An Automated Approach Towards Sparse Single-Equation Cointegration Modelling

An Automated Approach Towards Sparse Single-Equation Cointegration Modelling [*]

Stephan Smeekes Etienne Wijler

Maastricht University
Department of Quantitative Economics

September 25, 2018

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. We show that SPECS is able to consistently estimate an appropriate linear combination of the cointegrating vectors that may occur in the underlying DGP, while simultaneously enabling the correct recovery of sparsity patterns in the corresponding parameter space. 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.

Keywords: SPECS, Penalized Regression, Single-Equation Error-Correction Model, Cointegration, High-Dimensional Data.

JEL-Codes: C32, C52, C55

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 drawback when the dimension of the dataset increases.

While the VECM approach allows for a general and flexible modelling of potentially cointegrated series, and the optimality properties of a correctly specified full-system estimator are

[*] The first author thanks NWO for financial support. Previous versions of this paper have been presented at CFE-CM Statistics 2017 and the NESG 2018. We gratefully acknowledge the comments by participants at these conferences. In addition, we thank Robert Adámek, Alain Hecq, Luca Margaritella, Hanno Reuvers, Sean Telg and Ines Wilms for their valuable feedback, and Caterina Schiavoni for help with the data collection. All remaining errors are our own. Corresponding author: Etienne Wijler. Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: e.wijler@maastrichtuniversity.nl

theoretically attractive, 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 full-system estimators requires non-trivial a priori choices on the relevance of specific variables to keep the dimension manageable. Moreover, in many cases of practical relevance, 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. Additionally, this approach ignores the presence of cointegration among the variables, which may have detrimental effects on the subsequent analysis. In an attempt to resolve these issues, we propose SPECS as an alternative approach towards intuitive automated modelling of large non-stationary datasets.

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, thereby facilitating subsequent model interpretation without the need to a priori dismiss variables from the dataset. 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, ready-to-use R code is available online that implements an intuitive and easy-to-interpret algorithm for SPECS estimation.[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 are known to require a weak exogeneity assumption for asymptotically efficient inference (Johansen, 1992), where the assumption of 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, when the goal is to model a single variable of interest using a large set of potential regressors, SPECS provides a variable selection mechanism, allowing the researcher to discard the variables

[1] https://sites.google.com/view/etiennewijler.

(in levels and/or first differences) that are irrelevant for this particular analysis. Simulation results presented in this paper demonstrate strong selective capabilities in both low and high dimensions. Furthermore, a simulated nowcasting application highlights the importance of incorporating cointegration in the data; 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, 2018, for a recent overview). Particularly relevant to this paper are several recently developed methods that allow for sparse estimation of large-dimensional VECMs. Wilms and Croux (2016) propose to estimate the full system by penalized maximum likelihood, where shrinkage is performed on the cointegrating vector, the parameters regulating the short-run dynamics and the covariance matrix. Alternatively, Liang and Schienle (2015) develop an algorithm where parameter estimation and rank determination are performed jointly by employing a clever penalty that makes use of the well-known QR-decomposition of the long-run coefficient matrix. In addition, these authors show the resulting estimator to possess oracle-like properties. Liao and Phillips (2015) also provide an automated method of joint rank selection and parameter estimation with the use of an adaptive penalty where the penalty weights make use of the different convergence rates of the eigenvalues of the estimated long-run coefficient matrix. They derive comparable theoretical results, i.e. the oracle property, with an added extension to DGPs with weakly dependent innovations. While these novel contributions offer elegant solutions to the infamous curse of dimensionality, their implementation is non-standard and rather technical. Furthermore, while a correctly specified full system analysis is asymptotically efficient without necessitating the assumption of weak exogeneity, whether the potential loss of asymptotic efficiency in a more parsimonious single-equation model translates to inferior performance in finite samples ultimately remains an empirical question.

The paper is structured as follows. In Section 2 we discuss the data generating process and describe the SPECS estimator. The main theoretical results of the paper are presented in Section 3. Section 4 contains several simulation studies, followed by an empirical application in Section 5. We conclude in Section 6. All proofs of the major results are contained in Appendix A, while the supplementary Appendix B contains additional results.

Finally, a word on notation. We use $\|\cdot\|_p$ to denote the ℓ_p-norm, i.e. $\|v\|_p = \left(\sum_{i=1}^n |v|_i^p\right)^{1/p}$ for a vector $v \in \mathbb{R}^n$ or $\|V\|_p = \left(\sum_{j=1}^m \sum_{i=1}^n |v|_{ij}^p\right)^{1/p}$ for a matrix $V \in \mathbb{R}^{n \times m}$. The maximum (minimum) elements of a matrix A is denoted by $A_{\max}$ ($A_{\min}$), and we use $A \succ 0$ to denote that the matrix is positive definite. In addition, we let $A_\perp$ denote the orthogonal complement of A, such that $A_\perp' A = 0$. If v is a sparse vector and u is another vector of similar dimension, we define the support index of v as $S_v = \{i | v_i \neq 0\}$ and u_{S_v} as the sub-vector of u indexed by S_v. Similarly, for a matrix A, we use A_{S_v} to denote the matrix derived from A, containing the columns indexed by S_v. We use a similar notation for the complement of the support, i.e S_v^c, $u_{S_v^c}$ and $A_{S_v^c}$. Finally, convergence in distribution (probability) is denoted by $\xrightarrow{d}$ ($\xrightarrow{p}$).

2 The Single-Equation Penalized Error Correction Selector

2.1 Setup

Throughout the paper we let our single variable of interest be denoted by y_t, which we aim to model dynamically with the use of an N-dimensional time series $z_t = (y_t, x_t')'$, described by

$$z_t = \mu + \tau t + \zeta_t, \tag{1}$$

with the stochastic component given by

$$\Delta\zeta_t = \alpha\beta'\zeta_{t-1} + \sum_{j=1}^{p} \phi_j \Delta\zeta_{t-j} + \epsilon_t, \tag{2}$$

where $\epsilon_t = (\epsilon_{1,t}, \epsilon_{2,t}')'$. The model can be rewritten into a VECM form by substituting (1) into (2) to obtain

$$\Delta z_t = \alpha\beta' \left(z_{t-1} - \mu - \tau(t-1)\right) + \tau^* + \sum_{j=1}^{p} \phi_j \Delta z_{t-j} + \epsilon_t, \tag{3}$$

where $\tau^* = (I - \sum_{j=1}^{p} \phi_j)\tau$. 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 $\beta'\mu \neq 0$. 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 $\beta'\tau \neq 0$, the latter implying that the equilibrium error $\beta' z_t$ is a trend stationary process.

We impose the following assumption on the innovations.

Assumption 1. $\{\epsilon_t\}_{t\geq 1}$ is an N-dimensional martingale difference sequence with $\mathbb{E}\left(\epsilon_t\epsilon_t'\right) = \Sigma \succ 0$ and $\mathbb{E}\left|\epsilon_t\right|^{2+\eta} < \infty$ for $\eta > 0$.

Under this assumption, the innovations satisfy the multivariate invariance principle

$$T^{-1/2} \sum_{t=1}^{[T\cdot]} \epsilon_t \to B(\cdot), \tag{4}$$

where $B(\cdot)$ represents a vector Brownian motion with covariance matrix Σ (Phillips and Solo, 1992, p. 983).

For the VECM model to admit a vector moving average (VMA) representation we maintain the following assumptions.

Assumption 2. Define $A(z) := (1-z) - \alpha\beta' z - \sum_{j=1}^{p} \phi_j (1-z) z^j$.

(i) The determinantal equation $|A(z)|$ has all roots on or outside the unit circle.

(ii) α and β are $N \times r$ matrices with $r \leq N$ and $\text{rank}(\alpha) = \text{rank}(\beta) = r$. We adopt the convention that for $r = 0$, we have $\alpha\beta' = 0$ and $\alpha_\perp = \beta_\perp = I_N$.

(iii) The $((N-r)\times(N-r))$ matrix $\alpha_\perp'\left(I_N - \sum_{j=1}^{p} \phi_j\right)\beta_\perp$ is invertible.

The importance in deriving a single-equation model for y_t, our main variable of interest, is to ensure that the the variables modelling the variation in y_t remain exogenous. This is accomplished by orthogonalizing the errors driving the single-equation model, say $\epsilon_{y,t}$, from the errors driving the marginal equation of the endogenous variables x_t. Orthogonalization is achieved by decomposing $\epsilon_{1,t}$ into its best linear prediction based on $\epsilon_{2,t}$ and the corresponding orthogonal prediction error. To this end, partition the covariance matrix of ϵ_t as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{5}$$

such that we obtain

$$\epsilon_{1,t} = (0, \pi_0')\epsilon_t + \left(1, -\pi_0'\right)\epsilon_t = \hat{\epsilon}_{1,t} + \epsilon_{y,t} \tag{6}$$

where $\hat{\epsilon}_{1,t} = \pi_0'\epsilon_{2,t}$ with $\pi_0 = \Sigma_{22}^{-1}\Sigma_{21}$ and $\epsilon_{y,t} = \left(1, -\pi_0'\right)\epsilon_t$. Writing out (6) in terms of the observable time series results in the single-equation model

$$\begin{aligned} \Delta y_t &= \left(1, -\pi_0'\right)\left(\alpha\beta'\left(z_{t-1} - \mu - \tau(t-1)\right) + \tau^* + \sum_{j=1}^{p}\phi_j \Delta z_{t-j}\right) + \pi_0'\Delta x_t + \epsilon_{y,t} \\ &= \delta' z_{t-1} + \pi' w_t + \mu_0 + \tau_0(t-1) + \epsilon_{y,t}, \end{aligned} \tag{7}$$

where $\delta' = \left(1, -\pi_0'\right)\alpha\beta'$, $\pi = (\pi_0', \ldots, \pi_p')'$ with $\pi_j = (1, -\pi_0')\phi_j$ for $j = 1, \ldots, p$, $w_t = (\Delta x_t', \Delta z_{t-1}', \ldots, \Delta z_{t-p}')'$, $\mu_0 = (1, -\pi_0')(\tau^* - \mu)$, $\tau_0 = -(1, -\pi_0')\tau$, and $\epsilon_{y,t} = (1, -\pi_0')\epsilon_t$.

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 of (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.

In general, the implied cointegrating vector δ in the single-equation model for y_t contains a linear combination of the cointegrating vectors in β with their weights being given by $(1, -\pi_0')\,\alpha$. Since the marginal equations of 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) $\text{rank}(\alpha\beta') = 1$, i.e. there is a single cointegrating N-dimensional cointegrating vector β, and (ii) the vector of adjustment rates takes on the form $\alpha = (\alpha_1, 0')'$. 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. Accordingly, 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.

