results to even higher dimensional asymptotic frameworks. We consider this an interesting avenue for future research.

Remark 10. The VECM (18) can be rewritten into a non-stationary factor model with stationary idiosyncratic components, similar to Banerjee et al. (2014). Based on the VMA representation of $\boldsymbol{z}_t$ in (4), with $\boldsymbol{C}$ a matrix of reduced rank, we can rewrite the process as

$$\boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{u}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{u}_t, \tag{24}$$

where $\boldsymbol{\Lambda} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I} - \sum_{j=1}^p \boldsymbol{\Phi}_j\right) \boldsymbol{B}_\perp\right)^{-1}$, $\boldsymbol{f}_t = \boldsymbol{A}_\perp' \boldsymbol{s}_t$ and $\boldsymbol{u}_t = \boldsymbol{C}(L)\boldsymbol{\epsilon}_t + \boldsymbol{z}_0$. This representation is particularly relevant in relation to the growth rate of $N_\delta = N - r$. Typically, the theory for consistent estimation of (24) is derived under the assumption that the N_δ factors remain fixed, while letting both N and T go to infinity. Hence, in this framework, noting that $s_\delta \leq N_\delta$, the assumptions that $\frac{s_\delta}{T^{1/4}} \to 0$ and $\frac{N_\delta}{T^{1/4}} \to 0$ in Theorems 1–3 are automatically satisfied. Consequently, the convergence rates of the initial and final estimators are given by $\left\lVert \hat{\boldsymbol{\gamma}}_R - \boldsymbol{\gamma} \right\rVert_2 = O_p\left(\sqrt{\frac{M_\pi}{T}}\right)$ and $\left\lVert \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \right\rVert_2 = O_p\left(\sqrt{\frac{s_\pi}{T}}\right)$.

5. Simulations

In this section we analyse the selective capabilities and predictive performance of SPECS by means of simulations. We estimate the single-equation model according to the objective function (14) with the following settings for the penalty rates:

1. Ordinary Least Squares (OLS: $\lambda_G = 0$, $\lambda_I = 0$),
2. Autoregressive Distributed Lag (ADL: $\lambda_G = 0$, $\lambda_I > 0$, $\omega_i = \infty$ for $i = 1, \ldots, N$),
3. SPECS - no group penalty (SPECS$_1$: $\lambda_G = 0$, $\lambda_I > 0$),
4. SPECS - group penalty (SPECS$_2$: $\lambda_G > 0$, $\lambda_I > 0$).[8]

The OLS estimator is only included when feasible according to the dimension of the model to estimate and we additionally include a penalized autoregressive distributed lag model (ADL) with all variables entering in first differences. The latter model can be interpreted as the conditional model one would obtain when ignoring cointegration in the data and specifying a VAR in differences as a model for the full system. The resulting conditional model is the same as the CECM that we consider, but with the built-in restriction $\boldsymbol{\delta} = \boldsymbol{0}$.

We estimate the solutions for a grid of penalty values and construct the weights from an initial ridge estimator as proposed in Section 4.2. For ADL and SPECS$_1$, we consider 100 possible values for λ_I and choose the final model based on the BIC criterion. Alternatively, for SPECS$_2$, the model selection takes place over a two-dimensional grid consisting of 100 values for λ_I and 10 possible values for λ_G, with model selection again being based on the BIC criterion. The weights are defined by $\omega_i = \left|\hat{\gamma}_{R,i}\right|^{-k}$, where $k = 2$ for $i \in \{1, \ldots, N\}$ and $k = 1$ for $i \in \{N+1, \ldots, N+M\}$.

[8] As a helpful reminder, the reader may relate the subscript to the number of penalty categories included in the estimation; SPECS$_1$ only contains an individual penalty whereas SPECS$_2$ contains both a group penalty and individual penalty.

Table 2
Simulation design for the first study (Dimensionality and weak exogeneity).

Low dimension	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{\delta}$
WE	$a \cdot \begin{bmatrix} 1 \\ \mathbf{0}_{9\times 1} \end{bmatrix}$	$\begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{5\times 1} \end{bmatrix}$	$a \cdot \boldsymbol{B}$
No WE	$a \cdot \boldsymbol{B}$	$\begin{bmatrix} \tilde{\iota} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \tilde{\iota} \end{bmatrix}$	$(1+\rho)a \cdot \begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{5\times 1} \end{bmatrix}$
High dimension	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{\delta}$
WE	$a \cdot \begin{bmatrix} 1 \\ \mathbf{0}_{49\times 1} \end{bmatrix}$	$\begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$	$a \cdot \boldsymbol{B}$
No WE	$a \cdot \boldsymbol{B}$	$\begin{bmatrix} \tilde{\iota} & \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \tilde{\iota} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 1} & \tilde{\iota} \\ \mathbf{0}_{35\times 1} & \mathbf{0}_{35\times 1} & \mathbf{0}_{35\times 1} \end{bmatrix}$	$(1+\rho)a \cdot \begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$

Notes: The low-dimensional (high-dimensional) design corresponds to a system with $N = 10$ ($N = 50$) unique time series and $N' = 31$ ($N' = 151$) parameters to estimate. Furthermore, $\tilde{\iota} = (1, -\iota_4')'$ and $a = -0.5, -0.45, \ldots, 0$ regulates the adjustment rate towards the equilibrium.

We consider three different settings under which we analyse the performance of our estimators; the first setting aims to analyse the effects of dimensionality and weak exogeneity, the second setting explores the effect of the variables' orders of integration and the third setting considers the performance in non-sparse settings. Each setting is described in detail below.

5.1. Dimensionality and weak exogeneity

In the first part of our simulation study we focus on the effects of dimensionality and weak exogeneity on a (co)integrated dataset. Our simulation DGP takes the form

$$\Delta \boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{z}_{t-1} + \boldsymbol{\Phi}_1 \Delta \boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t, \tag{25}$$

with $t = 1, \ldots, T = 100$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and $\sigma_{ij} = 0.8^{|i-j|}$. Furthermore, $\boldsymbol{\Phi}_1$, the coefficient matrix regulating the short-run dynamics is generated as $0.4 \cdot \boldsymbol{I}_N$, where N varies depending on the specific DGP considered. Based on this DGP, the single-equation model takes on the form

$$\Delta y_t = \boldsymbol{\delta}'\boldsymbol{z}_{t-1} + \boldsymbol{\pi}_0'\Delta \boldsymbol{x}_t + \boldsymbol{\pi}_1'\Delta \boldsymbol{z}_{t-1} + \epsilon_{y,t},$$

with $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ as defined in (7). We consider a total of four different settings, corresponding to different combinations of (i) dimensionality (low/high) and (ii) weak exogeneity (present/absent). The corresponding parameter settings and implied cointegrating vector $\boldsymbol{\delta}$ are given in Table 2.

We measure the selective capabilities based on three metrics. The pseudo-power of the models measures the ability to appropriately pick up the presence of cointegration in the underlying DGP. For the OLS procedure we perform the Wald test proposed by Boswijk (1994). When the OLS fitting procedure is unfeasible due to the high-dimensionality, we perform the Wald test on the subset of variables included after fitting SPECS_1 and refer to this approach as Wald-PS (where PS stands for post-selection). Despite the caveats of oracle-based post-selection inference discussed after Corollary 1, the inclusion of Wald-PS still offers valuable insights regarding the performance one may expect of such a procedure in light of the aforementioned limitation. SPECS is used as an alternative to this cointegration test by simply checking whether at least one of the lagged levels is included in the model. The percentage of trials in which cointegration is found is then reported as the pseudo-power.

Second, for each trial the Proportion of Correct Selection (PCS) measures the proportion of correctly selected variables, while the Proportion of Incorrect Selection (PICS) describes, as the name may suggest, the proportion of incorrectly selected variables. They are given by

$$PCS = \frac{\left|\{j : \hat{\gamma}_j \neq 0\} \cap \{j : \gamma_j \neq 0\}\right|}{\left|\{j : \gamma_j \neq 0\}\right|}; \qquad PICS = \frac{\left|\{j : \hat{\gamma}_j \neq 0\} \cap \{j : \gamma_j = 0\}\right|}{\left|\{j : \gamma_j = 0\}\right|}.$$

The PCS and PICS are calculated for SPECS_1 and SPECS_2 and averaged over all trials.

Finally, we consider the predictive performance in a simulated nowcasting application, where we implicitly assume that the information on the latest realization of $\boldsymbol{x}_T$ arrives before the realization of y_T. These situations frequently occur in practice, see Giannone et al. (2008) and the references therein for an overview as well as the empirical application considered in Section 6. Due to the construction of the single-equation model, in which contemporaneous values of the conditioning variables contribute to the contemporaneous variation in the dependent variable, our proposed method is particularly well-suited to this application. For any of the considered fitting procedures, the nowcast is given by

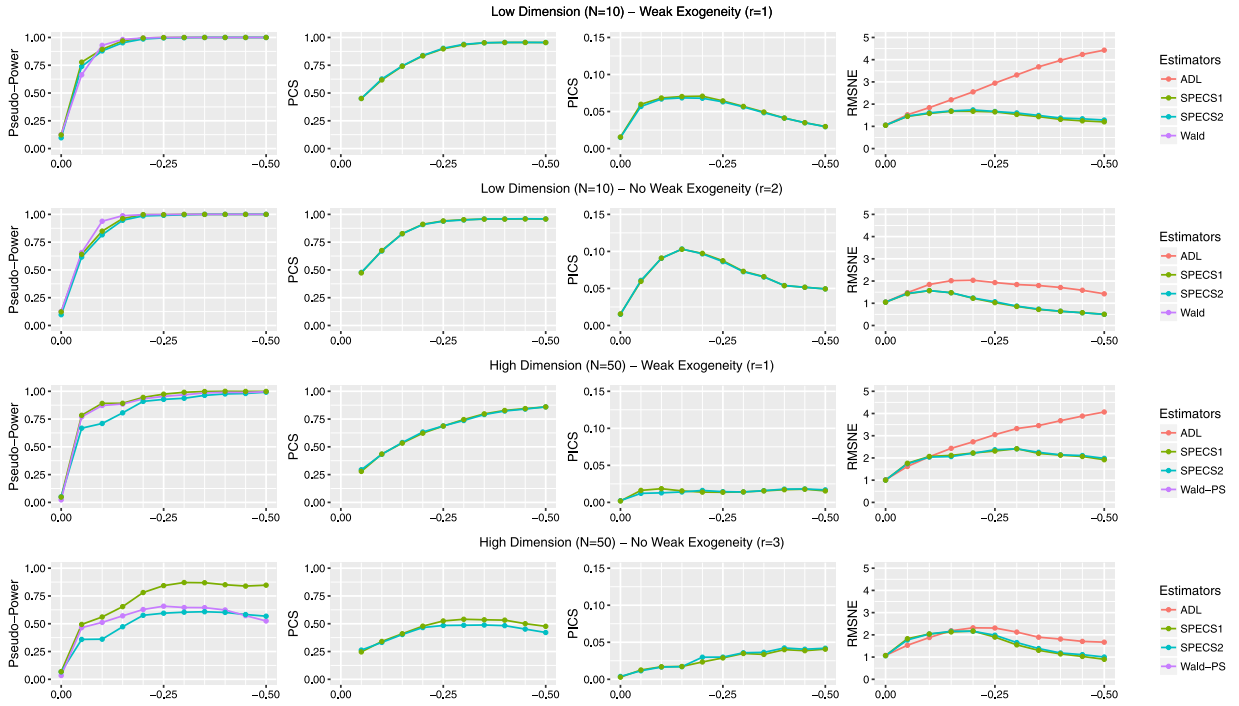

Fig. 1. Pseudo-Power, Proportion of Correct Selection (PCS), Proportion of Incorrect Selection (PICS) and Root Mean Squared Nowcast Error (RMSNE) for Low- and High-Dimensional specifications. The adjustment rate multiplier a is on the horizontal axis.

$\hat{y}_T = \hat{\boldsymbol{\delta}}'\boldsymbol{z}_{T-1} + \hat{\boldsymbol{\pi}}'\Delta\boldsymbol{x}_T + \hat{\boldsymbol{\phi}}'\Delta\boldsymbol{z}_{T-1}$, where by construction $\hat{\boldsymbol{\delta}} = \mathbf{0}$ in the ADL model. For each method we record the root mean squared nowcast error (RMSNE) relative to the OLS oracle procedure fitted on the relevant variables.

Fig. 1 visually displays the evolution of our performance metrics over a range of values for a, representing increasingly faster rates of adjustment towards the long-run equilibrium. The first row of plots shows near-perfect performance of SPECS over all metrics. The pseudo-size is slightly lower than the size of the Wald test when the latter is controlled at 5%, whereas the pseudo-power quickly approaches one. Following expectations, the pseudo-size for SPECS_2 is slightly lower as a result of the additional group penalty. Focusing on the selection of variables, we find that for faster adjustment rates, SPECS is able to exactly identify the sparsity pattern with very high frequency, as demonstrated by the PCS approaching 100% and the PICS staying near 0%. Furthermore, the MSNE obtained by our methods is close to the OLS oracle method and is substantially lower than the MSNE obtained by the ADL model for faster adjustment rates, while being almost identical absent of cointegration. The picture remains qualitatively similar when moving away from weak exogeneity while staying in a low-dimensional framework, although the gain in predictive performance over the ADL has decreased somewhat. We postulate that the ADL may benefit from a bias–variance tradeoff, given that the correctly specified single-equation model is sub-optimal in terms of efficiency absent of weak exogeneity compared to a full system estimator. Nonetheless, SPECS is clearly preferred.