2.2 Estimation Procedure

We propose to estimate (7) with SPECS, which incorporates an ℓ_1-penalty to attain sparse solutions. However, a property of ℓ_1-penalized regression is that its solutions are not equivariant to arbitrary scaling of the variables, which is why the convention is to standardize the data prior to estimation (see Hastie et al., 2008, p. 8). While this practice is fairly innocuous in the stationary setting, this is not the case when dealing with non-stationary variables, as the standard variance estimates are diverging such that care has to be taken when deriving the asymptotic theory. Let $Z_{-1} = (z_0, \dots, z_{T-1})'$, $W = (w_1, \dots, w_T)'$, and write $V = (Z_{-1}, W)$, $\gamma = (\delta', \pi')'$, $\theta = (\mu_0, \tau_0)'$ and $D = (\iota, \bar{t})$, where ι is an N-dimensional vector of ones and $\bar{t} = (0, \dots, T-1)'$. For any data matrix A, coefficient vector b and diagonal weighting matrix σ_A, define $\tilde{A} = A\sigma_A^{-1}$ and $b^s = \sigma_A b$. Then, we can rewrite (7) in standardized matrix form as

$$\begin{aligned}\Delta y &= Z_{-1}\delta + W\pi + \iota\mu_0 + \bar{t}\tau_0 + \epsilon_y = Z_{-1}\sigma_Z^{-1}\sigma_Z\delta + W\sigma_W^{-1}\sigma_W\pi + \iota\mu_0 + \bar{t}\tau_0 + \epsilon_y \\ &= \tilde{Z}_{-1}\delta^s + \tilde{W}\pi^s + \iota\mu_0 + \bar{t}\tau_0 + \epsilon_y = \tilde{V}\gamma^s + D\theta.\end{aligned} \tag{8}$$

We then estimate (8) with our shrinkage estimator, by minimizing the objective function

$$G_T(\gamma^s, \theta) = \left\| \Delta y - \tilde{V}\gamma^s - D\theta \right\|_2^2 + P_\lambda(\gamma^s). \tag{9}$$

The penalty function in (9) takes on the form

$$P_\lambda(\gamma^s) = \lambda_{G,T} \left\| \delta^s \right\|_2 + \lambda_{\delta,T} \sum_{i=1}^{N} \omega_{\delta,i}^{k_\delta} \left| \delta_i^s \right| + \lambda_{\pi,T} \sum_{j=1}^{M} \omega_{\pi,j}^{k_\pi} \left| \pi_j^s \right|, \tag{10}$$

where $\omega_{\delta,i}^{k_\delta} = 1/\left|\hat{\delta}_{Init,i}\right|^{k_\delta}$ and $\omega_{\pi,j}^{k_\pi} = 1/\left|\hat{\pi}_{Init,j}\right|^{k_\pi}$. The tuning parameters k_δ and k^π regulate the degree to which the initial estimates affect the penalty weights, and they should satisfy certain constraints that are specified in the theorems to follow. Throughout this paper we assume that the initial estimators are $\sqrt{T}$-consistent; for example we can use $\hat{\delta}_{OLS}$ and $\hat{\pi}_{OLS}$.[2]

We denote the minimizers of (9) by $\hat{\gamma}^s$ and $\hat{\theta}$ and the de-standardized minimizers by $\hat{\gamma} = \sigma_V^{-1}\hat{\gamma}^s$. The group penalty, regulated by $\lambda_{G,T}$, 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 individiual ℓ_1-penalties, regulated by $\lambda_{\delta,T}$ and $\lambda_{\pi,T}$ serve to enforce sparsity in the coefficient vector δ and π respectively. Furthermore, the penalties are weighted by an initial estimator to enable simultaneous estimation and selection consistency of the coefficients. Note that the deterministic components μ_0 and τ_0 are left unpenalized, as their inclusion in the model is desirable to enable identification of the limiting distribution of the estimators. As shown in Yamada (2017), the inclusion of an unpenalized constant and deterministic trend is equivalent to de-meaning and de-trending the data prior to estimation.

Remark 2. SPECS incorporates an ℓ_2 penalty to achieve sparsity on δ at the group level,

[2] In principal any consistent estimator would suffice, although the required growth rates of the penalty parameters in (10) are intrinsically related to the rate of convergence of the initial estimator.

while inclusion of ℓ_1 penalties ensures sparsity within and outside the group. The resulting optimization problem resembles that of the Sparse-Group Lasso (Simon et al., 2013), and the same algorithm can be employed here with only minor adjustments that account for the presence of just a single group. The *R* code that we make available online implements this algorithm to compute SPECS.

Remark 3. Standardization of unpenalized components does not affect the estimation of penalized components; a feature that can be directly verified by application of Lemma A.4 in Appendix A. Accordingly, we do not explicitly standardize the subset D containing the (deterministic) variables that are left unpenalized.

3 Theoretical Properties

In this Section we derive the theoretical properties of our SPECS estimator. We first establish the consistency and oracle properties of SPECS in Section 3.1. Thereafter we focus on a few specific cases of interest that deserve further attention in Section 3.2.

3.1 Consistency and Oracle Properties

Our first aim is to demonstrate that the SPECS estimator attains the same rate of convergence as the conventional least squares estimator.[3] Following standard convention in the cointegration literature, we first derive the consistency for a linear transformation of the coefficients to avoid singularities in the limits of sample moment matrices resulting from common stochastic trends (e.g. Lütkepohl, 2005, p. 290). In particular, under Assumption 2, the Granger Representation Theorem as displayed in Johansen (1995, p. 49) enables (3) to be written as a VMA process of the form

$$z_t = Cs_t + \mu + \tau t + C(L)\epsilon_t + z_0 = Cs_t + \mu + \tau t + u_t, \tag{11}$$

where $C = \beta_\perp \left(\alpha'_\perp \left(I_N - \sum_{j=1}^{p} \phi_j\right) \beta_\perp\right)^{-1} \alpha'_\perp$, $s_t = \sum_{i=1}^{t} \epsilon_i$, and $u_t = C(L)\epsilon_t + z_0$ a stationary process. In matrix notation, we write

$$Z_{-1} = S_{-1}C' + \iota\mu' + \bar{t}\tau' + U, \tag{12}$$

with $S_{-1} = (s_0, \dots, s_{T-1})'$ and $U = (u_1, \dots, u_T)'$. When cointegration is present in the data, the matrix C will be of rank $N - r$ such that the system may be separated into a stationary and non-stationary component. More specifically, we can define the linear transformation

$$Q := \begin{bmatrix} \beta' & 0 \\ 0 & I_M \\ \alpha'_\perp & 0 \end{bmatrix} \text{ with } Q^{-1} = \begin{bmatrix} \alpha(\beta'\alpha)^{-1} & 0 & \beta_\perp(\alpha'_\perp\beta_\perp)^{-1} \\ 0 & I_M & 0 \end{bmatrix}, \tag{13}$$

[3] As we derive our results for fixed N, we do not need to make an explicit assumption that the conditional model is sparse. Of course, in practical settings where T and N are of comparable size, sparsity is required for good performance. We return to this issue in our simulation study in Section 4.

such that we can decompose the system as $\begin{bmatrix}\xi_{1,t}\\ \xi_{2,t}\end{bmatrix} = Q\begin{bmatrix}z_{t-1}\\ w_t\end{bmatrix}$ with $\xi_{1,t} = \begin{bmatrix}\beta' z_{t-1}\\ w_t\end{bmatrix}$ being a stationary random vector and $\xi_{2,t} = \alpha'_\perp$ the non-stationary component.

Having defined the appropriate transformation, we are now able to state that SPECS attains the same rate of convergence as the OLS estimator. The proofs of all theorems in this section are provided in Appendix A.2.

Theorem 1 (Estimation Consistency). *Assume that* $\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$, $\frac{\lambda_{\delta,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$ *and* $\frac{\lambda_{\pi,T}\sigma_{W,\max}}{\sqrt{T}} \xrightarrow{p} 0$. *Let* $D_T = diag(TI_N, \sqrt{T}I_M)$ *and* $S_T = diag(\sqrt{T}I_{M+r}, TI_{N-r})$. *Then, under Assumption 1 and 2, the estimators* $\hat{\gamma}$ *satisfy:*

1. *No cointegration:* $D_T(\hat{\gamma} - \gamma) = O_p(1)$.

2. *Cointegration:* $S_T Q'^{-1}(\hat{\gamma} - \gamma) = O_p(1)$.

The conditions imposed on the penalty terms limit the amount of shrinkage to prevent excessive shrinkage bias from impeding consistent estimation. Clearly, the admissible growth rates of the penalties are dependent on the stochastic order of the possibly random quantities $\sigma_{Z,\max}$ and $\sigma_{W,\max}$. Consequently, the practice of standardizing the data by scaling each variable by its corresponding estimated standard deviation may influence the restrictions imposed on the growth rate of the penalty. To illustrate, consider the case where z_t contains N random walks (with no drift components). Then, for any $i \in \{1, \ldots, N\}$, the estimated standard deviation is

$$\hat{\sigma}_{Z,i} = \sqrt{\frac{\sum_{t=0}^{T-1} z_{it}^2}{T}} = \sqrt{T}\sqrt{\frac{\sum_{t=0}^{T-1} z_{it}^2}{T^2}} = O_p(\sqrt{T}),$$

such that also $\sigma_{Z,\max} = O_p(\sqrt{T})$. As a result, the requirements $\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$ and $\frac{\lambda_{\delta,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$ translate to $\lambda_{G,T} \to 0$ and $\lambda_{\delta,T} \to 0$. While theoretically feasible, the notion of requiring a vanishing penalty to maintain consistent estimation does not conform with the belief of a sparse DGP. Moreover, the presence of deterministic components in the variables, such as a trend/drift, impact the stochastic order of the standard deviation and, hence, the required growth rates of the penalty. Therefore, we advise against the standard convention of standardization by the estimated standard deviations.

Remark 4. By construction of Q, the resulting convergence stated in part (2) of Theorem 1 is equivalent to the statements $S_T^* Q^*(\hat{\delta} - \delta) = O_p(1)$ and $\sqrt{T}(\hat{\pi} - \pi) = O_p(1)$, where $S_T^* = \text{diag}(\sqrt{T}I_r, TI_{N-r})$ and $Q^* = \begin{bmatrix}(\alpha'\beta)^{-1}\alpha'\\ (\beta'_\perp\alpha_\perp)^{-1}\beta_\perp\end{bmatrix}$.

SPECS performs continuous model selection by estimating sparse solutions through the imposition of individual ℓ_1-penalties and a group penalty. In addition to consistently estimating the model parameters, an additional natural requirement of the estimator is to provide consistent selection of the relevant variables. This property is crucial 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. 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, 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.[4]

Theorem 2 (Selection Consistency)**.** *Assume that* $\frac{\lambda_{\delta,T}\sigma_{Z,\min}}{T^{1-k_\delta/2}} \to \infty$ *and* $\frac{\lambda_{\pi,T}\sigma_{W,\min}}{T^{1/2-k_\pi/2}} \to \infty$*. Then, under the same conditions as in Theorem 1, it holds that whenever* $\gamma_i = 0$*, we have*

$$\mathbb{P}(\hat{\gamma}_i = 0) \to 1.$$

Whereas the estimation consistency in Theorem 1 puts an upper limit on the amount of permissible shrinkage, the selection consistency in Theorem 2 requires a minimum amount of shrinkage to correctly remove irrelevant variables from the model. As before, the implied conditions regulating the growth rates of the penalties depend on the stochastic order of the possibly random quantities in σ_V. Assuming once more that σ_Z is a diagonal matrix containing the standard deviations of Z_{-1}, the condition for selection consistency of the lagged levels translates to $\frac{\lambda_{\delta,T}}{T^{1/2-k_\delta/2}} \to \infty$, as opposed to the $\frac{\lambda_{\delta,T}}{\sqrt{T}} \to 0$ required for estimation consistency. While any choice of $k_\delta > 0$ complies with these conditions from a theoretical point of view, we observe in simulations that the use of standard deviations as a means of standardization results in frequent removal of relevant non-stationary variables, thereby providing another argument against the use of standard deviations.

Remark 5. The only restriction imposed on the growth rate of the group penalty is $\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$, which is necessary to avoid the shrinkage bias induced by the group penalty from impeding estimation consistency. Since $\lambda_{G,T} = 0$ is an admissible value, it follows that SPECS provides both consistent estimation and selection without the addition of a group penalty as well.

Remark 6. A common implementation of the adaptive lasso in the stationary setting sets $k_\delta = k_\pi = 1$. However, in the presence of cointegration the coefficients regulating the long-run dynamics are $\sqrt{T}$-consistent, whereas the presence of common stochastic trends demand a higher rate to stabilize the data. Consequently, assuming $\sigma_V = I_{N+M}$, the conditions on λ_δ are $\frac{\lambda_\delta}{\sqrt{T}} \to 0$ and $\frac{\lambda_\delta}{T^{1-k_\delta/2}} \to \infty$. Hence, a choice of $k_\delta > 1$ is needed to maintain consistent selection of the lagged levels. Intuitively, one may argue that stricter penalization is necessitated by the correlation induced between the levels of variables through the presence of common stochastic trends.

Next, we establish that the limit distribution for the estimates of the non-zero population coefficients is the same as the OLS distribution. When $\delta \neq 0$, it follows from (11) that the

[4] A more detailed discussion of the interpretation of sparsity absent of weak exogeneity is provided in Section 3.2.2.

subset of variables indexed by S_δ has the representation

$$z_{S_\delta,t} = \beta_{\perp,S_\delta}\left(\alpha'_\perp\left(I_N - \sum_{j=1}^{p}\phi_j\right)\beta_\perp\right)^{-1}\alpha_\perp s_{t-1} + v_{S_\delta,t}, \tag{14}$$

where $\beta_{\perp,S_\delta}$ is a $(|S_\delta| \times (N-r))$-dimensional matrix. Let $\beta^0_{S_\delta}$ denote the left nullspace of $\beta_{\perp,S_\delta}$, i.e.

$$\beta^0_{S_\delta} = \left\{x \in \mathbb{R}^{|S_\delta|} | \beta'_{\perp,S_\delta}x = 0\right\}.$$

Note that by construction $\beta'_{\perp,S_\delta}\delta_{S_\delta} = 0$, such that $\dim(\beta^0_{S_\delta}) = r_2 > 0$, where the dimension of the null space is defined as the number of vectors in a corresponding basis.[5] For the case $|S_\delta| > r_2$, define β_{S_δ} as a basis matrix, i.e. a $(|S_\delta| \times r_2)$-dimensional matrix whose columns form a basis for $\beta^0_{S_\delta}$. Equivalently, define $\beta_{S_\delta,\perp}$ as a $(|S_\delta| \times (|S_\delta| - r_2))$-dimensional basis matrix for the orthogonal complement of β_{S_δ}.[6] With the use of these linear transformations, we are able to confirm the convergence to the appropriate asymptotic distribution in the following theorem.

Theorem 3 (Limit Distribution). *Define* $S_{T,S_\gamma} = diag(\sqrt{T}I_{|S_\pi|+r_2}, TI_{|S_\delta|-r_2})$,

$$Q_{S_\gamma} = \begin{bmatrix} \beta'_{S_\delta} & 0 \\ 0 & I_{|S_\pi|} \\ \beta'_{S_\delta,\perp} & 0 \end{bmatrix} \quad \text{and} \quad Q^{-1}_{S_\gamma} = \begin{bmatrix} \beta_{S_\delta}\left(\beta'_{S_\delta}\beta_{S_\delta}\right)^{-1} & 0 & \beta_{S_\delta,\perp}\left(\beta'_{S_\delta,\perp}\beta_{S_\delta,\perp}\right)^{-1} \\ 0 & I_{|S_\pi|} & 0 \end{bmatrix}.$$

Under the same assumptions as in Theorem 1 and 2 it holds that:

1. *No cointegration:* $\sqrt{T}(\hat{\pi}_{S_\pi} - \hat{\pi}_{OLS,S_\pi}) = o_p(1)$.
2. *Cointegration:* $S_{T,S_\gamma}Q'^{-1}_{S_\gamma}(\hat{\gamma}_{S_\gamma} - \hat{\gamma}_{OLS,S_\gamma}) = o_p(1)$.

Remark 7. When all variables in $z_{S_\delta,t}$ are stationary, it must hold that $\beta_{\perp,S_\delta} = 0$ such that $r_2 = \dim(\beta^0_{S_\delta}) = |S_\delta|$. In this special case we define $Q_{S_\gamma} = I_{|S_\gamma|}$ and $S_{T,S_\gamma} = \sqrt{T}$.

As a direct consequence of Theorem 3, we obtain the limit distribution of the SPECS estimator scaled by $\sqrt{T}$.

Corollary 1. *Under the same conditions as in Theorem 3, we have*

$$\sqrt{T}\left(\hat{\gamma}_{S_\gamma} - \gamma_{S_\gamma}\right) \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} \beta_{S_\delta}\Sigma_U^{-1}\beta'_{S_\delta} & 0 \\ 0 & \Sigma_{W_{S_\pi}} \end{bmatrix}\right), \tag{15}$$

where $\Sigma_U = \mathbb{E}\left(\beta'_{S_\delta}u_{S_\delta,t}u'_{S_\delta,t}\beta_{S_\delta}\right)$ *and* $\Sigma_{W_{S_\pi}} = \mathbb{E}\left(w_{S_\pi,t}w'_{S_\pi,t}\right)$. *Furthermore, the matrix* $\beta_{S_\delta}\Sigma_U^{-1}\beta'_{S_\delta}$ *is uniquely defined regardless of the choice of basis matrix* β_{S_δ}.

Remark 8. The oracle results in Theorem 3 suggest that one could test for cointegration by applying standard low-dimensional cointegration tests, such as the Wald test by Boswijk (1994),

[5] For details on the existence of a basis and its relation to the dimension of a finite-dimensional vector space, see Abadir and Magnus (2005, ex. 3.25, 3.29 and 3.30).

[6] Hence, $\beta_{\perp,S_\delta}$ are the rows of $\beta_\perp$ indexed by S_δ, whereas $\beta_{S_\delta,\perp}$ is a matrix whose columns form a basis for the orthogonal complement of β_{S_δ}.

on the selected variables with the same asymptotic distribution as if only the selected variables were considered from the start. 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, with the resulting implication that the finite-sample distribution will not get uniformly close to the respective asymptotic distribution when the sample size grows large. The practical implication is that for certain values of the parameters in the underlying DGP, relying on the oracle properties for post-selection test statistics may be misleading. While developing a valid post-selection cointegration test is certainly of interest, the field of valid post-selection inference is, while rapidly developing, 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 - 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.

Finally, all results thus far have focussed on the convergence and selection of the coefficients corresponding to the stochastic component in our model. Based on these results, we are able to obtain the behaviour of the estimated coefficients governing the deterministic components. However, the rate of convergence of the trend coefficient depends on three characteristics of the DGP, namely the presence of cointegration, the presence of a deterministic trend and whether the trend occurs within the long-run equilibrium. Consequently, we state the following corollary, the proof of which is delegated to the supplementary appendix.

Corollary 2. *Under the assumptions in Theorem 1 and 2, the estimators of the coefficients regulating the deterministic component, i.e. $\hat{\mu}_0$ and $\hat{\tau}_0$, are consistent. In particular, we have*

$$\sqrt{T}(\hat{\mu}_0 - \hat{\mu}_{0,OLS}) = o_p(1),$$
$$R_T(\hat{\tau}_0 - \hat{\tau}_{0,OLS}) = o_p(1),$$

$$\text{where } R_T = \begin{cases} T^{3/2} & \tau = 0 \\ T & \tau \neq 0, \beta'\tau = 0 \cdot \\ T^{1/2} & \tau \neq 0, \beta'\tau \neq 0 \end{cases}$$

In summary, under appropriate assumptions on the penalty rates, SPECS is able to consistently estimate the coefficients of the relevant stochastic variables with the same rate and asymptotic efficiency as the oracle least squares estimator and the inclusion of unpenalized deterministic components allows for an invariant limiting distribution in the same way de-meaning and de-trending is performed in the least squares case. In addition, the irrelevant variables are removed from the model with probability approaching one.

Remark 9. A possible extension to consider is allowing SPECS to select the appropriate deterministic specification by penalizing the coefficients corresponding to a set of deterministic components. While this certainly would be straightforward to implement, the extension of the current theoretical results to this new estimator is less trivial for two main reasons. The first difficulty is that the presence of a trend or drift component in a variable dominates its stochastic variation asymptotically, such that appropriately scaled estimates of sample covariance matrices converge to reduced rank matrices. This feature becomes problematic in instances where inverses or positive minimum eigenvalues are required. While the inclusion of unpenalized deterministic components allows one to effectively regress out the effect of those components (Yamada, 2017), this is not the case when the deterministic components are penalized as well. Secondly, the (pointwise) asymptotic distributions of the estimators are not uniquely identified when the trend coefficient is penalized. Based on the definition given in (9), a specification where $\tau_0 = 0$ can be implied by either (i) $\tau = 0$ or (ii) $\tau \neq 0$ and $\delta'\tau = 0$. It is well known that the limit distribution varies depending on whether a deterministic trend is present in the data (Park and Phillips, 1988, Theorems 3.2 and 3.3), such that identification of the distribution is not ensured when the data is not first de-trended.

3.2 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. 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.

3.2.1 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. Assumptions 1 and 2 under which our theory is developed are compatible with a wide variety of DGPs that include settings where the dataset contains an arbitrary mix of $I(1)$ and $I(0)$ variables. The dataset is simply transformed according to (7) and SPECS provides consistent estimation of the parameters and consistently identifies 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 $z_t = (z_{1,t}', z_{2,t}')'$, where $z_{1,t} \sim I(0)$ and $z_{2,t} \sim I(1)$ and possibly cointegrated. Let the dimensions of $z_{1,t}$ and $z_{2,t}$ be N_1 and N_2 respectively. Then, z_t admits the representation

$$\begin{aligned}\begin{bmatrix}\Delta z_{1,t}\\ \Delta z_{2,t}\end{bmatrix} &= \begin{bmatrix}-I_{N_1} & 0\\ 0 & A\end{bmatrix}\begin{bmatrix}z_{1,t-1}\\ z_{2,t-2}\end{bmatrix} + \Phi(L)\Delta z_{t-1} + \epsilon_t\\ &= Bz_{t-1} + \Phi(L)\Delta z_{t-1} + \epsilon_t,\end{aligned}$$

where $\Phi(L)$ corresponds to a p-dimensional matrix lag polynomial by Assumption 2 and ϵ_t satisfies the conditions in Assumption 1. In addition, we maintain the convention that $A = 0$ when $z_{2,t}$ does not cointegrate. Naturally, the single-equation derived from this VECM has the same form as in (7), with the crucial difference that some of the variables in z_{t-1} are stationary. More specifically, let π_0 be defined as in (6) with the decomposition $\pi_0 = (\pi_{0,1}', \pi_{0,2}')'$. Without loss of generality, if $y_t \sim I(0)$ we let $z_{1,t} = (y_t, x_{1,t}')'$, whereas if $y_t \sim I(1)$ we let $z_{2,t} = (y_t, x_{2,t}')'$. The single-equation model can then be represented as usual

$$\begin{aligned}\Delta y_t &= \left(1, -\pi_0'\right) \left(B z_{t-1} + \Phi(L)\Delta z_{t-1}\right) + \pi_0'\Delta x_t + \epsilon_{y,t} \\ &= \delta' z_{t-1} + \pi' w_t + \epsilon_{y,t}.\end{aligned} \tag{16}$$

or alternatively

$$\Delta y_t = \delta_2' z_{2,t-1} + \pi^{*\prime} w_t^* + \epsilon_{y,t}, \tag{17}$$

where $\pi^* = (\delta_1', \pi')'$ and $w_t^* = (z_{1,t-1}', w_t')'$. This representation highlights that the single-equation model can be decomposed into contributions from the non-stationary variables, i.e. $z_{2,t-1}$, and stationary variables, i.e. $\tilde{w}_t$. Moreover, from our theoretical results in Theorem 3 it follows that

$$\sqrt{T}(\hat{\delta}_1 - \hat{\delta}_{1,OLS}) = o_p(1). \tag{18}$$

In the extreme case, where the DGP consists of a collection of stationary variables and a collection of variables that are integrated of order one which do not cointegrate, we have $\beta_{\perp,S_\delta} = 0$ such that (18) follows directly from Remark 7.