The performance in the high-dimensional setting is displayed in rows 3 and 4 of Fig. 1. When the conditioning variables are weakly exogenous with respect to the parameters of interest, the selective capabilities remain strong. The pseudo-power demonstrates the attractive prospect of using our method as an alternative to cointegration testing, especially when taking into consideration that the traditional Wald test is infeasible in the current setting. In addition, the nowcasting performance remains far superior to that of the misspecified ADL. The last row depicts the performance absent of weak exogeneity. In this setting, exact identification of the implied cointegrating vector occurs less frequently, which seems to negatively impact the nowcasting performance. However, the misspecified ADL is still outperformed, despite the deterioration in the selective capabilities of our method.

5.2. Mixed orders of integration

We next analyse the performance of SPECS on datasets containing variables with mixed orders of integration. The aim of this section is to gain an understanding of the relative performance of SPECS when not all time series are (co)integrated and to compare the performance of SPECS to traditional approaches that rely on pre-testing. The latter goal is attained by adding an additional penalized ADL model to the comparison, namely one in which the data is first corrected for non-stationarity based on a pre-testing procedure in which an Augmented Dickey–Fuller (ADF) test is performed on the

Table 3
Simulation design for the second study (Mixed orders of integration).

Mixed order	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{\delta}$
$y \sim I(0)$	$\begin{bmatrix} 1 & 0 & \mathbf{0}_{1\times 24} \\ \mathbf{0}_{15\times 1} & a\boldsymbol{B}^* & \mathbf{0}_{15\times 24} \\ \mathbf{0}_{10\times 1} & \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 24} \\ \mathbf{0}_{24\times 1} & \mathbf{0}_{24\times 3} & \boldsymbol{I}_{24} \end{bmatrix}$	$\begin{bmatrix} -b & 0 & \mathbf{0}_{1\times 24} \\ \mathbf{0}_{15\times 1} & \boldsymbol{B}^* & \mathbf{0}_{15\times 24} \\ \mathbf{0}_{10\times 1} & \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 24} \\ \mathbf{0}_{24\times 1} & \mathbf{0}_{24\times 3} & -\boldsymbol{B}_{24\times 24} \end{bmatrix}$	$\begin{bmatrix} -1 \\ -\rho a\tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{44\times 1} \end{bmatrix}$
$y \sim I(1)$	$\begin{bmatrix} a\boldsymbol{B}^* & \mathbf{0}_{15\times 25} \\ \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 25} \\ \mathbf{0}_{25\times 3} & \boldsymbol{I}_{25} \end{bmatrix}$	$\begin{bmatrix} \boldsymbol{B}^* & \mathbf{0}_{15\times 25} \\ \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 25} \\ \mathbf{0}_{25\times 3} & -\tilde{\boldsymbol{B}}_{25\times 25} \end{bmatrix}$	$(1+\rho)a \cdot \begin{bmatrix} \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$,

Notes: see notes in Table 2. Additionally, we define $b = 1$ ($b \sim U(0, 0.2)$) and $\tilde{\boldsymbol{B}}$ as a diagonal matrix with $b_{ii} = 1$ ($b_{ii} \sim U(0, 0.2)$) in the absence (presence) of persistence, and $\boldsymbol{B}^* = (\mathbf{1}_{3\times 3} \otimes \tilde{\boldsymbol{\iota}})$.

individual series. We refer to this procedure as the ADL-ADF model. Based on the general DGP (25), we distinguish four different cases, corresponding to (i) different orders of the dependent variable ($I(0)/I(1)$) and (ii) different degrees of persistence in the stationary variables (low/high). The choice to include varying degrees of persistence is motivated by the conjecture that the performance of the pre-testing procedure incorporated in the ADL-ADF model may deteriorate when the degree of persistence increases, which in turn translates to a decrease in the overall performance of the procedure.

The parameter settings for the varying DGPs, displayed in Table 3, are chosen such that they allow for a subset of stationary variables in the system. In particular, we first consider a scenario in which the dependent variable itself admits a stationary autoregressive representation in levels. In addition, based on their cross-sectional ordering, the first 15 variables after y are cointegrated based on three cointegrating vectors, the next 10 variables are non-cointegrated random walks, and the last 24 variables all admit a stationary autoregressive structure in levels. The degree of persistence in the stationary variables is regulated by the diagonal matrix $\tilde{\boldsymbol{B}}$ in $\boldsymbol{B}$, with elements $b_{ii} = 1$ in the low persistence case and $b_{ii} \sim U(0, 0.2)$ in the high persistence case. It can be seen from the last column in Table 3, that in line with the stationarity of the dependent variable, the first element in $\boldsymbol{\delta}$ will always be equal to -1, whereas an additional five-dimensional cointegrating vector enters the single-equation model for positive values of a. For the scenario in which the dependent variable is integrated of order one, the first 15 variables (including y) are all cointegrated based on three cointegrating vectors, the next 10 variables are non-cointegrated random walks, whereas the last 15 variables all admit a stationary autoregressive representation. The persistence in the stationary variables is regulated similar to the previous case. Now, however, it is clear from the last column in Table 3 that $\boldsymbol{\delta} \neq \mathbf{0}$ only if $a > 0$, such that lagged levels only enter the single-equation when y is cointegrated with its neighbouring variables. We display the performance of the models in Fig. 2.

In the first row of Fig. 2, corresponding to $y \sim I(0)$ and low persistence, SPECS correctly selects the lagged dependent variable in all simulation trials, such that the pseudo-power is always 1. Interestingly, PCS also seems constant around 35%. Upon closer inspection, we find that SPECS chooses an alternative representation of the single-equation model in which the contribution of the non-trivial cointegrating vector seems to be absorbed in the lagged level of the dependent variable. While the resulting model differs from the implied oracle model, which is indeed accurately estimated by the OLS oracle procedure, the model choice seems motivated by a favourable bias–variance trade-off. In line with this conjecture, the nowcast performance of SPECS occasionally exceeds the OLS oracle procedure's where a larger number of parameters is estimated. The standard ADL nowcasts are again inferior, whereas the ADL-ADF model seems to benefit from correct identification of the stationarity of the dependent variable, which is particularly relevant given that the dependent variable itself is a main component in the optimal forecast. However, the nowcast accuracy of SPECS is almost identical to that of the ADL-ADF model, a finding that we interpret as reassuring and confirmatory of our claim that SPECS may be used without any pre-testing procedure. Moreover, the absence of strong persistence in the stationary variables idealizes the results of the ADL-ADF procedure.

In typical macroeconomic applications many time series that are considered as I(0) display much slower mean reversion and, consequently, are more difficult to correctly identify as being stationary.[9] Accordingly, in row 2 we display the result for a DGP where the stationary variables display more persistent behaviour. The performance of SPECS remains largely unaffected, whereas the nowcasting performance of the ADL-ADF model deteriorates drastically. We stress the relevance of this result, given that the estimation of ADL models after pre-testing for non-stationarity is fairly common practice. Somewhat surprisingly, the ADL model in differences nowcasts almost as well as SPECS here. Overall, however, the nowcast accuracy of SPECS remains the highest and, equally important, most stable across all specifications.

Continuing the analysis of mixed order datasets, rows 3 and 4 of Fig. 2 display the results for DGPs where the dependent variable is generated as being integrated of order one. The pseudo-power plot clearly reflects that $\boldsymbol{\delta} \neq \mathbf{0}$ only when $a > 0$. Furthermore, while SPECS performs well at removing the irrelevant variables, the relevant variables are not all selected correctly, resulting in somewhat lower values for the PCS metric. Nevertheless, the nowcast performance remains superior to that of the ADL model, especially in the presence of cointegration with fast adjustment rates.

[9] For example, the ten time series in the popular Fred-MD dataset which McCracken and Ng (2016) propose to be I(0), i.e. the series corresponding to a tcode of one, all display strong persistence or near unit root behaviour, with the smallest estimated AR(1) coefficient exceeding 0.86.

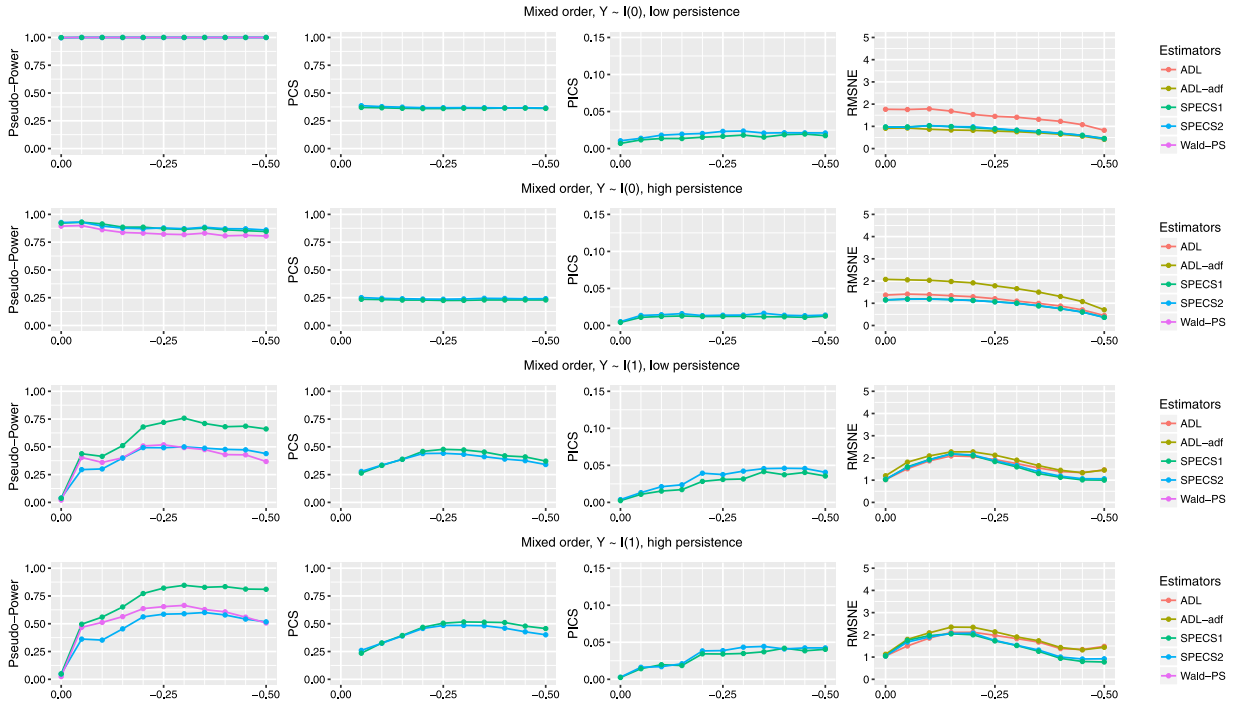

Fig. 2. Pseudo-Power, Proportion of Correct Selection (PCS), Proportion of Incorrect Selection (PICS) and Root Mean Squared Nowcast Error (RMSNE) for four Mixed Order specifications. The adjustment rate multiplier a is on the horizontal axis.

5.3. Non-sparse data generating processes

To avoid idealizing the results through a choice of DGPs that suits our estimator, this section considers the performance of the penalized regression estimators in two different non-sparse settings. First, we consider an explicitly constructed VECM that contains many small, but non-zero coefficients. Second, we consider a DGP that contains a non-stationary factor structure on which the single-equation model is likely misspecified.

The non-sparse VECM is generated according to (25) with $\boldsymbol{B} = \boldsymbol{I}_3 \otimes \tilde{\iota}$, where $\tilde{\iota} = (1, -\iota_4')'$, and $\boldsymbol{A} = a\boldsymbol{B}$ for $a = 0, -0.05, \ldots, -0.5$. Hence, $N = 15$ and the total number of parameters to estimate (including a constant and linear trend) is $N(p+2)+1 = 46$. A major difference with Section 5.1 is that we do not generate the covariance matrix of the errors as a Toeplitz-matrix, the latter being a crucial driver of sparsity in the preceding sections. Instead, we implement the procedure detailed in Chang (2004, p. 277–278), in which we generate a $(N \times N)$ matrix $\boldsymbol{U}$ with $u_{ij} \sim U(0, 1)$ to construct the orthonormal matrix $\boldsymbol{H} = \boldsymbol{U}\left(\boldsymbol{U}'\boldsymbol{U}\right)^{-1/2}$, and generate a set of N eigenvalues, $\lambda_1, \ldots, \lambda_N$, where $\lambda_1 = 0.01$, $\lambda_N = 1$ and $\lambda_2, \ldots, \lambda_{N-1} \sim U(0.1, 1)$ to construct $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_N)$. We then construct the covariance matrix as $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$. At each simulation trial, we generate a new $\boldsymbol{\Sigma}$ such that the results cannot be attributed to a specific random draw of the covariance matrix. Based on this construction, $\boldsymbol{\pi}_0$, as defined below (7), and $\boldsymbol{\delta}$ are non-sparse vectors with small elements; even in the setting with the strongest cointegration, i.e. $a = -0.5$, the median magnitude of the coefficients in $\boldsymbol{\delta}$ across all trials is only 0.12. As before, we set $T = 100$ and perform 1000 simulation trials.