Finally, in Assumption 2 we allow for the case where $\text{rank}(\beta) = N$. One, perhaps slightly cumbersome, interpretation of this scenario is a system in which every variable "trivially cointegrates", which intuitively motivates the applicability of our theoretical results. However, a more common interpretation follows from noting that when $r = N$ the system can be appropriately described by a stationary vector autoregressive model of the form

$$z_t = \Phi(L) z_{t-1} + \epsilon_t,$$

where ϵ_t complies with Assumption 1 and $\Phi(L)$ denotes an invertible matrix lag-polynomial of order p. Following the procedure detailed in section 2, the corresponding single-equation model can be derived as

$$\begin{aligned}y_t &= \pi' x_t + (1, -\pi')\Phi(L) z_{t-1} + \epsilon_{y,t} \\ &= \pi' x_t + (1, -\pi')\Phi(1) z_{t-1} + (1, -\pi')\tilde{\Phi}(L)\Delta z_{t-1} + \epsilon_{y,t},\end{aligned} \tag{19}$$

where the second equation follows from applying the Beveridge-Nelson decomposition to $\Phi(L)$. We can rewrite (19) as

$$\begin{aligned}\Delta y_t &= -y_{t-1} + \pi' x_{t-1} + \pi'\Delta x_t + (1, -\pi')\Phi(1) z_{t-1} + (1, -\pi')\tilde{\Phi}(L)\Delta z_{t-1} + \epsilon_{y,t} \\ &= \delta' z_{t-1} + \pi'\Delta x_t + \Phi^*(L)\Delta z_{t-1} + \epsilon_{y,t},\end{aligned} \tag{20}$$

where $\delta = (1, -\pi')\,(-I + \Phi(1))$ and $\Phi^*(L) = (1, -\pi')\tilde{\Phi}(L)$. Hence, the single-equation model that we estimate can be derived from a stationary system as well. Given that all variables in (20) are stationary time series, 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).

3.2.2 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. 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 $\alpha = \begin{bmatrix} \alpha_1' \\ \alpha_2' \end{bmatrix}$, with α_1 an $(r \times 1)$ row-vector, we obtain

$$\delta = \beta(\alpha_1 - \alpha_2\Sigma_{22}^{-1}\Sigma_{21}).$$

It follows that $\delta_i = 0$ if the condition

$$\beta_i'\left(\alpha_1 - \alpha_2\Sigma_{22}^{-1}\Sigma_{21}\right) = 0 \tag{21}$$

is satisfied, where β_i is the i-th row-vector of β. While this condition may hold in a variety of non-trivial ways, some general cases can be derived that lead to sparsity in δ. For example, a variable x_i that does not cointegrate with any of the variables in the system, i.e. $\beta_i = 0$, will carry a zero coefficient in the derived long-run equilibrium in the single-equation model.

An additional special case is the addition of I(0) variables to the system. Consider the estimation of a standard VECM of the form (3) without any short-run dynamics. Assume, however, that the last variable in the dataset, say w_t, is a stationary white noise series that is mistakenly considered to be integrated of order one. This is described by the following representation

$$\begin{bmatrix} \Delta z_t \\ \Delta w_t \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \beta' & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_{t-1} \\ w_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{z,t} \\ \epsilon_{w,t} \end{bmatrix}.$$

Letting the last row-vector of β be denoted by $\beta_w = (0, \dots, 0, 1)'$, condition (21) then translates to $\beta_w'\Sigma_{22}^{-1}\Sigma_{21} = 0$. A sufficient condition for this to hold is when $\mathbb{E}(\epsilon_{w,t}\epsilon_{z,t}) = 0$, implying that exogenous stationary variables will not be considered as part of the cointegration vector δ. This statement does not come at a surprise, but it also highlights that stationary variables whose errors are correlated with other variables in the system might end up being part of the cointegration vector in the equation for Δy_t. As this correlation contains information about Δy_t, we consider this property desirable for applications such as nowcasting. It does, however, demonstrate that care has to be taken when the aim is direct interpretation of the implied cointegrating vector in the absence of weak exogeneity.

Finally, we explore a slightly less trivial case by considering a VECM model in which Σ, the covariance matrix of the errors, follow a Toeplitz structure with $\sigma_{ij} = \rho^{|i-j|}$. After partitioning Σ as in (5), we can rewrite

$$\Sigma_{21} = \begin{bmatrix} \rho^1 \\ \vdots \\ \rho^N \end{bmatrix} = \begin{bmatrix} \rho^0 & \dots & \rho^{N-1} \\ \vdots & \ddots & \vdots \\ \rho^{N-1} & \dots & \rho^0 \end{bmatrix} \begin{bmatrix} \rho^1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \Sigma_{22}\pi_0, \tag{22}$$

thus showing that $\pi_0 = \Sigma_{22}^{-1}\Sigma_{21} = (\rho, 0, \dots, 0)'$.[7] As $\delta' = (1, -\pi_0')\alpha\beta'$, 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 δ.

4 Simulations

In this section we analyze the selective capabilities and predictive performance of SPECS by means of simulations. We estimate the single-equation model according to the objective function (9) with the following settings for the penalty rates:

1. Ordinary Least Squares (OLS: $\lambda_{G,T} = 0$, $\lambda_{\delta,T} = 0$, $\lambda_{\pi,T} = 0$),

2. Autoregressive Distributed Lag (ADL: $\lambda_{G,T} = 0$, $\lambda_{\delta,T} = \infty$, $\lambda_{\pi,T} > 0$),

3. SPECS - no group penalty (SPECS$_1$: $\lambda_{G,T} = 0$, $\lambda_{\delta,T} > 0$, $\lambda_{\pi,T} > 0$),

4. SPECS - group penalty (SPECS$_2$: $\lambda_{G,T} > 0$, $\lambda_{\delta,T} > 0$, $\lambda_{\pi,T} > 0$)[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 $\delta = 0$.

For the sake of computational efficiency we estimate the solutions for $\lambda_{\delta,T}$ and $\lambda_{\pi,T}$ over a one-dimensional grid, i.e. both penalties are governed by a single universal parameter λ_T. We weigh the universal parameter by initial estimates obtained from a ridge regression. Specifically, we adopt $\omega_{\delta,i}^{k_\delta} = 1/\left|\hat{\delta}_{ridge,i}\right|^{k_\delta}$ and $\omega_{\pi,j}^{k_\pi} = 1/|\hat{\pi}_{ridge,j}|^{k_\pi}$, where $k_\delta = 2$ and $k_\pi = 1$ in accordance with the assumptions in Theorems 1 and 2. We consider 100 possible values for λ_T and choose the final model based on the BIC criterion. For SPECS$_2$, the model selection takes place over

[7] It is straightforward to show that this property carries over to covariance matrices with a block-diagonal Toeplitz structure, with each block $\Sigma^{(k)}$ having the form $\sigma_{i,j}^{(k)} = \rho_{(k)}^{|i-j|}$. The number of non-zero elements in the resulting vector π_0 will equal the number of blocks in the covariance matrix.

[8] As a useful mnemonic, 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 and individual penalty.

Table 1 Simulation Design for the First Study (Dimensionality and Weak Exogeneity)

Low Dimension	α	β	δ
WE	$\alpha_1 \cdot \begin{bmatrix} 1 \\ \mathbf{0}_{9\times 1} \end{bmatrix}$	$\begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{5\times 1} \end{bmatrix}$	$\alpha_1 \cdot \beta$
No WE	$\alpha_1 \cdot \beta$	$\begin{bmatrix} \tilde{\iota} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \tilde{\iota} \end{bmatrix}$	$(1+\rho)\alpha_1 \cdot \begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{5\times 1} \end{bmatrix}$
High Dimension	α	β	δ
WE	$\alpha_1 \cdot \begin{bmatrix} 1 \\ \mathbf{0}_{49\times 1} \end{bmatrix}$	$\begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$	$\alpha_1 \cdot \beta$
No WE	$\alpha_1\beta$	$\begin{bmatrix} \tilde{\iota} & \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \tilde{\iota} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 1} & \tilde{\iota} \\ \mathbf{0}_{35\times 1} & \mathbf{0}_{35\times 1} & \mathbf{0}_{35\times 1} \end{bmatrix}$	$(1+\rho)\alpha_1 \cdot \begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$

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, $\tilde{\iota} = (1, -\iota_4')'$, $\beta^* = (\mathbf{1}_{3\times 3} \otimes \tilde{\iota})$, and $\alpha_1 = -0.5, -0.45, \ldots, 0$ regulates the adjustment rate towards the equilibrium.

a two-dimensional grid consisting of 100 values for λ_T and 10 possible values for $\lambda_{G,T}$. We note that while the use of the single universal penalty λ_T significantly reduces the dimension of the search space, this heuristic may negatively impact the performance of SPECS. Since this choice of implementation does not impact the ADL model, the relative performance gain of SPECS over the ADL model would likely be underestimated.

We now consider three different settings under which we analyze the performance of our SPECS estimator.

4.1 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. The general DGP from which we simulate our data is given by the equation

$$\Delta z_t = \alpha\beta' z_{t-1} + \phi_1 \Delta z_{t-1} + \epsilon_t, \tag{23}$$

with $t = 1, \ldots, T = 100$, $\epsilon_t \sim \mathscr{N}(0, \Sigma)$ and $\sigma_{ij} = 0.8^{|i-j|}$. Furthermore, ϕ_1, the coefficient matrix regulating the short-run dynamics is generated as $0.4 \cdot I_N$, where N varies depending on the specific DGP considered. Based on this DGP, the single-equation model takes on the form

$$\Delta y_t = \delta' z_{t-1} + \pi_0' \Delta x_t + \pi_1' \Delta z_{t-1} + \epsilon_{y,t},$$

with π_0 and π_1 as defined in (7). We consider a total of four different settings, corresponding to (i) different combinations of dimensionality (low/high) and (ii) weak exogeneity (present/absent). The corresponding parameter settings, and their implied cointegrating vector δ, are tabulated in Table 1.

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 mentioned in Remark 8, 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) describes the proportion of correctly selected variables:

$$PCS = \frac{|\{\hat{\gamma}_j \neq 0\} \cap \{\gamma_j \neq 0\}|}{|\{\gamma_j \neq 0\}|}.$$

Alternatively, the Proportion of Incorrect Selection (PICS) describes, as the name may suggest, the proportion of incorrectly selected variables:

$$PICS = \frac{|\{\hat{\gamma}_j \neq 0\} \cap \{\gamma_j = 0\}|}{|\{\gamma_j = 0\}|}.$$

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 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 5. 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 $\hat{y}_T = \hat{\delta}' z_{T-1} + \hat{\pi}' \Delta x_T + \hat{\phi}' \Delta z_{T-1}$, where by construction $\hat{\delta} = 0$ 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 subset of relevant variables.

Figure 1 visually displays the evolution of our performance metrics over a range of values for α_1, 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. Focussing 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 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

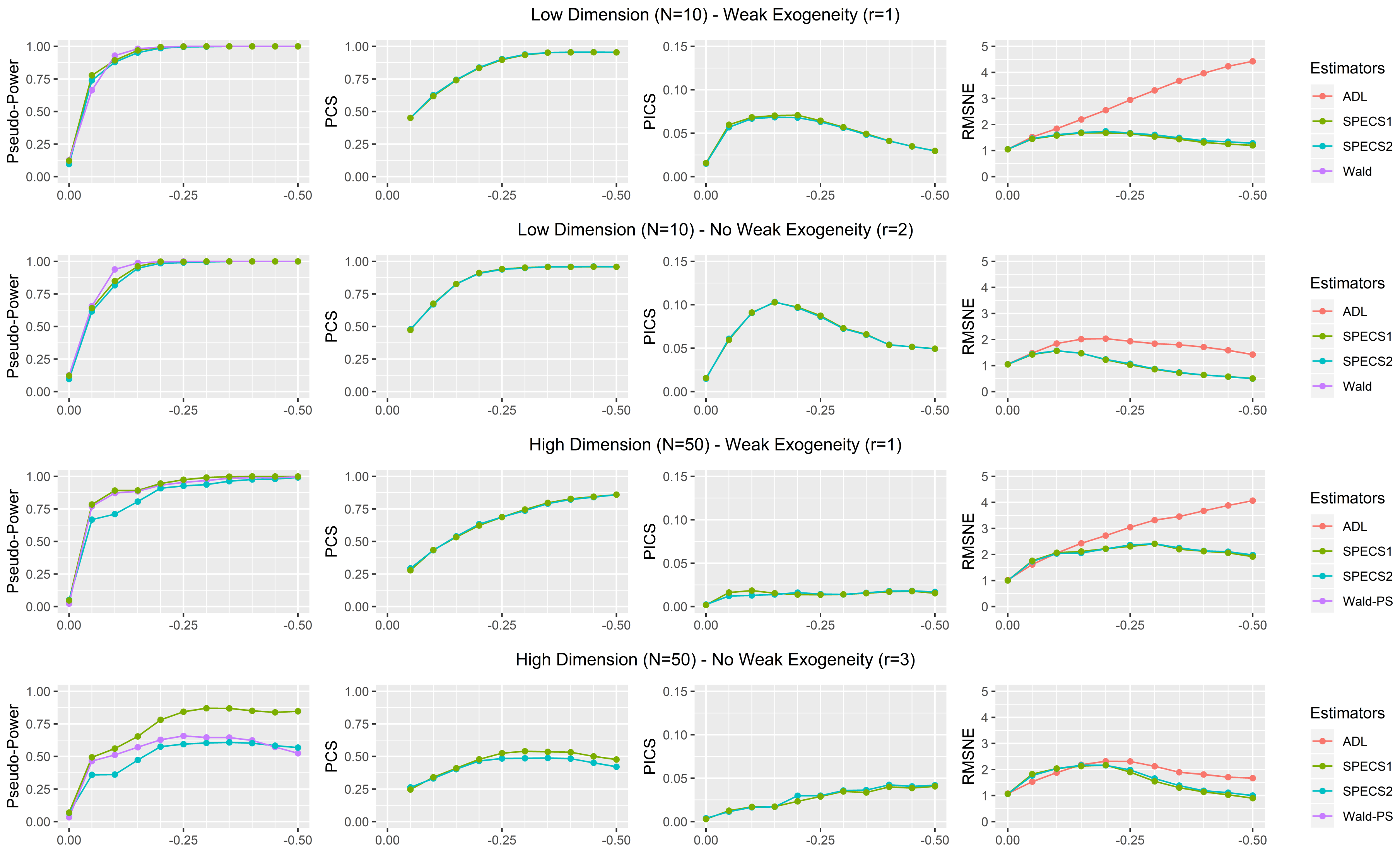

Figure 1: Pseudo-Power, Proportion of Correct Selection (PCS), Proportion of Incorrect Selection (PICS) and Root Mean Squared Nowcast Error (RMSNE) for Low- and High-Dimensional specifications. The adjustment rate multiplier α_1 is on the horizontal axis.

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 Figure 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.

4.2 Mixed Orders of Integration

We move on to an analysis of 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 individual series. We refer to this procedure as the ADL-ADF model. Based on the general DGP (23), 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 2, 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 B in β, with elements $b_{ii} = 1$ ($b_{ii} \sim U(0, 0.2)$ in the low-dimensional (high-dimensional) case. It can be seen from the last column in Table 2, that due to the stationary of the dependent variable, the first element in δ will always be equal to -1, whereas an additional five-dimensional cointegrating vector enters the single-equation model for positive values of α_1. For the scenario

Table 2 Simulation Design for the Second Study (Mixed Orders of Integration)

Mixed Order	α	β	δ
$y \sim I(0)$	$\begin{bmatrix} 1 & 0 & \mathbf{0}_{1\times 24} \\ \mathbf{0}_{15\times 1} & \alpha_1\beta^* & \mathbf{0}_{15\times 24} \\ \mathbf{0}_{10\times 1} & \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 24} \\ \mathbf{0}_{24\times 1} & \mathbf{0}_{24\times 3} & I_{24} \end{bmatrix}$	$\begin{bmatrix} -b & 0 & \mathbf{0}_{1\times 24} \\ \mathbf{0}_{15\times 1} & \beta^* & \mathbf{0}_{15\times 24} \\ \mathbf{0}_{10\times 1} & \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 24} \\ \mathbf{0}_{24\times 1} & \mathbf{0}_{24\times 3} & -\boldsymbol{B}_{24\times 24} \end{bmatrix}$	$\begin{bmatrix} -1 \\ -\rho\alpha_1\tilde{\iota} \\ \mathbf{0}_{44\times 1} \end{bmatrix}$
$y \sim I(1)$	$\begin{bmatrix} \alpha_1\beta^* & \mathbf{0}_{15\times 25} \\ \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 25} \\ \mathbf{0}_{25\times 3} & I_{25} \end{bmatrix}$	$\begin{bmatrix} \beta^* & \mathbf{0}_{15\times 25} \\ \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 25} \\ \mathbf{0}_{25\times 3} & -\boldsymbol{B}_{25\times 25} \end{bmatrix}$	$(1+\rho)\alpha_1 \cdot \begin{bmatrix} \tilde{\iota} \\ \mathbf{0}_{45\times 1} \end{bmatrix},$

Notes: see notes in Table 1. Additionally, we define $b = 1$ ($b \sim U(0, 0.2)$) and $\boldsymbol{B}$ as a diagonal matrix with $b_{ii} = 1$ ($b_{ii} \sim U(0, 0.2)$) in the absence (presence) of persistence.

in which the dependent variable is integrated of order one, the first 15 variables, y included, are all cointegrated and 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 2 that $\delta = 0$ only if $\alpha_1 > 0$, 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 Figure 2.

In the first two rows of Figure 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 plot displays a constant line at 1. Interestingly, the 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 we indeed find to be accurately estimated by the OLS oracle procedure, the model choice seems to be motivated by a favourable bias-variance trade-off. In line with this conjecture, the nowcast performance of SPECS occasionally exceeds that of the oracle procedure in which a larger number of parameters must be estimated. Focussing on the ADL models, we observe that 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

[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.

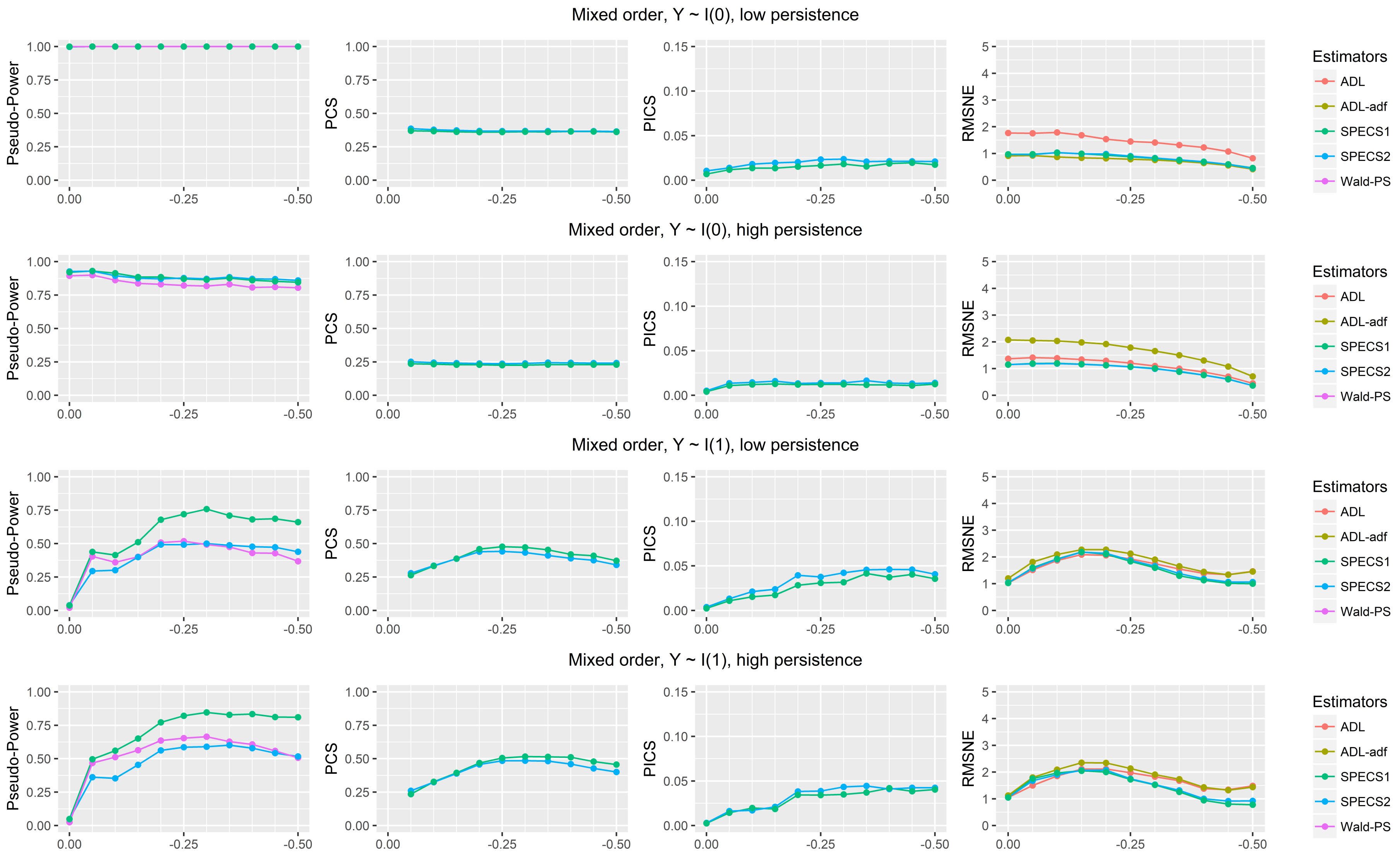

Figure 2: Pseudo-Power, Proportion of Correct Selection (PCS), Proportion of Incorrect Selection (PICS) and Root Mean Squared Nowcast Error (RMSNE) for four Mixed Order specifications. The adjustment rate multiplier α_1 is on the horizontal axis.