The results are displayed in Fig. 3, which contain a number of interesting results. Unsurprisingly, all estimators obtain a substantially lower (pseudo-)power in the current framework. The ℓ_1-regularized estimators seem more sensitive to this than the traditional Wald estimator considered in Boswijk (1994). In line with the weak power, we observe that the PCS for both SPECS$_1$ and SPECS$_2$ is low, with on average only 0.75 out of 15 variables being included in levels.[10] Appropriate inference in the current setting is a difficult task and direct application of SPECS without alteration does not seem to be a feasible strategy. The development of a uniformly valid post-selection inference procedure, such as the desparsified lasso of Van de Geer et al. (2014), may alleviate some of these issues. While we consider this an interesting avenue of research, it is outside the scope of the current paper.

While these results may seem discouraging, the results on the nowcast accuracy display a different story. The mean-squared nowcast errors, relative to the OLS oracle procedure, are almost always below one and are similar for the SPECS and penalized ADL estimators. This highlights that the signal of the long-run component is so weak, that the estimation of a misspecified model which ignores cointegration benefits from a favourable bias–variance tradeoff. Therefore, the

[10] The PICS is zero for all $a > 0$, simply because the DGP is non-sparse, and is omitted accordingly.

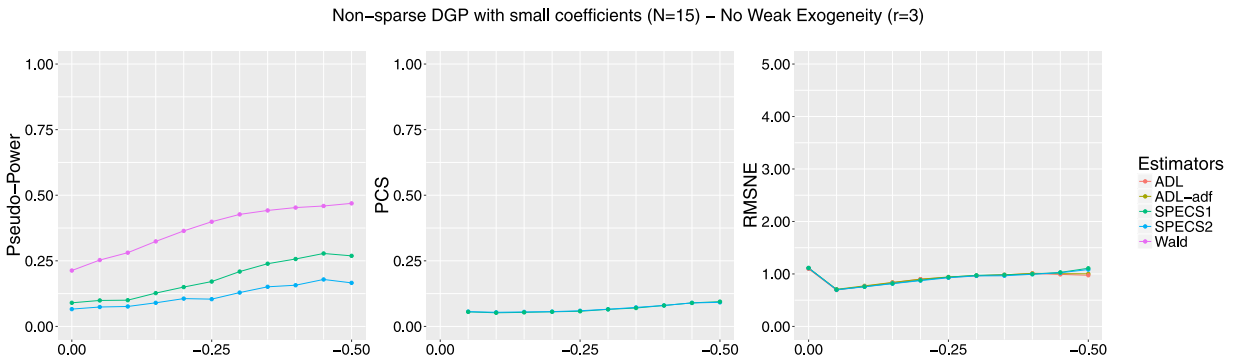

Fig. 3. Pseudo-Power, Proportion of Correct Selection (PCS), Proportion of Incorrect Selection (PICS) and Root Mean Squared Nowcast Error (RMSNE). The adjustment rate multiplier a is on the horizontal axis.

Table 4
Nowcasting performance on a DGP with a non-stationary factor.

	Root Mean Squared Nowcast Error		
	$SPECS_1$	$SPECS_2$	$SPECS_1$ - OLS
No dynamics	1.07	1.11	0.99
Dynamics	1.02	1.02	1.01

conclusion remains that SPECS obtains superior predictive performance relative to methods that ignore cointegration when the long-run component provides a strong signal, without sacrificing performance absent of cointegration or in the presence of very weak cointegration.

The second, and final, DGP that we consider contains a non-stationary factor structure and corresponds to setting III in Palm et al. (2011, p. 92). We allow for contemporaneous correlation and dynamic structures in both the error processes driving the "observable" data and the idiosyncratic component in the factor structure. The DGP is given by $\boldsymbol{z}_t = \boldsymbol{\lambda} f_t + \boldsymbol{\omega}_t$, where $\boldsymbol{z}_t$ is a (50×1) time series process, f_t is a single scalar factor and

$$f_t = \phi f_{t-1} + \zeta_t, \qquad \omega_{i,t} = \theta_i \omega_{i,t-1} + v_{i,t}.$$

Furthermore,

$$\boldsymbol{v}_t = \boldsymbol{A}_1 \boldsymbol{v}_{t-1} + \boldsymbol{\epsilon}_{1,t} + \boldsymbol{B}_1 \boldsymbol{\epsilon}_{1,t-1}, \qquad \zeta_t = \alpha_2 \zeta_{t-1} + \epsilon_{2,t} + \beta_2 \epsilon_{2,t-1},$$

where $\boldsymbol{\epsilon}_{1,t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ again generated as in Chang (2004), and $\epsilon_{2,t} \sim \mathcal{N}(0, 1)$.

The comparison focuses exclusively on the nowcasting performance for a setting without dynamics ($\boldsymbol{A}_1 = \boldsymbol{B}_1 = \boldsymbol{0}$ and $\alpha_2 = \beta_2 = 0$) and a setting with dynamics ($\alpha_2 = \beta_2 = 0.4$) in which the construction of $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ is analogous to Palm et al. (2011, p. 93). We report the RMSNEs of SPECS relative to the ADL in Table 4. Given that the single-equation model is misspecified in this setup, it is unreasonable to expect SPECS to outperform. Indeed, we observe that the RMSNEs are all very close to one and, while in most cases the ADL model performs slightly better, the difference seems negligible. Hence, the risk of using SPECS to estimate a misspecified model in the sense considered here, does not seem to be higher than the use of the alternative ADL model, whereas the relative merits of SPECS when applied to a wide range of correctly specified models are evident from the first part of the simulations.

6. Empirical application

Inspired by Choi and Varian (2012), we consider nowcasting Dutch unemployment with SPECS based on Google Trends data. Google Trends are time series consisting of normalized indices depicting the volume of search queries entered in Google, originating from a certain geographical area. The Dutch unemployment rates are made available by Statistics Netherlands, an autonomous administrative body focusing on the collection and publication of statistical information. These rates are published on a monthly basis with new releases being made available on the 15th of each new month. This misalignment of publication dates clearly illustrate a practically relevant scenario where improvements upon forward looking predictions of Dutch unemployment rates may be obtained by utilizing contemporaneous Google Trends series.

We collect a novel dataset containing seasonally unadjusted Dutch unemployment rates from the website of Statistics Netherlands[11] and a set of manually selected Google Trends time series containing unemployment related search queries, such as "Vacancy", "Resume" and "Unemployment Benefits". The dataset comprises of monthly observations ranging from

[11] http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLEN&PA=80479eng&LA=EN.

Table 5
Mean-Squared Nowcast Error relative to the ADL model for varying number of lagged differences p.

p	# of parameters	ADL-ADF	SPECS_1	SPECS_2
1	262	1.27	0.99	1.07
3	436	1.06	0.82*	0.88
6	697	0.90	0.90	0.84*

*Denotes rejection by the Diebold–Mariano test at a 10% significance level.

January 2004 to December 2017. While the full dataset contains 100 unique search queries, a number of these contain zeros for large sub-periods, indicating insufficient search volumes for those particular series. Consequently, we remove all series that are perfectly correlated over any sub-period consisting of 20% of the total sample.[12]

The benchmark model we consider is an ADL model fitted to the differenced data. In detail, let y_t and $\boldsymbol{x}_t$ be the scalar unemployment rate and the vector of Google Trends series observed at time t, respectively, and define $\boldsymbol{z}_t = (y_t, \boldsymbol{x}_t')'$. The benchmark ADL estimator fits

$$\Delta y_t = \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \sum_{j=1}^{p} \boldsymbol{\pi}_j' \Delta \boldsymbol{z}_{t-j} + \mu_0 + \tau_0(t-1) + \epsilon_t.$$

However, this estimator ignores the order of integration of individual time series by differencing the whole dataset, while it is common practice to transform individual series to stationarity based on a preliminary test for unit roots. Hence, similar to Sections 5.2 and 5.3, we include an additional ADL model where the decision to difference is based on a preliminary ADF test and refer to this method as ADL-ADF.[13] Finally, SPECS estimates

$$\Delta y_t = \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \sum_{j=1}^{p} \boldsymbol{\pi}_j' \Delta \boldsymbol{z}_{t-i} + \mu_0 + \tau_0(t-1) + \epsilon_t.$$

All tuning parameters are obtained by time series cross-validation and we use $k = 1.1$ based on a preliminary analysis.[14] The first nowcast is made by fitting the models on a window containing the first two-thirds of the complete sample, i.e. $t = 1, \ldots, T_c$ with $T_c = \lceil \frac{2}{3} T \rceil$, based on which the nowcast for Δy_{T_c+1} is produced. This procedure is repeated by rolling the window forward by one observation until the end of the sample is reached, producing a total of 54 pseudo out-of-sample nowcasts. Table 5 reports the MSNE relative to the ADL model for $p = 1, 3, 6$.

The ADL-ADF estimator does not perform better than the regular ADL model for $p = 1, 3$, indicating that the potential for errors in pre-testing might lead to unfavourable results. SPECS performs well and is able to obtain smaller mean-squared nowcast errors than the ADL benchmark across almost all specifications, with the combination SPECS_2 and $p = 1$ being the exception. Moreover, for SPECS_1 ($p = 3$) and SPECS_2 ($p = 6$), we find the differences in MSNE to be significant at the 10% level according to the Diebold–Mariano test. The overall (unreported) MSNE is lowest for the SPECS_1 estimator based on $p = 3$ lagged differences. Given that the addition of lagged levels to the models improves the nowcast performance, the premise of cointegrating relationships between Dutch unemployment rates and Google Trends series seems likely. To further explore the presence of cointegration among our time series we group our variables in five categories; (1) Application Training, (2) General, (3) Job Search, (4) Recruitment Agencies (RA) and (5) Social Security. We narrow down our focus to the nowcasts of models with three lagged difference included, $p = 3$, estimated by SPECS_1. In Fig. 4 we visually display the share of nowcasts in which the lagged levels of each variable are included in the estimated model. In addition, it depicts the selection stability of those variables, where a green colour indicates that given variables are included in the respective nowcast, and red vice versa. The figure also displays the actual unemployment rates compared to the nowcasted values.

Fig. 4 highlights that only few variables are consistently selected for all nowcasts, although in each category we can distinguish some variables that are included at higher frequencies. The variable whose lagged levels are always selected is "Vakantiebaan", which is a search query for a temporary job during the summer holiday. We postulate that this variable is selected by SPECS to account for seasonality in the Dutch unemployment rates. In an unreported exercise we estimate the model with the addition of a set of eleven unpenalized dummies representing different months of the year. While the variable "Vakantiebaan" is never selected, the mean squared nowcast error increases substantially. Hence, we opt to adhere to our standard model under the caveat that for at least one of the lagged levels included, seasonality effects rather than cointegration seem a more appropriate explanation for its inclusion. Other frequently included variables are queries for vacancies ("uwv.vacatures", 78%), unemployment ("werkloos", 76%) and social benefits ("ww uitkering",

[12] We have made the dataset available in the specs *R* package (https://cran.r-project.org/web/packages/specs/index.html).

[13] We note that none of the time series were found to be integrated of order 2. The outcome of the ADF test is reported for each time series in the online Appendix C.4.

[14] Comparing the nowcast accuracy for varying $k \in [0, 4]$, we found the highest accuracy for $k = 1.1$.