Table 3 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

the nowcasting performance of the ADL-ADF model deteriorates drastically. We stress the relevance of this result, given that this estimation method in combination with similar pre-testing procedure is fairly common practice. Somewhat surprisingly, the ADL model in differences nowcasts almost as well as SPECS for this particular setting. Overall, however, the nowcast accuracy of SPECS remains the most accurate and, equally important, most stable across all specifications.

Continuing the analysis of mixed order datasets, rows 3 and 4 of Figure 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 $\delta \neq 0$ only when $\alpha_1 > 0$. 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.

4.3 A Dense Factor Model

Finally, to avoid idealizing the results through a choice of DGPs that suits our procedure, we consider a more adverse setting by generating the data with a non-stationary factor structure, while allowing 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 that we adopt corresponds to setting III in Palm et al. (2011, p. 92). For completeness, the DGP is given by

$$y_t = \Lambda F_t + \omega_t,$$

where y_t is a (50×1) time series process, F_t is a single scalar factor and

$$\begin{aligned} F_t &= \phi F_{t-1} + f_t, \\ \omega_{i,t} &= \theta_i \omega_{i,t-1} + v_{i,t}. \end{aligned}$$

Furthermore,

$$\begin{aligned} v_t &= A_1 v_{t-1} + \epsilon_{1,t} + B_1 \epsilon_{1,t-1}, \\ f_t &= \alpha_2 f_{t-1} + \epsilon_{2,t} + \beta_2 \epsilon_{2,t-1}, \end{aligned}$$

where $\epsilon_{1,t} \sim \mathcal{N}(0, \Sigma)$ and $\epsilon_{2,t} \sim \mathcal{N}(0, 1)$.

The comparison focusses exclusively on the nowcasting performance for a setting without dynamics ($A_1 = B_1 = \alpha_2 = \beta_2 = 0$) and a setting with dynamics ($\alpha_2 = \beta_2 = 0.4$). The construction of A_1 and B_1 is analogous to Palm et al. (2011, p. 93). We report the RMSNEs of SPECS relative to the ADL in Table 3. 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 model are clear from the first part of the simulations.

5 Empirical Application

Inspired by Choi and Varian (2012), we consider the possibility of nowcasting Dutch unemployment with our method based on Google Trends data. Google Trends are hourly updated time series consisting of normalized indices depicting the volume of search queries entered in Google originating from a certain geographical area that were entered into Google. The Dutch unemployment rates are made available by Statistics Netherlands, an autonomous administrative body focussing 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[10] 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 January 2004 to December 2017. While the full dataset contains 100 unique search queries, a number of these contain zeroes 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.[11]

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

$$\Delta y_t = \pi_0' \Delta x_t + \sum_{j=1}^{p} \pi_j' \Delta z_{t-j} + \epsilon_t.$$

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, we include another ADL model where the decision to difference is based on a preliminary ADF test referred to as ADL-ADF.[12] Finally,

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

[11] The dataset is available with the R code at https://sites.google.com/view/etiennewijler.

[12] 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 Appendix B.2.

p	N'	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*

Table 4 This table reports the number of parameters estimated, N', as well as the Mean-Squared Nowcast Error relative to the ADL model for varying number of lagged differences p. We use * to denote rejection by the Diebold-Mariano test at the 10% significance level.

SPECS estimates

$$\Delta y_t = \delta' z_{t-1} + \pi_0' \Delta x_t + \sum_{j=1}^{p} \pi_j' \Delta z_{t-i} + \epsilon_t.$$

All tuning parameters are obtained by time series cross-validation (Hyndman, 2016) and we use $k_\delta = 1.1$ which performed well based on a preliminary analysis.[13] 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. In Table 4 we report 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, (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 Figure 3 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 a given variables is included in a given nowcast, and red vice versa. The figure also displays the actual unemployment rates compared to the nowcasted values.

Figure 3 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

[13] We compared the nowcast accuracy for varying $k_\delta \in [0, 2]$ and observed that the lowest nowcast accuracy was obtained for $k_\delta = 1.1$, whereas for values of $k_\delta > 1.5$ almost all lagged levels were consistently excluded. In the latter case, the nowcast accuracy of SPECS was similar to that of the ADL benchmark.

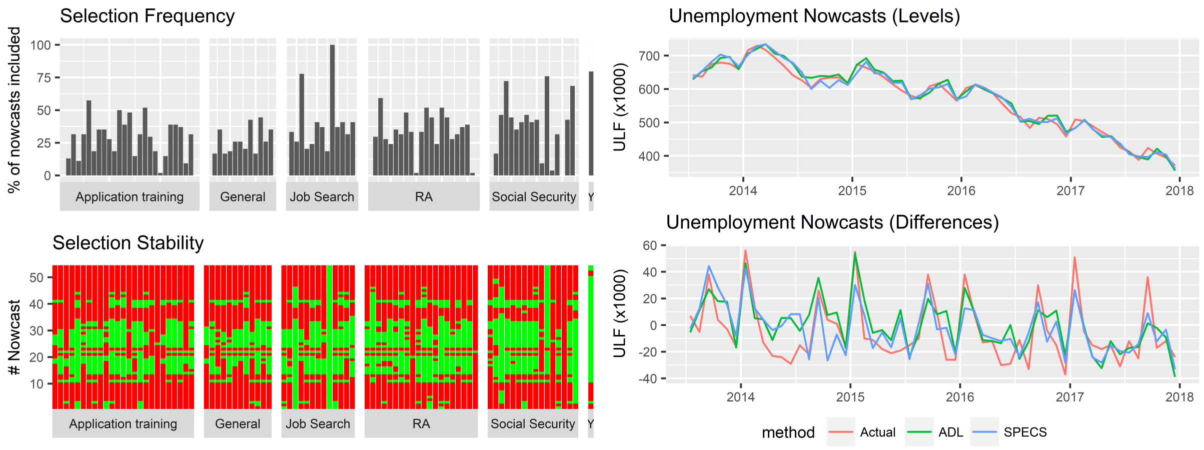

Figure 3: *Top-left*: Selection frequency, measured as the percentage of all nowcasts the variable was selected. *Bottom-left*: Selection stability with green indicating a variable was included in the nowcast model and red indicating exclusion. *Right*: Actual versus predicted unemployed labour force (ULF) in levels and differences.

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", 72%), where the stated percentages indicate the percentage 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 Figure 3. 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 Figure 3 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.

6 Conclusion

In this paper we propose the use of SPECS as an automated approach to cointegration modelling. 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. A simulation exercise confirms strong selective and predictive capabilities in both low and high dimensions with impressive gains over a benchmark penalized ADL model that ignores cointegration in the dataset. The assumption of weak exogeneity is important for efficient estimation and interpretation of the model. However, while our estimator is not entirely insensitive to this assumption, the simulation results demonstrate that the selective capabilities remain adequate 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 results in our simulation study. As a result, we believe that our proposed estimator, which is easily implemented with readily available tools at 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.

Appendix A Proofs

A.1 Preliminary Results

Similar to (8), we write the conditional error correction model in matrix notation as

$$\Delta y = Z_{-1}\delta + W\pi + \iota\mu_0 + \bar{t}\tau_0 + \epsilon_y,$$

where by construction $\mathbb{E}(\epsilon_t \epsilon_{y,t}) = 0$. Following the discussion in Section 3.2.1, we may equivalenty write this as

$$\Delta y = Z_{1,-1}\delta_1 + W_2\pi_2 + \iota\mu_0 + \bar{t}\tau_0 + \epsilon_y,$$

where $Z_{1,-1}$ contains the subset of variables in Z_{-1} that are $I(1)$ and $W_2 = (Z_{2,-1}, W)$ with $Z_{2,-1}$ the subset of $I(0)$ variables. For notational convencience we proceed under the assumption that all variables are integrated of order one such that $Z_{-1} = Z_{1,-1}$. We stress, however, that this assumption is without loss of generality, as one may replace the matrices in the proof below by their decomposed variants without additional complications. Under Assumption 2, the moving average representation of the N-dimensional time series z_t is given by

$$Z_{-1} = S_{-1}C' + \iota\mu' + \bar{t}\tau' + U_{-1}, \tag{A.1}$$

where $S_{-1} = (s_0, \ldots, s_{T-1})'$, with $s_t = \sum_{i=1}^{t} \epsilon_i$, $C = \beta_\perp \left(\alpha'_\perp \left(I_N - \sum_{j=1}^{p} \phi_j \right) \beta_\perp \right)^{-1} \alpha_\perp$, and $U_{-1} = (u_0, \ldots, u_{T-1})'$, with $u_t = C(L)\epsilon_t + z_0$ consisting of a linear process plus initial conditions.

We first present a number of useful intermediary results that will aid the proofs of our main results. The first of such results details the weak convergence of integrated processes. Based on Assumption 1, the following results are well-known in the literature.

Lemma A.1. *Let $B(r)$ denote a Brownian Motion with covariance matrix Σ and define $D = (\iota, \bar{t})$ and $M_D = I - D(D'D)^{-1}D'$. Then, under Assumption 1,*

(a) $T^{-2}S'_{-1}S_{-1} \xrightarrow{d} \int_0^1 B(r)B(r)'dr$

(b) $T^{-3/2}S'_{-1}\iota \xrightarrow{d} \int_0^1 B(r)dr$

(c) $T^{-5/2}S'_{-1}\bar{t} \xrightarrow{d} \int_0^1 rB(r)dr$

(d) $T^{-1}S'_{-1}\epsilon_y \xrightarrow{d} \int_0^1 B(r)dB_{\epsilon_y}(r)$

(e) $T^{-3/2}S'_{-1}U_{-1} \xrightarrow{d} \left(\int_0^1 B(r)dr \right) z'_0$

(f) $T^{-1}U'_{-1}U_{-1} \xrightarrow{p} \sum_{j=0}^{\infty} C_j \Sigma C'_j$.

In addition, these results carry through for $S^*_{-1} = M_D S_{-1}$ by replacing $B(r)$ for $B^*(r) = B(r) - \int_0^1 B(s)ds - 12\left(r - \frac{1}{2}\right)\int_0^1 \left(s - \frac{1}{2}\right) B(s)ds$ in the corresponding limit distributions.

Proof. Under Assumption 1, Phillips and Solo (1992) show that ϵ_t satisfies a multivariate invariance principle. Consequently, the convergence results (a)-(e) are directly implied by Lemma 2.1 in Park and Phillips (1989), whereas (f) is a standard result for linear processes (e.g. Brockwell and Davis, 1991, p. 404). The claim that the convergence holds true after de-meaning and de-trending, i.e. after pre-multiplication of the data matrix by M_D, can be found in most standard time series textbooks, see for example Davidson (2000, p. 354). □

Absent of cointegration in the data, the matrix C will be of full rank. In this setting, the following convergence results are well-established in the literature.

Lemma A.2. *Let M_D be defined as in Lemma A.1. Then, under Assumptions 1 and 2,*

(a) $T^{-2}Z'_{-1}M_DZ_{-1} \xrightarrow{d} C\left(\int_0^1 B^*(r)B^{*\prime}(r)dr\right)C'$,

(b) $T^{-3/2}Z'_{-1}M_DW \xrightarrow{p} 0$,

(c) $T^{-1}W'M_DW \xrightarrow{p} \Sigma_w$,

(d) $T^{-1}Z'_{-1}M_D\epsilon_y \xrightarrow{d} \int_0^1 B^*(r)dB_{\epsilon_y}(r)$,

(e) $T^{-1/2}W'M_D\epsilon_y \xrightarrow{d} \mathscr{N}\left(0, \sigma^2_{\epsilon_y}\Sigma_w\right)$,

where $B^*(r)$ as in Lemma A.1.