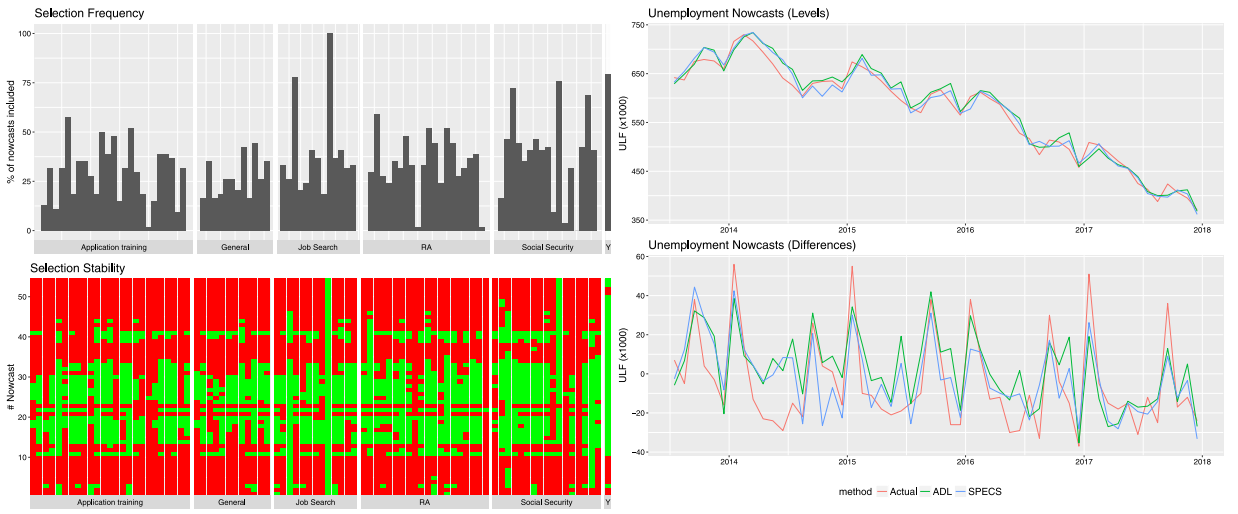

Fig. 4. *Top-left*: Selection frequency, measured as the percentage of all nowcasts the variable was selected. *Bottom-left*: Selection stability with green indicating a variable was included in the nowcast model and red indicating exclusion. *Right*: Actual versus predicted unemployed labour force (ULF) in levels and differences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

72%), where the stated percentages indicate the proportion of nowcast models in which the respective variables are selected. Furthermore, the last bar represents the frequency in which the lagged level of the Dutch unemployment rate is selected, which occurs for 43 out of 54 nowcasts (80%). The frequent selection of the lagged level of unemployment rates in conjunction with the other lagged levels is indicative of the presence of cointegration among unemployment and Google Trends series. However, we do not attach any structural meaning to the found equilibria based on the difficulty of interpretation when one does not assume the presence of weak exogeneity.

In an attempt to gain insights into the temporal stability of our estimator, we visually display the selection stability in the bottom-left part of Fig. 4. Generally, we see that for the early and later period of the sample very few time series enter the model in levels, whereas for the middle part of the sample the majority of variables are selected. The exact reason for these patterns to occur is unknown and raises questions on the stability of Google trends as informative predictors of Dutch unemployment rates. Standard feasible explanations concern structural instability in the DGP, seasonality effects or data idiosyncrasies. However, there are additional peculiarities specific to the use of Google trends such as normalization, data hubris and search algorithm dynamics, all of which might result in unstable performance (cf. Lazer et al., 2014). Since the focus of this application is on the relative performance between our estimator and a common benchmark model, rather than on a structural analysis of the relation between Google Trends and unemployment rates, we leave this issue aside as it is outside the scope of the paper. Instead, we focus on the relative empirical performance of our methods, which, notwithstanding the aforementioned caveats, we deem convincingly favourable for SPECS. Finally, on the right of Fig. 4 we display the realized and predicted unemployment rates in levels and differences. Both the penalized ADL model and SPECS seem to follow the actual unemployment rates with reasonable accuracy, with the largest nowcast errors occurring in the first half of 2014. Prior to this period the unemployment rates had been steadily rising in the aftermath of the economic recession, whereas 2014 marks the start of a recovery period. Given that the models are fit on historical data, it is natural that the estimators overestimate the unemployment rate shortly after the start of the economic recovery. Perhaps not entirely coincidental, the start of the period over which the majority of lagged levels are included by SPECS coincides with this recovery period as well, thereby hinting towards structural instability in the DGP as a plausible cause for the observed selection instability.

7. Conclusion

In this paper, we propose the use of SPECS as an automated approach for sparse single-equation error correction modelling in high-dimensional settings. SPECS is an intuitive estimator that applies penalized regression to a conditional error-correction model. We show that SPECS possesses the oracle property and is able to consistently select the long-run and short-run dynamics in the underlying DGP. These results are derived with the aid of a novel bound on the minimum eigenvalue of the sample covariance matrix containing integrated process, which may be of independent interest. Additionally, in pursuit of suitable weights that aid in the identification of the subset of relevant variables, we derive the consistency of the ridge estimator applied to the same model and demonstrate how ridge regression may be used to construct these weights.

We document favourable finite sample performance of SPECS by means of simulations and an empirical application. The simulation exercise confirms strong selective and predictive capabilities in both low and high dimensions with convincing

gains over a benchmark penalized ADL model that ignores cointegration in the dataset. Furthermore, the simulation results demonstrate that the selective capabilities of SPECS remain adequate absent of weak exogeneity and the nowcasting performance remains superior to the benchmark. Finally, we consider an empirical application in which we nowcast the Dutch unemployment rate with the use of Google Trends series. Across all three different dynamic specifications considered, SPECS attains higher nowcast accuracy, thus confirming the findings from our simulation study. As a result, we believe that our proposed estimator, which is easily implemented with readily available tools at a low computational cost, offers a valuable tool for practitioners by enabling automated model estimation on relatively large and potentially non-stationary datasets and, most importantly, allowing to take into account potential (co)integration without requiring pre-testing procedures.

Finally, we highlight several important sources through which the assumptions and asymptotic framework may be generalized further. Sharper and more direct eigenvalue bounds can be utilized to cast SPECS into an even higher-dimensional setting. Similarly, a suitable compatibility condition can be used to validate the lasso as an initial estimator, resulting in improved weights and, again, a less restrictive asymptotic framework. These topics remain subject to our continuing investigation.

Appendix A. Main proofs

Before presenting the proofs, we start by defining several quantities of interest, some of which are simply repeated for the sake of convenience. First, recall that, under the assumption that $\boldsymbol{z}_0 = \boldsymbol{0}$, the moving average representation of the observed time series is given by

$$\boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{C}(L)\boldsymbol{\epsilon}_t, \qquad \boldsymbol{Z}_{-1} = \boldsymbol{S}_{-1}\boldsymbol{C}' + \boldsymbol{\iota}_T\boldsymbol{\mu}' + \boldsymbol{t}\boldsymbol{\tau}' + \boldsymbol{E}_{-1}\boldsymbol{C}'(L), \tag{26}$$

where $\boldsymbol{S}_{-1} = (\boldsymbol{s}_0, \dots, \boldsymbol{s}_{T-1})'$, $\boldsymbol{C} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I}_N - \sum_{j=1}^p \boldsymbol{\Phi}_j\right)\boldsymbol{B}_\perp\right)^{-1}\boldsymbol{A}_\perp'$, $\boldsymbol{t} = (0, \dots, T-1)'$ and $\boldsymbol{E}_{-1} = (\boldsymbol{\epsilon}_0, \dots, \boldsymbol{\epsilon}_{T-1})'$. Furthermore, by the Beveridge–Nelson decomposition $\boldsymbol{C}(z) = \boldsymbol{C}(1) + (1-z)\boldsymbol{C}^*(z)$, where $\boldsymbol{C}^*(z) = \sum_{l=0}^\infty \boldsymbol{C}_l^*$ with $\boldsymbol{C}_l^* = -\sum_{k=l+1}^\infty \boldsymbol{C}_l$. Assumption 3 implies that,

$$\sum_{l=0}^\infty \left\|\boldsymbol{C}_l^*\right\|_\infty = \sum_{l=0}^\infty \left\|\sum_{k=l+1}^\infty \boldsymbol{C}_l\right\|_\infty \le \sum_{l=0}^\infty \sum_{k=l+1}^\infty \|\boldsymbol{C}_l\|_\infty = \sum_{l=1}^\infty l\,\|\boldsymbol{C}_l\|_\infty < \infty,$$

a property that is used to bound several quantities of interest in the proofs of our theoretical results. Letting $\boldsymbol{M} = \boldsymbol{I}_T - \boldsymbol{D}\left(\boldsymbol{D}'\boldsymbol{D}\right)^{-1}\boldsymbol{D}'$, we define

$$\tilde{\boldsymbol{Z}}_{-1} = \boldsymbol{M}\boldsymbol{Z}_{-1} = \boldsymbol{M}\boldsymbol{S}_{-1}\boldsymbol{C}' + \boldsymbol{M}\boldsymbol{E}_{-1}\boldsymbol{C}'(L) = \tilde{\boldsymbol{S}}_{-1}\boldsymbol{C}' + \tilde{\boldsymbol{E}}_{-1}\boldsymbol{C}'(L),$$

where $\tilde{\boldsymbol{Z}}_{-1} = \left(\tilde{\boldsymbol{z}}_0, \dots, \tilde{\boldsymbol{z}}_{T-1}\right)'$, and $\tilde{\boldsymbol{S}}_{-1}$, $\tilde{\boldsymbol{E}}_{-1}$ admitting a similar decomposition. From this representation, one can derive the stationary processes

$$\boldsymbol{B}'\tilde{\boldsymbol{z}}_t = \boldsymbol{B}'\boldsymbol{C}(L)\tilde{\boldsymbol{\epsilon}}_t = \boldsymbol{C}^\beta(L)\tilde{\boldsymbol{\epsilon}}_t, \qquad \text{and} \qquad \Delta\tilde{\boldsymbol{z}}_t = \boldsymbol{C}\tilde{\boldsymbol{\epsilon}}_t + (1-L)\boldsymbol{C}(L)\tilde{\boldsymbol{\epsilon}}_t = \boldsymbol{C}^\Delta(L)\tilde{\boldsymbol{\epsilon}}_t.$$

Letting $\tilde{\boldsymbol{I}} = (0 \cdot \boldsymbol{\iota}_{N-1}, \boldsymbol{I}_{N-1})$, we get the moving average representation

$$\tilde{\boldsymbol{w}}_t = \begin{bmatrix} \Delta\tilde{\boldsymbol{x}}_t \\ \Delta\tilde{\boldsymbol{z}}_{t-1} \\ \vdots \\ \Delta\tilde{\boldsymbol{z}}_{t-p} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{I}}\boldsymbol{C}^\Delta(L) \\ \boldsymbol{C}^\Delta(L)L \\ \vdots \\ \boldsymbol{C}^\Delta(L)L^p \end{bmatrix} \tilde{\boldsymbol{\epsilon}}_t = \boldsymbol{C}^w(L)\tilde{\boldsymbol{\epsilon}}_t, \qquad \tilde{\boldsymbol{W}} = \tilde{\boldsymbol{E}}C^{w\prime}(L), \tag{27}$$

where $\tilde{\boldsymbol{E}} = (\tilde{\boldsymbol{\epsilon}}_1, \dots, \tilde{\boldsymbol{\epsilon}}_T)'$. An additional useful representation follows from partitioning the data as $\boldsymbol{M}\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2)$, where $\boldsymbol{V}_1 = \left(\tilde{\boldsymbol{Z}}_{-1,S_\delta}, \tilde{\boldsymbol{W}}_{S_\pi}\right)$ contains the relevant variables. In congruence with Section 2.3, the $(|S_\delta| \times r^*)$-dimensional matrix $\boldsymbol{B}_{S_\delta}$ is defined as a basis matrix for the cointegration space of $\boldsymbol{z}_{S_\delta,t}$ and $\boldsymbol{B}_{S_\delta,\perp}$ is an $(|S_\delta| \times |S_\delta| - r^*)$-dimensional matrix for its left null space, i.e. $\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{B}_{S_\delta} = \boldsymbol{0}$. Moreover, without loss of generality, we assume that the columns of $\boldsymbol{B}_{S_\delta,\perp}$ are standardized to have unit L_1-norms. The $\boldsymbol{Q}$-transformation is defined in (10) and the $\boldsymbol{Q}$-transformed data are given by (11). Denote the tth row of $\boldsymbol{V}_1\boldsymbol{Q}'$ by $\boldsymbol{v}_t = \left(\boldsymbol{v}_{1,t}', \boldsymbol{v}_{2,t}'\right)'$, where

$$\begin{aligned}
\boldsymbol{v}_{1,t} &= \begin{bmatrix} \boldsymbol{B}_{S_\delta}'\tilde{\boldsymbol{z}}_{S_\delta,t-1} \\ \tilde{\boldsymbol{w}}_{S_\pi,t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{S_\delta}'\boldsymbol{C}_{S_\delta}(L)L \\ \boldsymbol{C}_{S_\pi}^w(L) \end{bmatrix} \tilde{\boldsymbol{\epsilon}}_t =: \boldsymbol{C}^v(L)\tilde{\boldsymbol{\epsilon}}_t, \\
\boldsymbol{v}_{2,t} &= \boldsymbol{B}_{S_\delta,\perp}'\tilde{\boldsymbol{z}}_{S_\delta,t-1} = \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\tilde{\boldsymbol{s}}_{t-1} + \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}(L)\tilde{\boldsymbol{\epsilon}}_{t-1}.
\end{aligned}$$

Let $s_\pi = |S_\pi| + r^*$ and $s_\delta = |S_\delta| - r^*$ and define the scaling matrix $\boldsymbol{S}_T = \operatorname{diag}\left(\sqrt{T}\boldsymbol{I}_{s_\pi}, \frac{T}{\sqrt{s_\delta}}\boldsymbol{I}_{s_\delta}\right)$. Then, we define the appropriately scaled sample covariance matrix as $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}_T^{-1}\left(\sum_{t=1}^T \boldsymbol{v}_t\boldsymbol{v}_t'\right)\boldsymbol{S}_T^{-1} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{bmatrix}$. Based on these quantities, we proceed to describe a set of lemmas and propositions that are required for the proofs of the main theorems in this paper.

A.1. Preliminary lemmas

In this section, we list a set of preliminary results that are used in the proofs of our main theorems in Appendix A.2. The proofs of all lemmas are delegated to the supplementary Appendix C.1. The first result will simplify the calculations on the stochastic components after regressing out $\boldsymbol{D}$.

Lemma A.1. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ denote arbitrary deterministic matrices of dimensions $(N_A \times d_A)$ and $(N_B \times d_B)$, respectively, with $\|\boldsymbol{A}\|_1 \leq K$ and $\|\boldsymbol{B}\|_1 \leq K$. Define two martingale difference sequences $\{\epsilon_j^w\}_{j=\infty}^{\infty}$ and $\{\epsilon_j^u\}_{j=\infty}^{\infty}$ of dimensions N_A and N_B, respectively, where each sequence satisfies Assumption 1. Any form of dependence between these two sequences is allowed and they may correspond to each other. Next, define a stationary $(T \times N_A)$ matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T)'$ with $\boldsymbol{w}_t = \boldsymbol{C}^w(L)\boldsymbol{\epsilon}_{t-l}^w$, where $\boldsymbol{C}^w(L)$ is an $(N_A \times N_A)$-dimensional matrix lag polynomial satisfying $\sum_{l=0}^{\infty} \left\|\boldsymbol{C}_l^w\right\|_{\infty}$. Similarly, let $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)'$ with $\boldsymbol{u}_t = \boldsymbol{C}^u(L)\boldsymbol{\epsilon}_{t-l}^u$, where $\boldsymbol{C}^u(L)$ is an $(N_B \times N_B)$-dimensional matrix lag polynomial satisfying $\sum_{l=0}^{\infty} \left\|\boldsymbol{C}_l^u\right\|_{\infty}$. Define the $(T \times N_A)$-dimensional partial sum matrix $\boldsymbol{S}_{-1} = (\boldsymbol{0}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{T-1})'$ with $\boldsymbol{s}_t = \sum_{j=1}^{T-1} \boldsymbol{\epsilon}_j^w$. Then, letting $\boldsymbol{P} = \boldsymbol{I}_T - \boldsymbol{M}$,*

(1) $\left\|\boldsymbol{A}'\boldsymbol{S}_{-1}'\boldsymbol{M}\boldsymbol{S}_{-1}\boldsymbol{A}\right\|_2 \leq \left\|\boldsymbol{A}'\boldsymbol{S}_{-1}'\boldsymbol{S}_{-1}\boldsymbol{A}\right\|_2$ *and* $\left\|\boldsymbol{B}'\boldsymbol{U}'\boldsymbol{M}\boldsymbol{U}\boldsymbol{B}\right\|_2 \leq \left\|\boldsymbol{B}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{B}\right\|_2$,

(2) $\left\|\boldsymbol{A}'\boldsymbol{S}_{-1}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{B}\right\|_F = O_p\left(\sqrt{d_A d_B}T\right)$ (3) $\left\|\boldsymbol{A}'\boldsymbol{W}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{B}\right\|_F = O_p\left(\sqrt{d_A d_B}\right)$.

The second result describes a set on which SPECS obtains its selection consistency.

Lemma A.2. *Let $\boldsymbol{\gamma}_{S_\gamma} = (\boldsymbol{\delta}_{S_\delta}', \boldsymbol{\pi}_{S_\pi}')'$ denote the $|S_\gamma|$-dimensional vector containing all non-zero coefficients and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{N+M})'$. Furthermore, define $\boldsymbol{s}_1$ as the subgradient of $\left\|\hat{\boldsymbol{\gamma}}\right\|_1$ and $\boldsymbol{s}_2 = \left(\tilde{\boldsymbol{s}}_2', \boldsymbol{0}'\right)'$, where $\tilde{\boldsymbol{s}}_2$ is the subgradient of $\left\|\hat{\boldsymbol{\delta}}\right\|_2$. Then, $\mathbb{P}\left(sign\left(\hat{\boldsymbol{\gamma}}\right) = sign(\boldsymbol{\gamma})\right) \geq \mathbb{P}(\mathscr{A}_T \cap \mathscr{B}_T)$, where*

$$\mathscr{A}_T = \bigcap_{i=1}^{|S_\gamma|} \left\{ \left| \left[\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right]_i \right| < \left|\gamma_{S_\gamma,i}\right| - \frac{\lambda_I}{2} \left| \left[\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{s}_{1,S_\gamma} \right]_i \right| - \frac{\lambda_G}{2} \left| \left[\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{s}_{2,S_\gamma} \right]_i \right| \right\},$$

$$\mathscr{B}_T = \bigcap_{i=1}^{|S_\gamma^c|} \left\{ \left| \left[\boldsymbol{V}_2'\boldsymbol{M}_V\boldsymbol{\epsilon}_y \right]_i \right| < \frac{\lambda_I}{2} \left[\left(\boldsymbol{\Omega}_2\boldsymbol{\iota} - \left| \boldsymbol{V}_2'\boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{s}_{1,S_\gamma} \right| \right) \right]_i - \frac{\lambda_G}{2} \left| \left[\boldsymbol{V}_2'\boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{s}_{2,S_\gamma} \right]_i \right| \right\},$$

with $\boldsymbol{\Omega}_1 = diag(\boldsymbol{\omega}_{S_\gamma})$, $\boldsymbol{\Omega}_2 = diag(\boldsymbol{\omega}_{S_\gamma^c})$, and $\boldsymbol{M}_V = \boldsymbol{I}_T - \boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{V}_1'$.

Next, we derive bounds on the empirical process that frequently appears throughout the proofs.

Lemma A.3. *Under Assumptions 1–3, the stochastic order of the empirical process is*

$$\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 = O_p\left(s_\delta + \sqrt{s_\pi}\right). \tag{28}$$

Pursuing a minimum eigenvalue bound on $\hat{\boldsymbol{\Sigma}}$, we show that its off-diagonal blocks converge to zero.

Lemma A.4. *Under Assumptions 1–4, it holds that $\left\|\hat{\boldsymbol{\Sigma}}_{12}\right\|_2 \xrightarrow{p} 0$ as $T, s_\delta, s_\pi \to \infty$.*

Combining Assumption 5 with Lemma A.4, we obtain the following immediate result.

Lemma A.5. *Under Assumptions 1–5, there exists a constant $\phi^* > 0$, such that, as $T, s_\delta, s_\pi \to \infty$, $\mathbb{P}\left(\lambda_1\left(\hat{\boldsymbol{\Sigma}}\right) \geq \phi^*\right) \to 1$.*

Finally, Lemmas A.3 and A.5 have natural counterparts based on the full dataset.

Lemma A.6. *Let $\hat{\boldsymbol{\Sigma}}_R$ be as defined in (17). Recall the definitions $N_\delta = N - r$, $M_\pi = M + r$ and assume that $\frac{N_\delta}{T^{1/4}} \to 0$ and $\frac{M_\pi}{\sqrt{T}} \to 0$. Then, under Assumptions 1–3 and 7,*

1. $\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_R\right) \geq \phi_R\right) \to 1$, *as $T, N_\delta, M_\pi \to \infty$, and*
2. $\left\|\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{M}\boldsymbol{\epsilon}_y\right\|_2 = O_p\left(N_\delta + \sqrt{M_\pi}\right)$.

A.2. Proofs of Theorems 1 and 2

In this section we present the proofs of Theorems 1 and 2. The proofs of Theorem 3 and Corollary 1 are delegated to the Supplementary Appendix C.2.

Proof of Theorem 1. Based on Lemma A.2, it suffices to show that $\mathbb{P}(\mathscr{A}_T \cap \mathscr{B}_T) \to 1$ as $T, N \to \infty$ or, equivalently, that $\mathbb{P}(\mathscr{A}_T^c) \to 0$ and $\mathbb{P}(\mathscr{B}_T^c) \to 0$. Thus, we start by deriving that $\mathbb{P}(\mathscr{A}_T^c) \to 0$.

Recall the definitions of $\boldsymbol{S}_T = \text{diag}\left(\sqrt{T}\boldsymbol{I}_{s_\pi}, \frac{T}{\sqrt{s_\delta}}\boldsymbol{I}_{s_\delta}\right)$ and define $\boldsymbol{Q}$ as in (10), with $\|\boldsymbol{Q}\|_\infty \leq 1$ by the normalization on $\boldsymbol{B}_{S_\delta}$ and $\boldsymbol{B}_{S_\delta,\perp}$. Then, for T large enough, we may write the set $\mathscr{A}_T^c$ as

$$\begin{aligned}
\mathscr{A}_T^c &= \bigcup_{i=1}^{|S_\gamma|} \left\{ \left| \left[\boldsymbol{Q}'\boldsymbol{S}_T^{-1} \left(\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1} \right)^{-1} \boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right]_i \right| \right. \\
&\geq |\boldsymbol{\gamma}_{S_\gamma,i}| - \frac{\lambda_I}{2} \left| \left[\boldsymbol{Q}'\boldsymbol{S}_T^{-1} \left(\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1} \right)^{-1} \boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} \right]_i \right| \\
&\quad \left. - \frac{\lambda_G}{2} \left| \left[\boldsymbol{Q}'\boldsymbol{S}_T^{-1} \left(\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1} \right)^{-1} \boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{s}_{2,S_\gamma} \right]_i \right| \right\} \\
&= \bigcup_{i=1}^{|S_\gamma|} \left\{ \left| \left[\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right]_i \right| \right. \\
&\geq |\boldsymbol{\gamma}_{S_\gamma,i}| - \frac{\lambda_I}{2} \left| \left[\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} \right]_i \right| - \frac{\lambda_G}{2} \left| \left[\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{s}_{2,S_\gamma} \right]_i \right| \left. \right\} \\
&\subseteq \left\{ \left\| \boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right\|_2 \right. \\
&\geq \min_{1\leq i\leq |S_\gamma|} |\boldsymbol{\gamma}_{S_\gamma,i}| - \frac{\lambda_I}{2} \left\| \boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} \right\|_2 - \frac{\lambda_G}{2} \left\| \boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{s}_{2,S_\gamma} \right\|_2 \left. \right\}
\end{aligned} \tag{29}$$

We proceed by bounding the three quantities in (29) separately. First, by Assumption 4(1), $\frac{s_\delta}{T} \leq \frac{1}{\sqrt{T}} \Rightarrow \left\|\boldsymbol{S}_T^{-1}\right\|_2 = \frac{1}{\sqrt{T}}$ for large enough T. Moreover, letting $s = (s_\delta + s_\pi)$,

$$\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma}\right\|_2 \leq \left\|\boldsymbol{S}_T^{-1}\right\|_2 \|\boldsymbol{Q}\|_2 \|\boldsymbol{\Omega}_1\|_2 \left\|\boldsymbol{s}_{1,S_\gamma}\right\|_2 \leq \frac{\sqrt{s}}{T^{1/2-\xi}}.$$

Then, by Assumption 5, it holds that

$$\left\| \boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right\|_2 \leq \left\|\boldsymbol{S}_T^{-1}\right\|_2 \|\boldsymbol{Q}\|_2 \left\|\hat{\boldsymbol{\Sigma}}^{-1}\right\|_2 \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \leq \frac{\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2}{\sqrt{T}\phi} \tag{30}$$

on a set with probability converging to one. Furthermore, on the same set,

$$\left\| \boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} \right\|_2 \leq \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\right\|_2 \left\|\hat{\boldsymbol{\Sigma}}^{-1}\right\|_2 \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma}\right\|_2 \leq \frac{\sqrt{s}}{\phi T^{1-\xi}} \tag{31}$$

$$\left\| \boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{s}_{2,S_\gamma} \right\|_2 \leq \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\right\|_2^2 \left\|\hat{\boldsymbol{\Sigma}}^{-1}\right\|_2 \left\|\boldsymbol{s}_{2,S_\gamma}\right\|_2 \leq \frac{1}{\phi T} \tag{32}$$

Based on (30) and (31), we obtain probability bounds for $\mathscr{A}_T^c$ as follows:

$$\begin{aligned}
\mathbb{P}\left(\mathscr{A}_T^c\right) &\leq \mathbb{P}\left(\frac{\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2}{\sqrt{T}\phi} \geq |\gamma_{\min}| - \frac{\lambda_I\sqrt{s}}{2\phi T^{1-\xi}} - \frac{\lambda_G}{2\phi T} \right) + o(1) \\
&= \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \geq \phi\,|\gamma_{\min}|\,\sqrt{T} - \frac{\lambda_I\sqrt{s}}{2T^{1/2-\xi}} - \frac{\lambda_G}{2\phi\sqrt{T}} \right) + o(1).
\end{aligned} \tag{33}$$