Proof. These results are standard and details of the proof are omitted. Briefly, one can plug in the definitions of the matrices Z_{-1} and W based on (A.1), and apply Lemma A.1 to show the results (a)-(d). Result (e) follows from an application of a central limit theorem for linear process as in Theorem 3.4 in Phillips and Solo (1992). □

When cointegration is present in the data, the matrix C will be of rank $N - r$, which will be problematic in applications where the inverse is required. A workaround is to transform the system into a stationary and non-stationary component. From (A.1), it follows that

$$Z_{-1}\beta = \iota\mu'\beta + \bar{t}\tau'\beta + U_{-1}\beta$$

is a (trend-)stationary process and

$$Z_{-1}\alpha_\perp = S^{-1}C'\alpha_\perp + \iota\mu'\alpha_\perp + \bar{t}\tau'\alpha_\perp + U_{-1}\alpha_\perp$$

contains the stochastic trends.[14] Accordingly, define the linear transformation

$$Q := \begin{bmatrix} \beta' & 0 \\ 0 & I_M \\ \alpha'_\perp & 0 \end{bmatrix} \text{ with } Q^{-1} = \begin{bmatrix} \alpha(\beta'\alpha)^{-1} & 0 & \beta_\perp(\alpha'_\perp\beta_\perp)^{-1} \\ 0 & I_M & 0 \end{bmatrix},$$

and let $V = (Z_{-1}, W)$. Then,

$$VQ = \begin{bmatrix} Z_{-1}\beta & W & Z_{-1}\alpha_\perp \end{bmatrix} = \begin{bmatrix} V_1 & V_2 \end{bmatrix},$$

with $V_1 = (Z_{-1}\beta, W)$. We maintain the convention that for the case $r = N$, we define $\beta_\perp = \alpha_\perp = 0$ and $V = V_1$. Based on this decomposition, we recall a number of convergence results under the remark that the results involving V_2 are relevant only for the case $r < N$.

Lemma A.3. *Let M_D be defined as in Lemma A.1. Then, under Assumptions 1 and 2,*

(a) $T^{-2}V'_2M_DV_2 \xrightarrow{d} \alpha'_\perp C\left(\int_0^1 B^*(r)B^{*\prime}(r)dr\right)C'\alpha_\perp$

(b) $T^{-3/2}V'_2M_DV_1 \xrightarrow{p} 0$

[14] Note that $C'\alpha_\perp$ simplifies to $\alpha_\perp$ when $\phi_j = 0$ for $j = 1, \ldots, p$.

(c) $T^{-1}V_1'M_DV_1 \xrightarrow{p} \Sigma_{V_1}$

(d) $T^{-1}V_2'M_D\epsilon_y \xrightarrow{d} \alpha_\perp' C\left(\int_0^1 B^*(r)dB_{\epsilon_y}(r)\right)$

(e) $T^{-1/2}V_1'M_D\epsilon_y \xrightarrow{d} \mathscr{N}\left(0, \sigma_{\epsilon_y}^2\Sigma_{V_1}\right)$

Proof. These results correspond to Lemma 1 in Ahn and Reinsel (1990) and we refer the reader to the original paper for their proofs. □

The final preliminary result that will be used is an extension of the Frisch-Wraugh-Lovell theorem to penalized regression.

Lemma A.4. *Let M_D be defined as in Lemma A.1 and consider the solutions to the following two lasso regressions:*

$$\left(\hat{\gamma}', \hat{\theta}'\right)' = \underset{\gamma,\theta}{arg\ min} \ \|\Delta y - V\gamma - D\theta\|_2^2 + P_\lambda(\gamma), \tag{A.2}$$

$$\breve{\gamma} = \underset{\gamma}{arg\ min} \ \|M_D\Delta y - M_DV\gamma\|_2^2 + P_\lambda(\gamma), \tag{A.3}$$

where

$$P_\lambda(\gamma) = \lambda_G\left(\sum_{i=1}^N |\gamma_i|^2\right)^{1/2} + \sum_{i=1}^N \lambda_{2,i}|\gamma_i| + \sum_{j=1}^M \lambda_{3,j}|\gamma_{N+j}|.$$

Based on (A.2) *and* (A.3) *we have*

(i) $\hat{\gamma} = \breve{\gamma}$;

(ii) $\hat{\theta} = (D'D)^{-1}D'(\Delta y - V'\hat{\gamma})$.

Proof of Lemma A.4. The proof is provided in Yamada (2017) for the standard lasso. In our case the only difference is the addition of the derivative of the group penalty in the subgradient vector. Once this contribution is added the proof is entirely analogous. □

A.2 Proofs of Theorems

Proof of Theorem 1. The proof largely follows along the lines of Liao and Phillips (2015). Recall from (9) that we obtain the standardized estimates $\hat{\gamma}^s$ by minimizing

$$G_T(\gamma^s, \theta) = \left\|\Delta y - \tilde{V}\gamma^s - D\theta\right\|_2^2 + P_\lambda(\gamma^s),$$

which by Lemma A.4 are equivalent to those obtain from minimizing

$$G_T(\gamma^s) = \left\|M_D\left(\Delta y - \tilde{V}\gamma^s\right)\right\|_2^2 + P_\lambda(\gamma^s), \tag{A.4}$$

where we defined $\tilde{V} = V\sigma_V^{-1}$ and $\gamma^s = \sigma_V\gamma$, with $\sigma_V = \text{diag}(\sigma_Z, \sigma_W)$ a diagonal weighting matrix, which results in the decomposition $\gamma^s = (\delta^{s\prime}, \pi^{s\prime})' = (\delta'\sigma_Z, \pi'\sigma_W)'$. By construction we have $G_T(\hat{\gamma}^s) < G_T(\gamma^s)$, from which it follows that

$$(\hat{\gamma}^s - \gamma^s)'\tilde{V}'M_D\tilde{V}(\hat{\gamma}^s - \gamma^s) - 2(\hat{\gamma}^s - \gamma^s)'\tilde{V}'M_D\epsilon_y \leq P_\lambda(\gamma^s) - P_\lambda(\hat{\gamma}^s),$$

which is equivalent to

$$(\hat{\gamma} - \gamma)'V'M_DV(\hat{\gamma} - \gamma) - 2(\hat{\gamma} - \gamma)'V'M_D\epsilon_y \leq P_\lambda(\gamma^s) - P_\lambda(\hat{\gamma}^s). \tag{A.5}$$

The strategy to derive consistency of the estimators consists of appropriately bounding both sides of (A.5) from which the results in Theorem 1 can be obtained. We first proceed under the assumption that there is no cointegration present in the underlying DGP, i.e. $\delta = 0$. Define the scaling matrix $D_T = \text{diag}(TI_N, \sqrt{T}I_N)$. Then, a lower bound for the first left-hand side term of (A.5) is given by

$$(\hat{\gamma} - \gamma)D_TD_T^{-1}V'M_DVD_T^{-1}D_T(\hat{\gamma} - \gamma) \geq \|D_T(\hat{\gamma} - \gamma)\|_2^2 \phi_{\min},$$

where $\phi_{\min}$ is the smallest eigenvalue of $D_T^{-1}V'M_DVD_T^{-1}$. Let A be a $(N \times N)$ matrix and define $\rho(A) : \mathbb{R}^{N\times N} \to \mathbb{C}$ as the function that extracts its minimum eigenvalue. Then, by the continuous mapping theorem, it follows that

$$\phi_{\min} \xrightarrow{d} \rho_{\min}\left(\begin{bmatrix} C\left(\int_0^1 B^*(r)B^{*\prime}(r)dr\right)C' & 0 \\ 0 & \Sigma_W \end{bmatrix}\right) > 0, \quad \text{a.s.} \tag{A.6}$$

The almost sure positiveness of the minimum eigenvalue is motivated as follows. Absent of cointegration, C is full rank and $\int_0^1 B^*(r)B^{*\prime}(r)dr \succ 0$ almost surely by Lemma A2 in Phillips and Hansen (1990), such that $C\left(\int_0^1 B^*(r)B^{*\prime}(r)dr\right)C' \succ 0$. Additionally, $\Sigma_W \succ 0$ as a consequence of Assumption 1. Then, as a direct consequence of (A.6), it also holds that $\mathbb{P}(\phi_{\min} > 0) \to 1$.

The second term in (A.5) is bounded by

$$(\hat{\gamma} - \gamma)'D_TD_T^{-1}V'M_D\epsilon_y \leq \|D_T(\hat{\gamma} - \gamma)\|_2 \left\|D_T^{-1}V'M_D\epsilon_y\right\|_2 = \|D_T(\hat{\gamma} - \gamma)\|_2\, a_T,$$

where $a_T = \left\|D_T^{-1}V'M_D\epsilon_y\right\|_2 = O_p(1)$ by Lemma A.2.

Next, we derive an upper bound for the right-hand side of (A.5). For ease of exposition, we write $\lambda_{2,i} = \omega_{\delta,i}^{k_\delta}\lambda_{\delta,T}$ and $\lambda_{3,j} = \omega_{\pi,j}^{k_\pi}\lambda_{\pi,T}$. First, note that

$$\lambda_{G,T}\left(\|\delta^s\|_2 - \left\|\hat{\delta}^s\right\|_2\right) \leq \lambda_{G,T}\left\|\hat{\delta}^s - \delta^s\right\|_2 \leq \lambda_{G,T}\|\hat{\gamma}^s - \gamma^s\|_2 \leq T^{-1/2}\lambda_{G,T}\|D_T(\gamma^s - \hat{\gamma}^s)\|_2,$$

where $T^{-1/2}\lambda_{G,T} \to 0$ by assumption. To bound the difference between the individual penalties, we define $\lambda_\gamma = (\lambda_2', \lambda_3')'$ and λ_{S_γ} as an $(N+M)$-dimensional vector with $\lambda_{S_\gamma,i} = \lambda_{\gamma,i}\mathbb{1}\{\gamma_i \neq 0\}$. Then,

$$\sum_{i=1}^{N}\lambda_{2,i}\left(|\delta_i^s| - \left|\hat{\delta}_i^s\right|\right) + \sum_{j=1}^{M}\lambda_{3,j}\left(|\pi_j^s| - |\hat{\pi}_j^s|\right) \leq \sum_{i\in S_\delta}\lambda_{2,i}\left(|\delta_i^s| - \left|\hat{\delta}_i^s\right|\right) + \sum_{j\in S_\pi}\lambda_{3,j}\left(|\pi_j^s| - |\hat{\pi}_j^s|\right)$$

$$\leq \sum_{i\in S_\delta}\lambda_{2,i}\left|\hat{\delta}_i^s - \delta_i^s\right| + \sum_{j\in S_\pi}\lambda_{3,j}\left|\hat{\pi}_j^s - \pi_j^s\right| = \lambda_{S_\gamma}'\sigma_VD_T^{-1}D_T\left|\hat{\gamma}_i - \gamma_i\right| \leq \left\|D_T^{-1}\sigma_V\lambda_{S_\gamma}\right\|_2\|D_T(\hat{\gamma}_i^s - \gamma_i^s)\|_2.$$

Furthermore, it is straightforward to see that $\left\|D_T^{-1}\sigma_V\lambda_{S_\gamma}\right\|_2 = o_p(1)$ if

$$\frac{\lambda_{3,j}\sigma_{W,jj}}{\sqrt{T}} = \frac{\lambda_{\pi,T}\sigma_{W,jj}}{\sqrt{T}\left|\hat{\pi}_{OLS,j}^{k_\pi}\right|} = o_p(1),$$

for all $j \in S_\pi$. Since $\hat{\pi}_{OLS,j} \xrightarrow{p} \pi_j$ by the consistency of the OLS estimator, we require the condition $\frac{\lambda_{\pi,T}\sigma_{W,\max}}{\sqrt{T}} \xrightarrow{p} 0$.