Then, to establish that $\mathbb{P}\left(\mathscr{A}_T^c\right) \to 0$, by Lemma A.3 it suffices that $\frac{|\gamma_{\min}|\sqrt{T}}{s_\delta+\sqrt{s_\pi}} \to \infty$, $\frac{|\gamma_{\min}|T^{1-\xi}}{\lambda_I\sqrt{s}} \to \infty$ and $\frac{|\gamma_{\min}|T}{\lambda_G} \to \infty$. The first condition corresponds to part (3) of Assumption 4. For the second and third condition, it follows from part (2) of Assumption 6 that, for sufficiently large T,

$$\frac{|\gamma_{\min}|\,T^{1-\xi}}{\lambda_I\sqrt{s}} \geq \frac{\left(s_\delta+\sqrt{s_\pi}\right)T^{1/2-\xi}}{\lambda_I\sqrt{s}} \to \infty, \quad \text{and} \quad \frac{|\gamma_{\min}|\,T}{\lambda_G} \geq \frac{\left(s_\delta+\sqrt{s_\pi}\right)\sqrt{T}}{\lambda_G} \to \infty,$$

Next, we show that $\mathbb{P}\left(\mathscr{B}_T^c\right) \to 0$. It follows from Lemma A.2 that $\mathscr{B}_T^c = \mathscr{B}_{z,T}^c \cup \mathscr{B}_{w,T}^c$, where

$$\begin{aligned}
\mathscr{B}_{z,T}^c &= \bigcup_{i=1}^{|S_\delta^c|} \left\{ \left| \tilde{\boldsymbol{z}}_{S_\delta^c,i}'\boldsymbol{M}_V\boldsymbol{\epsilon}_y \right| \geq \frac{\lambda_I}{2}\omega_{S_\delta^c,i} - \frac{\lambda_I}{2} \left| \tilde{\boldsymbol{z}}_{S_\delta^c,i}'\boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} \right| - \frac{\lambda_G}{2} \left| \tilde{\boldsymbol{z}}_{S_\delta^c,i}'\boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{s}_{2,S_\gamma} \right| \right\} \\
\mathscr{B}_{w,T}^c &= \bigcup_{i=1}^{|S_\pi^c|} \left\{ \left| \tilde{\boldsymbol{w}}_{S_\pi^c,i}'\boldsymbol{M}_V\boldsymbol{\epsilon}_y \right| \geq \frac{\lambda_I}{2}\omega_{S_\pi^c,i} - \frac{\lambda_I}{2} \left| \tilde{\boldsymbol{w}}_{S_\pi^c,i}'\boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} \right| - \frac{\lambda_G}{2} \left| \tilde{\boldsymbol{w}}_{S_\pi^c,i}'\boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{s}_{2,S_\gamma} \right| \right\}
\end{aligned} \tag{34}$$

and $\tilde{\boldsymbol{z}}_{S_\delta^c,i}$ and $\tilde{\boldsymbol{w}}_{S_\pi^c,i}$ represent the ith columns of $\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}$ and $\tilde{\boldsymbol{W}}_{S_\pi^c}$, respectively. For $\mathscr{B}_{z,T}^c$, note that

$$\mathscr{B}_{z,T}^c \subseteq \left\{ \left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{M}_V \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_I}{2} \omega_{S_\delta^c,\min} - \frac{\lambda_I}{2} \left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{s}_{1,S_\gamma} \right\|_2 - \frac{\lambda_G}{2} \left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{s}_{2,S_\gamma} \right\|_2 \right\}. \tag{35}$$

We proceed by bounding each individual term in (35). First, on a set with probability converging to 1,

$$\begin{aligned} \left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{M}_V \boldsymbol{\epsilon}_y \right\|_2 &\leq \left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{\epsilon}_y \right\|_2 + \left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 \\ &\leq \left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{\epsilon}_y \right\|_2 + \frac{\left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2}{\sqrt{\phi}} \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2, \end{aligned} \tag{36}$$

where the last inequality follows from the fact that

$$\begin{aligned} \left\| \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 &= \left(\boldsymbol{\epsilon}_y' \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right)^{1/2} = \left(\boldsymbol{\epsilon}_y' \boldsymbol{V}_1 \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \left(\boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{V}_1 \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \right)^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right)^{1/2} \\ &= \left\| \left(\boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{V}_1 \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \right)^{-1/2} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 \leq \frac{\left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2}{\sqrt{\phi}} \end{aligned}$$

by Lemma A.5. By the same argument, it follows that

$$\left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{s}_{1,S_\gamma} \right\|_2 \leq \frac{\left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2}{\sqrt{\phi}} \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{s}_{1,S_\gamma} \right\|_2 \leq \frac{\sqrt{s} \left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2}{\sqrt{\phi} T^{1/2-\xi}}, \tag{37}$$

$$\left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{s}_{2,S_\gamma} \right\|_2 \leq \frac{\left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2}{\sqrt{\phi}} \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{s}_{2,S_\gamma} \right\|_2 \leq \frac{\left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2}{\sqrt{\phi} T^{1/2}}. \tag{38}$$

Then, plugging (36)–(38) into (35), we obtain

$$\begin{aligned} \mathbb{P}\left(\mathscr{B}_{z,T}^c \right) \leq\ & \mathbb{P}\left(\left\| \tilde{\boldsymbol{Z}}'_{-1,S_\delta^c} \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_I \omega_{S_\delta^c,\min}}{4} - \frac{\lambda_I \sqrt{s} \left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2}{4\sqrt{\phi} T^{1/2-\xi}} - \frac{\lambda_G \left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2}{4\sqrt{\phi} T^{1/2}} \right) \\ & + \mathbb{P}\left(\left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\sqrt{\phi} \lambda_I \omega_{S_\delta^c,\min}}{4 \left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2} - \frac{\lambda_I \sqrt{s}}{4T^{1/2-\xi}} - \frac{\lambda_G}{4T^{1/2}} \right) + o(1). \end{aligned} \tag{39}$$

We proceed by deriving the stochastic order of the common term $\left\| \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2$. Letting $\boldsymbol{U}_{-1,S_\delta^c}$ denote the matrix containing the columns of $\boldsymbol{E}_{-1} \boldsymbol{C}^{w\prime}(L)$ indexed by S_δ^c, and using that $\|\boldsymbol{M}\|_2 = 1$,

$$\begin{aligned} &\mathbb{P}\left(\left\| T^{-1} N^{-1/2} \tilde{\boldsymbol{Z}}_{-1,S_\delta^c} \right\|_2 \geq K_\epsilon \right) = \mathbb{P}\left(\left\| \boldsymbol{M} \boldsymbol{S}_{-1} \boldsymbol{C}'_{S_\delta^c} + \boldsymbol{M} \boldsymbol{U}_{-1,S_\delta} \right\|_2 \geq K_\epsilon \right) \\ &\leq \mathbb{P}\left(\left\| \boldsymbol{C}_{S_\delta^c} \right\|_2 \left\| T^{-1} N^{-1/2} \boldsymbol{S}_{-1} \right\|_2 \geq \frac{K_\epsilon}{2} \right) + \mathbb{P}\left(\left\| T^{-1} N^{-1/2} \boldsymbol{U}_{-1,S_\delta^c} \right\|_2 \geq \frac{K_\epsilon}{2} \right). \end{aligned}$$

Furthermore, by Markov's inequality and Assumption 1, for $K_\epsilon \geq \sqrt{\frac{4 \left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 K}{\epsilon}}$,

$$\begin{aligned} \mathbb{P}\left(\left\| \boldsymbol{C}_{S_\delta^c} \right\|_2 \left\| T^{-1} N^{-1/2} \boldsymbol{S}_{-1} \right\|_2 \geq \frac{K_\epsilon}{2} \right) &\leq \frac{4 \left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 \sum_{i=1}^N \sum_{t=1}^{T-1} \mathbb{E}\left(s_{i,t} \right)^2}{K_\epsilon^2 T^2 N} \leq \frac{4 \left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 K}{K_\epsilon^2} \leq \epsilon, \\ \mathbb{P}\left(\left\| T^{-1} N^{-1/2} \boldsymbol{U}_{-1,S_\delta^c} \right\|_2 \geq \frac{K_\epsilon}{2} \right) &\leq \frac{4 \sum_{i=1}^{|S_\delta^c|} \sum_{t=1}^{T-1} \mathbb{E}\left(u_{S_\delta^c,i,t} \right)^2}{K_\epsilon^2 T^2 N} \leq \frac{4 \phi_{\max} \sum_{i=1}^{|S_\delta^c|} \sum_{l=0}^{\infty} \left\| \boldsymbol{c}_{S_\delta^c,l,i} \right\|_2^2}{K_\epsilon^2 T N} \\ &\leq \frac{4 \phi_{\max} \sum_{l=0}^{\infty} \left\| \boldsymbol{C}_{S_\delta^c,l} \right\|_2^2}{K_\epsilon^2 T} \to 0. \end{aligned}$$

Hence, for all $\epsilon > 0$ there exist $K_\epsilon, T^*, N^* > 0$ such that $\mathbb{P}\left(\left\|\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}\right\|_2 \geq T\sqrt{N}K_\epsilon\right) \leq \epsilon$ for all $T > T^*$ and $N > N^*$. Then, for sufficiently large T, N, the first RHS term of (39) is bounded by

$$\begin{aligned}
&\mathbb{P}\left(\left\|\tilde{\boldsymbol{Z}}'_{-1,S_\delta^c}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I \omega_{S_\delta^c,\min}}{4} - \frac{\lambda_I\sqrt{s}\left\|\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}\right\|_2}{4\sqrt{\phi}T^{1/2-\xi}} - \frac{\lambda_G\left\|\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}\right\|_2}{4\sqrt{\phi}T^{1/2}}\right) \\
&\leq \mathbb{P}\left(\left\|\boldsymbol{C}_{S_\delta^c}\boldsymbol{S}'_{-1}\boldsymbol{M}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I \omega_{S_\delta^c,\min}}{8} - \frac{\lambda_I K_\epsilon\sqrt{s}T^{1/2+\xi}\sqrt{N}}{8\sqrt{\phi}} - \frac{\lambda_G K_\epsilon\sqrt{TN}}{8\sqrt{\phi}}\right) \\
&\quad + \mathbb{P}\left(\left\|\boldsymbol{U}'_{-1,S_\delta^c}\boldsymbol{M}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I \omega_{S_\delta^c,\min}}{8} - \frac{\lambda_I K_\epsilon\sqrt{s}T^{1/2+\xi}\sqrt{N}}{8\sqrt{\phi}} - \frac{\lambda_G K_\epsilon\sqrt{TN}}{8\sqrt{\phi}}\right) + \epsilon.
\end{aligned} \tag{40}$$

As $\{s_{i,t-1}\epsilon_{y,t}\}$ is a m.d.s., it follows from Burkholder's inequality and the C_r-inequality that for $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\left\|\boldsymbol{C}_{S_\delta^c}\boldsymbol{S}'_{-1}\boldsymbol{\epsilon}_y\right\|_2}{T\sqrt{N}} \geq K_\epsilon\right) \leq \frac{\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2^2 \sum_{i=1}^N \mathbb{E}\left(\sum_{t=2}^T s_{i,t-1}\epsilon_{y,t}\right)^2}{K_\epsilon^2 T^2 N} \\
&\leq \frac{K\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2^2 \sigma_y^2 \sum_{i=1}^N \sum_{t=1}^{T-1}\mathbb{E}(s_{i,t})^2}{K_\epsilon^2 T^2 N} \leq \frac{K^*\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2^2 \sigma_y^2}{K_\epsilon^2} \leq \epsilon, \\
&\mathbb{P}\left(\frac{\left\|\boldsymbol{U}'_{-1,S_\delta^c}\boldsymbol{\epsilon}_y\right\|_2}{T\sqrt{N}} \geq K_\epsilon\right) \leq \frac{\sum_{i=1}^{|S_\delta^c|}\mathbb{E}\left(\sum_{t=2}^T\sum_{l=0}^\infty \boldsymbol{c}'_{S_\delta,l,i}\boldsymbol{\epsilon}_{t-1-l}\epsilon_{y,t}\right)^2}{K_\epsilon^2 T^2 N} \\
&\leq \frac{K\sigma_y^2\sum_{i=1}^{|S_\delta^c|}\sum_{t=2}^T\sum_{l=0}^\infty \mathbb{E}\left(\boldsymbol{c}'_{S_\delta,l,i}\boldsymbol{\epsilon}_{t-1-l}\right)^2}{K_\epsilon^2 T^2 N} \leq \frac{K\sigma_y^2\phi_{\max}\sum_{l=0}^\infty\left\|\boldsymbol{C}_{S_\delta,l}\right\|_2^2}{K_\epsilon^2 TN} \to 0,
\end{aligned} \tag{41}$$

for $K_\epsilon \geq \sqrt{\frac{K^*\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2^2\sigma_y^2}{\epsilon}}$. Then, part (2)–(3) of Lemma A.1, it follows that $\left\|\tilde{\boldsymbol{Z}}'_{-1,S_\delta^c}\boldsymbol{\epsilon}_y\right\|_2 = O_p(\sqrt{N}T)$. As $\omega^{-1}_{S_\delta^c,\min} = o_p\left(\frac{\lambda_I}{T\sqrt{N}}\right)$, $\omega^{-1}_{S_\delta^c,\min} = o_p\left(\frac{T^\xi}{\sqrt{sTN}}\right)$, and $\omega^{-1}_{S_\delta^c,\min} = o_p\left(\frac{\lambda_I}{\lambda_G\sqrt{TN}}\right)$ by Assumption 6, we have that

$$\mathbb{P}\left(\left\|\tilde{\boldsymbol{Z}}'_{-1,S_\delta^c}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I\omega_{S_\delta^c,\min}}{4} - \frac{\lambda_I\sqrt{s}\left\|\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}\right\|_2}{4\sqrt{\phi}T^{1/2-\xi}} - \frac{\lambda_G\left\|\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}\right\|_2}{4\sqrt{\phi}T^{1/2}}\right) \to 0.$$

Next, we focus on the second RHS term of (39). First, again using that $\left\|\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}\right\|_2 = O_p(T\sqrt{N})$,

$$\begin{aligned}
&\mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}'_1\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\sqrt{\phi}\lambda_I\omega_{S_\delta^c,\min}}{4\left\|\tilde{\boldsymbol{Z}}_{-1,S_\delta^c}\right\|_2} - \frac{\lambda_I\sqrt{s}}{4T^{1/2-\xi}} - \frac{\lambda_G}{4T^{1/2}}\right) \\
&\leq \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}'_1\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\sqrt{\phi}\lambda_I\omega_{S_\delta^c,\min}}{4K_\epsilon T\sqrt{N}} - \frac{\lambda_I\sqrt{s}}{4T^{1/2-\xi}} - \frac{\lambda_G}{4T^{1/2}}\right) + \epsilon.
\end{aligned} \tag{42}$$

Then, based on Lemma A.3, for the RHS of (42) to converge to zero, it is sufficient that

$$\omega^{-1}_{S_\delta^c,\min} = o_p\left(\frac{\lambda_I}{(s_\delta + \sqrt{s_\pi})T\sqrt{N}}\right), \quad \omega^{-1}_{S_\delta^c,\min} = o_p\left(\frac{1}{\sqrt{s}T^{1/2+\xi}\sqrt{N}}\right) \quad \text{and} \quad \omega^{-1}_{S_\delta^c,\min} = o_p\left(\frac{\lambda_I}{\lambda_G\sqrt{TN}}\right).$$

All three conditions are satisfied under Assumption 6. Consequently, both RHS terms of (39) converge to zero, thereby concluding that $\mathbb{P}\left(\mathscr{B}^c_{z,T}\right) \to 0$.

It remains to prove that $\mathbb{P}\left(\mathscr{B}^c_{w,T}\right) \to 0$, where $\mathscr{B}^c_{w,T}$ is defined in (34). First, note that

$$\begin{aligned}
\mathscr{B}^c_{w,T} \subseteq \Bigg\{&\left\|\tilde{\boldsymbol{W}}'_{S_\pi^c}\boldsymbol{M}_V\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I}{2}\omega_{S_\pi^c,\min} - \frac{\lambda_I}{2}\left\|\tilde{\boldsymbol{W}}'_{S_\pi^c}\boldsymbol{V}_1\left(\boldsymbol{V}'_1\boldsymbol{V}_1\right)^{-1}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma}\right\|_2 \\
&- \frac{\lambda_G}{2}\left\|\tilde{\boldsymbol{W}}'_{S_\pi^c}\boldsymbol{V}_1\left(\boldsymbol{V}'_1\boldsymbol{V}_1\right)^{-1}\boldsymbol{s}_{2,S_\gamma}\right\|_2\Bigg\}.
\end{aligned}$$

Furthermore, on a set with probability converging to one,

$$\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}'\boldsymbol{M}_V\boldsymbol{\epsilon}_y\right\|_2 \leq \left\|\tilde{\boldsymbol{W}}_{S_\pi^c}'\boldsymbol{\epsilon}_y\right\|_2 + \frac{\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2 \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2}{\sqrt{\phi}}, \tag{43}$$

$$\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}'\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma}\right\|_2 \leq \frac{\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2 \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma}\right\|_2}{\sqrt{\phi}} \leq \frac{\sqrt{s}\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2}{\sqrt{\phi}T^{1/2-\xi}} \tag{44}$$

$$\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}'\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{s}_{2,S_\gamma}\right\|_2 \leq \frac{\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2 \left\|\boldsymbol{S}_T^{-1}\boldsymbol{s}_{2,S_\gamma}\right\|_2}{\sqrt{\phi}} \leq \frac{\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2}{\sqrt{\phi}T^{1/2}}. \tag{45}$$

Then, plugging (43)–(45) into $\mathscr{B}_{w,T}^c$ from (34), we obtain

$$\begin{aligned}\mathbb{P}\left(\mathscr{B}_{w,T}^c\right) &\leq \mathbb{P}\left(\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}'\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I\omega_{S_\pi^c,\min}}{4} - \frac{\lambda_I\sqrt{s}\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2}{4\sqrt{\phi}T^{1/2-\xi}} - \frac{\lambda_G\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2}{4\sqrt{\phi}T^{1/2}}\right) \\ &\quad + \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I\sqrt{\phi}\omega_{S_\pi^c,\min}}{4\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2} - \frac{\lambda_I\sqrt{s}}{4T^{1/2-\xi}} - \frac{\lambda_G}{4T^{1/2}}\right) + o(1) \\ &= \mathbb{P}\left(\mathscr{B}_{w_1,T}^c\right) + \mathbb{P}\left(\mathscr{B}_{w_2,T}^c\right) + o(1).\end{aligned} \tag{46}$$

Next, we derive the order of $\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2$. From definition (27), and using that $\|\boldsymbol{M}\|_2 = 1$, it follows that

$$\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2 = \left\|\boldsymbol{M}\boldsymbol{E}\boldsymbol{C}_{S_\pi^c}'(L)\right\|_2 \leq \left\|\boldsymbol{E}\boldsymbol{C}_{S_\pi^c}'(L)\right\|_2 = \left\|\boldsymbol{W}_{S_\pi^c}\right\|_2.$$

Recalling that $w_{i,t} = \sum_{l=0}^\infty \boldsymbol{c}_{l,i}^{w\prime}\boldsymbol{\epsilon}_{t-l}$, it holds that

$$\mathbb{E}\left(w_{i,t}\right)^2 = \sum_{l=0}^\infty \boldsymbol{c}_{l,i}^{w\prime}\boldsymbol{\Sigma}_\epsilon\boldsymbol{c}_{l,i}^w \leq \phi_{\max}\sum_{l=0}^\infty \left\|\boldsymbol{c}_{l,i}^w\right\|_2^2 \leq \phi_{\max}\sum_{l=0}^\infty \left\|\boldsymbol{C}_l^w\right\|_2^2,$$

by Assumption 3. Then, for any $\epsilon > 0$, it follows that, for $K_\epsilon \geq \left(\phi_{\max}\sum_{l=0}^\infty \left\|\boldsymbol{C}_l^w\right\|_2^2\right)^{-1/2}$,

$$\mathbb{P}\left(\frac{\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}\right\|_2}{\sqrt{TM}} \geq K_\epsilon\right) \leq \frac{\sum_{i=1}^M\sum_{t=1}^T\mathbb{E}\left(w_{i,t}\right)^2}{K_\epsilon^2 TM} \leq \frac{\phi_{\max}\sum_{l=0}^\infty\left\|\boldsymbol{C}_l^w\right\|_2^2}{K_\epsilon^2} \leq \epsilon. \tag{47}$$

Furthermore, it is straightforward to verify that $\{w_{i,t}\epsilon_{y,t}\}$ is a martingale difference sequence. Thus, by the Markov bound and Burkholder's inequality,

$$\begin{aligned}\mathbb{P}\left(\left\|\boldsymbol{W}_{S_\pi^c}'\boldsymbol{\epsilon}_y\right\|_2 \geq K_\epsilon\sqrt{TM}\right) &\leq \frac{\sum_{i=1}^M\mathbb{E}\left(\sum_{t=1}^T w_{i,t}\epsilon_{y,t}\right)^2}{K_\epsilon^2 TM} \leq \frac{K\sum_{i=1}^M\sum_{t=1}^T\mathbb{E}\left(w_{i,t}\epsilon_{y,t}\right)^2}{K_\epsilon^2 TM} \\ &\leq \frac{K\sum_{i=1}^M\sum_{t=1}^T\sum_{l_1,l_2=0}^\infty\sum_{j_1,j_2=1}^M\left|c_{l_1,i,j_1}^w\right|\left|c_{l_2,i,j_2}^w\right|\mathbb{E}\left|\epsilon_{j_1,t-1_1}\epsilon_{j_2,t-1_2}\epsilon_{y,t}^2\right|}{K_\epsilon^2 TM} \\ &\leq \frac{K^*\sum_{i=1}^M\left(\sum_{l=0}^\infty\left\|\boldsymbol{c}_{l,i}^w\right\|_1\right)^2}{K_\epsilon^2 M} \leq \frac{K^*\left(\sum_{l=0}^\infty\left\|\boldsymbol{C}_l^w\right\|_\infty\right)^2}{K_\epsilon} \leq \epsilon,\end{aligned} \tag{48}$$

for $K_\epsilon \geq \left(\frac{K^*\left(\sum_{l=0}^\infty\|\boldsymbol{c}_l^w\|_\infty\right)^2}{\epsilon}\right)^{1/2}$. Then, part (3) of Lemma A.1 shows that $\left\|\boldsymbol{W}_{S_\pi^c}'\boldsymbol{M}\boldsymbol{\epsilon}_y\right\|_2 = O_p(\sqrt{TM})$. Using (47) to further simplify (46),

$$\mathbb{P}\left(\left\|\tilde{\boldsymbol{W}}_{S_\pi^c}'\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I\omega_{S_\pi^c,\min}}{4} - \frac{\lambda_I K_\epsilon\sqrt{s}T^\xi\sqrt{M}}{4\sqrt{\phi}} - \frac{\lambda_G K_\epsilon\sqrt{M}}{4\sqrt{\phi}}\right) + \epsilon, \tag{49}$$

such that (48) implies $\mathbb{P}\left(\mathscr{B}^c_{w_1,T}\right) \to 0$, if $\omega^{-1}_{S^c_\pi,\min} = o_p\left(\frac{\lambda_I}{\sqrt{TM}}\right)$, $\omega^{-1}_{S^c_\pi,\min} = o_p\left(\frac{1}{\sqrt{s}T^\xi\sqrt{M}}\right)$ and $\omega^{-1}_{S^c_\pi,\min} = o_p\left(\frac{\lambda_I\omega_{S^c_\pi,\min}}{\lambda_G\sqrt{M}}\right)$. as ensured by Assumption 6. Similarly, for $\mathscr{B}^c_{w_2,T}$ as defined in (46) we get

$$\mathbb{P}\left(\mathscr{B}^c_{w_2,T}\right) \leq \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_I\sqrt{\phi}\omega_{S^c_\pi,\min}}{4K_\epsilon\sqrt{TM}} - \frac{\lambda_I\sqrt{s}}{4T^{1/2-\xi}} - \frac{\lambda_G}{4T^{1/2}}\right) + \epsilon, \tag{50}$$

such that, by Lemma A.3, sufficient conditions for $\mathbb{P}\left(\mathscr{B}^c_{w_2,T}\right) \to 0$ are given by

$$\omega^{-1}_{S^c_\pi,\min} = o_p\left(\frac{\lambda_I}{(s_\delta+\sqrt{s_\pi})\sqrt{TM}}\right), \quad \omega^{-1}_{S^c_\pi,\min} = o_p\left(\frac{1}{\sqrt{s}T^\xi\sqrt{M}}\right) \quad \text{and} \quad \omega^{-1}_{S^c_\pi,\min} = \left(\frac{\lambda_I}{\lambda_G\sqrt{M}}\right).$$

All three conditions are satisfied under Assumption 6. Hence, we may conclude that $\mathbb{P}\left(\mathscr{B}^c_{w,T}\right) \to 0$. □

Proof of Theorem 2. First, we recall the definitions $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2)$, $\boldsymbol{V}_1 = \left(\tilde{\boldsymbol{Z}}_{-1,S_\delta}, \tilde{\boldsymbol{W}}_{S_\pi}\right)$, $\boldsymbol{\Omega} = \operatorname{diag}(\boldsymbol{\omega})$ and $\boldsymbol{\Omega}_1 = \operatorname{diag}\left(\boldsymbol{\omega}_{S_\gamma}\right)$. Based on the first order conditions, it follows from (5) that

$$\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} = \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y - \frac{1}{2}\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\left(\lambda_I\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} + \lambda_G\boldsymbol{s}_{2,S_\gamma}\right), \tag{51}$$

on a set with probability converging to one based on Theorem 1. By pre-multiplying (51) by $\boldsymbol{S}_T\boldsymbol{Q}'^{-1}$ and taking the Euclidean norm on both sides, it follows that

$$\begin{aligned}\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 &\leq \left\|\left(\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\right)^{-1}\right\|_2 \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\left(\boldsymbol{V}_1'\boldsymbol{\epsilon}_y - \frac{\lambda_I}{2}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma} - \frac{\lambda_G}{2}\boldsymbol{s}_{2,S_\gamma}\right)\right\|_2 \\ &\leq \phi^{-1}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 + \frac{\lambda_I}{2}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma}\right\|_2 + \frac{\lambda_G}{2}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{s}_{2,S_\gamma}\right\|_2\right) + o_p(1),\end{aligned} \tag{52}$$

by Lemma A.5. We derive the stochastic order for the three RHS terms of (52). First, $\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 = O_p\left(s_\delta + \sqrt{s_\pi}\right)$, by Lemma A.3. By Assumption 6, on a set with probability converging to one, the second term and third term on the RHS of (52) are bounded by

$$\frac{\lambda_I}{2}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}_{1,S_\gamma}\right\|_2 \leq \frac{\lambda_I}{2}\left\|\boldsymbol{S}_T^{-1}\right\|_2\|\boldsymbol{Q}\|_2\|\boldsymbol{\Omega}_1\|_2\left\|\boldsymbol{s}_{1,S_\gamma}\right\|_2 \leq \frac{\lambda_I\sqrt{s}}{2T^{1/2-\xi}} = o\left(s_\delta + \sqrt{s_\pi}\right), \tag{53}$$

$$\frac{\lambda_G}{2}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{s}_{2,S_\gamma}\right\|_2 \leq \frac{\lambda_G}{2}\left\|\boldsymbol{S}_T^{-1}\right\|_2\|\boldsymbol{Q}\|_2\left\|\boldsymbol{s}_{2,S_\gamma}\right\|_2 \leq \frac{\lambda_G}{\sqrt{T}} \to 0. \tag{54}$$

Hence, plugging these result into (52), we conclude that, as required,

$$\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 = O_p\left(s_\delta + \sqrt{s_\pi}\right). \quad \square$$

Appendix B. Bounds on minimum eigenvalues

In this Appendix, we provide sufficient conditions for Assumption 5. We first present some preliminary results in Appendix B.1 and main eigenvalue bounds in Appendix B.2. The proofs of these lemmas and theorems are delegated to the Supplementary Appendices C.1 and C.3, respectively.

B.1. Preliminary results

We first present a general result linking the eigenvalues of two matrices together.

Lemma B.1. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ denote two s-dimensional square non-negative definite matrices. Then,*

(1) for all $i = 1, \ldots, s$, it holds that $|\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{B})| \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2$,
(2) if $\|\boldsymbol{A} - \boldsymbol{B}\|_{\max} \leq \delta$, then $\lambda_{\min}(\boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A}) - s\delta$.

The following result demonstrates the issue of collinearity of integrated variables in high dimensions.

Lemma B.2. *Define an s-dimensional white noise sequence $\boldsymbol{u}_t \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_s)$ and let $\boldsymbol{h}_t = \sum_{j=1}^t \boldsymbol{u}_j$. Then, as $s, T \to \infty$, for any $\phi > 0$,*

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^T \boldsymbol{h}_t\boldsymbol{h}_t'\right) > \phi\right) \to 0. \tag{55}$$

B.2. Main results

We first give a bound on the minimum eigenvalue of the covariance matrix of the stationary variables. This follows standard arguments in the literature, but is given for completeness.

Theorem B.1. *Define* $\boldsymbol{\Sigma}_{11} = \mathbb{E}\left(\boldsymbol{v}_{1,t}\boldsymbol{v}_{1,t}'\right)$ *and assume that* $\lambda_{\min}(\boldsymbol{\Sigma}_{11}) \geq 2\phi$ *for some* $\phi > 0$*. Then, under Assumptions 1–3 and 4(2), as* $T, s_\delta, s_\pi \to \infty$ *we have that* $\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) \geq \phi\right) \to 1$.

Contrary to $\hat{\boldsymbol{\Sigma}}_{11}$, the matrix $\hat{\boldsymbol{\Sigma}}_{22} = \frac{s_\delta}{T^2}\boldsymbol{B}_{S_\delta,\perp}'\left(\sum_{t=1}^{T}\tilde{\boldsymbol{z}}_{S_\delta,t}\tilde{\boldsymbol{z}}_{S_\delta,t}'\right)\boldsymbol{B}_{S_\delta,\perp}$ does not converge in probability to a deterministic matrix. Accordingly we aim to bound $\hat{\boldsymbol{\Sigma}}_{22}$ directly, under varying additional assumptions on the DGP and the growth rate of s_δ.

Theorem B.2. *Let* $\hat{\boldsymbol{\Sigma}}_{22}$ *be as defined in Assumption 5 and assume that* $\boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$*. Then, under Assumptions 1–3, there exists a constant* $\zeta > 0$ *such that, as* $s_\delta, T \to \infty$ *with* $\frac{s_\delta}{T^{1/2}} \to 0$*, we have that* $\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) \geq \zeta\right) \to 1$.

It is possible to extend Theorem B.2 to general distributions, based on an argument that relies on strong Gaussian approximations, at the additional cost of a further restriction on the growth rate of s_δ.

Theorem B.3. *Let* $\hat{\boldsymbol{\Sigma}}_{22}$ *be as defined in Assumption 5 and set* $\boldsymbol{M} = \boldsymbol{I}_T$ *assuming that* $\boldsymbol{\mu} = \boldsymbol{\tau} = \boldsymbol{0}$*. Assume that* $\boldsymbol{\epsilon}_t = \boldsymbol{D}\boldsymbol{\epsilon}_{u,t}$*, where* $\boldsymbol{D}$ *is a* $T \times T$*-matrix with* $\|\boldsymbol{D}\| \leq K < \infty$*, and* $\epsilon_{u,s,i} \perp\!\!\!\perp \epsilon_{u,t,j}$ *for all* i, j, s, t *with* $i \neq j$*. Let* $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})_{i,j=1}^N$ *and assume that* $\max_{1\leq i\leq N}\mathbb{E}\left|\sum_{t=1}^{T}\left(\epsilon_{u,t,i}^2 - \sigma_{u,ii}^2\right)\right|^2 = O\left(T^{1/2}\right)$*. Then, under Assumptions 1–3, a constant* $\zeta > 0$ *exists, independent of* s_δ*,* N *and* T*, such that, as* $s_\delta, N, T \to \infty$ *with* $\frac{s_\delta N}{T^{1/4}} \to 0$*,* $\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) > \zeta\right) \to 1$.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.07.021. This supplementary material provides additional results which contain the following: Appendix C.1 provides proofs of all lemmas that were presented in Appendix A and of the main paper. Appendix C.2 provides the proofs of Corollary 1 and Theorem 3 and Appendix C.3 contains the proofs of the main theorems in of the main paper. Finally, Appendix C.4 provides a detailed data description of the empirical application considered in Section 6.

References

Banerjee, A., Dolado, J., Mestre, R., 1998. Error-correction mechanism tests for cointegration in a single-equation framework. J. Time Series Anal. 19 (3), 267–283.

Banerjee, A., Marcellino, M., Masten, I., 2014. Forecasting with factor-augmented error correction models. Int. J. Forecast. 30 (3), 589–612.

Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19 (2), 521–547.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., 2013. Valid post-selection inference. Ann. Statist. 41, 802–837.

Boswijk, H., 1994. Testing for an unstable root in conditional and structural error correction models. J. Econometrics 63, 37–60.

Breheny, P., Huang, J., 2009. Penalized methods for bi-level variable selection. Stat. Interface 2 (3), 369.

Bühlmann, P., Van De Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Science & Business Media.

Chang, Y., 2004. Bootstrap unit root tests in panels with cross-sectional dependency. J. Econometrics 120 (2), 263–293.

Chernozhukov, V., Härdle, W.K., Huang, C., Wang, W., 2018. LASSO-driven inference in time and space. arXiv e-print 1806.05081, arXiv.

Chetverikov, D., Liao, Z., Chernozhukov, V., 2016. On cross-validated lasso. arXiv e-print 1605.02214, arXiv.

Choi, H., Varian, H., 2012. Predicting the present with Google Trends. Econ. Rec. 88 (s1), 2–9.

Chou, W., Denis, K.F., Lee, C.F., 1996. Hedging with the nikkei index futures: The convential model versus the error correction model. Q. Rev. Econ. Finance 36 (4), 495–505.

Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation and testing. Econometrica 55, 251–276.

Engle, R.F., Yoo, B.S., 1987. Forecasting and testing in co-integrated systems. J. Econometrics 35 (1), 143–159.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. J. Monetary Econ. 55 (4), 665–676.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning. Springer.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: Principles and Practice. OTexts.

Johansen, S., 1992. Cointegration in partial systems and the efficiency of single-equation analysis. J. Econometrics 52 (3), 389–402.

Johansen, S., 1995. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press.

Kock, A.B., 2016. Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. Econometric Theory 32, 243–259.

Kock, A.B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. J. Econometrics 186, 325–344.

Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. The parable of Google Flu: traps in big data analysis. Science 343, 1203–1205.

Lee, J.H., Shi, Z., Gao, Z., 2018. On LASSO for predictive regression. arXiv e-prints 1810.03140, arXiv.

Lee, J.D., Sun, D.L., Sun, Y., Taylor, J.E., 2016. Exact post-selection inference, with application to the lasso. Ann. Statist. 44, 907–927.

Leeb, H., Pötscher, B.M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21 (1), 21–59.

Liang, C., Schienle, M., 2019. Determination of vector error correction models in high dimensions. J. Econometrics 208 (2), 418–441.

Liao, Z., Phillips, P.C.B., 2015. Automated estimation of vector error correction models. Econometric Theory 31, 581–646.

Masini, R.P., Medeiros, M.C., Mendes, E.F., 2019. Regularized estimation of high-dimensional vector autoregressions with weakly dependent innovations. arXiv e-print 1912.09002, arXiv.
McCracken, M.W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. J. Bus. Econom. Statist. 34, 574–589.
Medeiros, M.C., Mendes, E.F., 2016. ℓ_1-Regularization of high-dimensional time series models with non-gaussian and heteroskedastic errors. J. Econometrics 191, 255–271.
Onatski, A., Wang, C., 2019. Extreme canonical correlations and high-dimensional cointegration analysis. J. Econometrics 212 (1), 307–322.
Palm, F.C., Smeekes, S., Urbain, J.-P., 2010. A sieve bootstrap test for cointegration in a conditional error correction model. Econometric Theory 26 (3), 647–681.
Palm, F.C., Smeekes, S., Urbain, J.-P., 2011. Cross-sectional dependence robust block bootstrap panel unit root tests. J. Econometrics 163, 85–104.
Phillips, P.C., Hansen, B.E., 1990. Statistical inference in instrumental variables regression with I(1) processes. Rev. Econom. Stud. 57, 99–125.
Phillips, P.C., Ouliaris, S., 1990. Asymptotic properties of residual based tests for cointegration. Econometrica 58, 165–193.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.
Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group lasso. J. Comput. Graph. Statist. 22 (2), 231–245.
Smeekes, S., Wijler, E., 2018b. Macroeconomic forecasting using penalized regression methods. Int. J. Forecast. 34 (3), 408–430.
Smeekes, S., Wijler, E., 2020. Unit roots and cointegration. In: Fuleky, P. (Ed.), Macroeconomic Forecasting in the Era of Big Data. In: Advanced Studies in Theoretical and Applied Econometrics, vol. 52, Springer, pp. 541–584, chapter 17.
Vaiter, S., Deledalle, C., Peyré, G., Fadili, J., Dossal, C., 2012. The degrees of freedom of the group lasso for a general design. arXiv e-print 1212.6478, arXiv.
Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42, 1166–1202.
Wilms, I., Basu, S., Bien, J., Matteson, D.S., 2017. Sparse identification and estimation of high-dimensional vector autoregressive moving averages. arXiv e-print 1707.09208, arXiv.
Wilms, I., Croux, C., 2016. Forecasting using sparse cointegration. Int. J. Forecast. 32, 1256–1267.
Yamada, H., 2017. The frisch–waugh–lovell theorem for the lasso and the ridge regression. Comm. Statist. Theory Methods 46 (21), 10897–10902.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. Roy. Statist. Soc. Ser. B 68 (1), 49–67.
Zhang, R., Robinson, P., Yao, Q., 2019. Identifying cointegration by eigenanalysis. J. Amer. Statist. Assoc. 114 (526), 916–927.
Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7, 2541–2563.
Zou, H., 2006. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.
Zou, H., Hastie, T., Tibshirani, R., 2007. On the "degrees of freedom" of the lasso. Ann. Statist. 35 (5), 2173–2192.