Combining the bounds obtained thus far we can rewrite (A.5) as

$$\phi_{\min}\left\|D_T(\hat{\gamma}-\gamma)\right\|_2^2 - 2a_T\left\|D_T(\hat{\gamma}-\gamma)\right\|_2 \leq \left(T^{-1/2}\lambda_G + \left\|D_T^{-1}\sigma_V\lambda_{S_\gamma}\right\|_2\right)\left\|D_T(\hat{\gamma}-\gamma)\right\|_2,$$

from which it follows that

$$\left\|D_T(\hat{\gamma}-\gamma)\right\|_2 \leq \phi_{\min}^{-1}2a_T + \phi_{\min}^{-1}\left(T^{-1/2}\lambda_G + \left\|D_T^{-1}\sigma_V\lambda_{S_\gamma}\right\|_2\right) = O_p(1),$$

which demonstrates the consistency of our estimator absent of cointegration.

Next, we assume there exists cointegration between the variables in the DGP, i.e. $\delta \neq 0$. Let Q be defined as in (13) and define the scaling matrix $S_T = \operatorname{diag}(\sqrt{T}I_{M+r}, TI_{N-r})$. By arguments analogous to the case without cointegration, we obtain a lower bound for the first left-hand side term of (A.5) as

$$(\hat{\gamma}-\gamma)'Q^{-1}S_TS_T^{-1}QV'M_DVQ'S_T^{-1}S_TQ'^{-1}(\hat{\gamma}-\gamma) \leq \psi_{\min}\left\|S_TQ'^{-1}(\hat{\gamma}-\gamma)\right\|_2^2,$$

where $\psi_{\min}$ is the smallest eigenvalue of $S_T^{-1}QV'M_DVQ'S_T^{-1}$. By Lemma A.3 and the continuous mapping theorem, we have

$$\psi_{\min} \xrightarrow{d} \rho_{\min}\left(\begin{bmatrix}\Sigma_{V_1} & 0 \\ 0 & \alpha_\perp' C\left(\int_0^1 B^*(r)B^{*\prime}(r)dr\right)C'\alpha_\perp\end{bmatrix}\right) > 0, \quad \text{a.s.}$$

The almost sure positiveness is implied by the fact that the matrix $\Sigma_{V_1} \succ 0$ as a consequence of Assumption 1. Additionally, by Assumption 2, $\alpha_\perp' C$ is an $(r \times N)$-dimensional matrix of full-row rank r and $\int_0^1 B^*(r)B^{*\prime}(r)dr \succ 0$ by Lemma A2 in Phillips and Hansen (1990). Consequently, $\mathbb{P}(\psi_{\min} > 0) \to 1$.

The second term of (A.5) is bounded by

$$\begin{aligned}(\hat{\gamma}-\gamma)'Q^{-1}S_TS_T^{-1}QV'M_D\epsilon_y &\leq \left\|S_TQ'^{-1}(\hat{\gamma}-\gamma)\right\|_2\left\|S_T^{-1}QV'M_D\epsilon_y\right\|_2 \\ &= \left\|S_TQ'^{-1}(\hat{\gamma}-\gamma)\right\|_2 b_T,\end{aligned}$$

where $b_T = O_p(1)$ according to Lemma A.3. The bounds for the right-hand side of (A.5) are the same as for the case $\delta = 0$, but with D_T replaced by $S_TQ'^{-1}$. In particular, we obtain

$$\lambda_{G,T}\left(\left\|\delta^s\right\|_2 - \left\|\hat{\delta}^s\right\|_2\right) \leq T^{-1/2}\lambda_{G,T}\left\|S_TQ'^{-1}(\gamma^s - \hat{\gamma}^s)\right\|_2,$$

and

$$\sum_{i=1}^{N} \lambda_{2,i} \left(|\delta_i^s| - \left|\hat{\delta}_i^s\right| \right) + \sum_{j=1}^{M} \lambda_{3,j} \left(|\pi_j^s| - |\hat{\pi}_j^s| \right) \leq \left\| S_T^{-1} Q \sigma_V \lambda_{S_\gamma} \right\|_2 \left\| S_T Q'^{-1} (\hat{\gamma}_i^s - \gamma_i^s) \right\|_2 .$$

Furthermore, we can bound

$$\left\| S_T^{-1} Q \sigma_V \lambda_{S_\gamma} \right\|_2 \leq T^{-1/2} \left\| \lambda_{S_\gamma} \right\|_2 \left\| \sigma_V \right\|_2 \left\| Q \right\|_2 ,$$

which is easily seen to be bounded in probability when $\frac{\lambda_{3,j} \sigma_{W,jj}}{\sqrt{T}} = o_p(1)$, for $j \in S_\pi$, and

$$\frac{\lambda_{2,i} \sigma_{Z,ii}}{\sqrt{T}} = \frac{\lambda_{\delta,T} \sigma_{Z,ii}}{\sqrt{T} \left| \hat{\delta}_{OLS,i}^{k_\delta} \right|} = o_p(1),$$

for $i \in S_\delta$. Since $\hat{\delta}_{OLS,i} \xrightarrow{p} \delta_i$ by the consistency of the OLS estimator, we require the additional condition $\frac{\lambda_{\delta,T} \sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$.

Combining the bounds for the case $\delta \neq 0$ we can rewrite (A.5) as

$$\begin{aligned} &\psi_{\min} \left\| S_T Q'^{-1} (\hat{\gamma} - \gamma) \right\|_2^2 - 2 b_T \left\| S_T Q'^{-1} (\hat{\gamma} - \gamma) \right\|_2 \\ &\leq \left(T^{-1/2} \lambda_G + \left\| S_T^{-1} Q \sigma_V \lambda_{S_\gamma} \right\|_2 \right) \left\| S_T Q'^{-1} (\hat{\gamma} - \gamma) \right\|_2 , \end{aligned}$$

which can be rewritten as

$$\left\| S_T Q'^{-1} (\hat{\gamma} - \gamma) \right\|_2 \leq \psi_{\min}^{-1} 2 b_T + \psi_{\min}^{-1} \left(T^{-1/2} \lambda_G + \left\| S_T^{-1} Q \sigma_V \lambda_{S_\gamma} \right\|_2 \right) = O_p(1),$$

thereby completing the proof for the case of cointegration. □

Proof of Theorem 2. We first proceed by deriving the selection consistency for the case $\delta = 0$. Assume that $\hat{\delta}_i^s \neq 0$ is a minimizer of (9) and thus, by application of Lemma A.4, also minimizes (A.4). Let z_i denote the i-th column vector of Z_{-1}. The first order conditions for $\hat{\delta}_i^s$ to be a minimum state

$$\left. \frac{d G_T(\gamma^s)}{d \delta_i^s} \right|_{\gamma^s = \hat{\gamma}^s} = \tilde{z}_i' M_D \left(\Delta y - \tilde{V} \hat{\gamma}^s \right) - \frac{\lambda_G}{2} \hat{\delta}_i^{\ s} \left\| \hat{\delta}^s \right\|_2^{-1} - \frac{\lambda_{2,i} \mathrm{sign}(\hat{\delta}_i^s)}{2} = 0.$$

After multiplying by $\frac{\sigma_{Z,ii}}{T}$ we get

$$\frac{z_i' M_D \left(\Delta y - Z \hat{\delta} - W \hat{\pi} \right)}{T} - \frac{\lambda_G \sigma_{Z,ii} \hat{\delta}_i^{\ s} \left\| \hat{\delta}^s \right\|_2^{-1}}{2T} - \frac{\lambda_{2,i} \sigma_{Z,ii} \mathrm{sign}(\hat{\delta}_i^s)}{2T} = 0 \qquad \text{(A.7)}$$

The first term can be rewritten as

$$\frac{z_i' M_D \left(\Delta y - Z \hat{\delta} - W \hat{\pi} \right)}{T} = \frac{z_i' M_D \left(\epsilon_y - V D_T^{-1} D_T (\hat{\gamma} - \gamma) \right)}{T} = O_p(1),$$

where the stochastic boundedness follows from the convergence in Lemma A.2 and the result that $D_T(\hat{\gamma} - \gamma) = O_p(1)$ under the assumptions in Theorem 1. Regarding the second term in

(A.7), note that $\hat{\delta}_i^s \left\|\hat{\delta}^s\right\|_2^{-1} = O_p(1)$, because all estimates share the same rate of convergence. Then,

$$\frac{\lambda_G \sigma_{Z,ii} \hat{\delta}_i^{\,s} \left\|\hat{\delta}^s\right\|_2^{-1}}{2T} \xrightarrow{p} 0,$$

since by our assumptions in Theorem 1, $\frac{\lambda_G \sigma_{Z,\max}}{\sqrt{T}} \to 0$. Finally, for the last term in (A.7) we obtain

$$\frac{\lambda_{2,i}\sigma_{z,ii}}{2T} = \frac{\lambda_{\delta,T}\sigma_{Z,ii}}{2T\left|\hat{\delta}_{OLS,i}\right|^{k_\delta}} = \frac{\lambda_{\delta,T}\sigma_{Z,ii}}{T^{1-k_\delta}} \frac{1}{2\left|T\hat{\delta}_{OLS,i}\right|^{k_\delta}} \to \infty$$

under the assumption that $\frac{\lambda_{\delta,T}\sigma_{Z,\min}}{T^{1-k_\delta}} \to \infty$. This implies that

$$\mathbb{P}(\hat{\delta}_i^s = 0) = 1 - \mathbb{P}(\hat{\delta}_i^s \neq 0) \geq 1 - \mathbb{P}\left(\frac{dG_T(\gamma^s)}{d\delta_i^s}\Big|_{\gamma^s=\hat{\gamma}^s} = 0\right) \to 1. \tag{A.8}$$

Then, by noting that $\mathbb{P}(\hat{\delta}_i^s = 0) = \mathbb{P}(\hat{\delta}_i = 0)$, the selection consistency for $\hat{\delta}_i$ absent of cointegration follows.

Next, assume $\hat{\pi}_j^s \neq 0$ while $\pi_j = 0$ and let w_j be the j-th column of W. For $\hat{\pi}_j^s$ to be a minimum of (A.4) the first order conditions, after appropriate scaling, state

$$\frac{w_j' M_D\left(\Delta y - V\hat{\gamma}\right)}{\sqrt{T}} - \frac{\lambda_{3,j}\sigma_{W,jj}\operatorname{sign}(\hat{\pi}_j^s)}{2\sqrt{T}} = 0. \tag{A.9}$$

The first term can be rewritten as

$$\frac{w_j' M_D\left(\Delta y - V\hat{\gamma}\right)}{\sqrt{T}} = \frac{w_j' M_D\left(\epsilon_y - VD_T^{-1}D_T(\hat{\gamma} - \gamma)\right)}{\sqrt{T}} = O_p(1),$$

where the stochastic boundedness follows from the Lemma A.2 and $D_T(\hat{\gamma} - \gamma) = O_p(1)$ by Theorem 1. For the second term in (A.9) we have

$$\frac{\lambda_{3,j}\sigma_{W,jj}}{2\sqrt{T}} = \frac{\lambda_{\pi,T}\sigma_{W,jj}}{2\sqrt{T}\left|\hat{\pi}_{OLS,j}\right|^{k_\pi}} = \frac{\lambda_{\pi,T}\sigma_{W,jj}}{T^{1/2-k_\pi/2}} \frac{1}{2\left|\sqrt{T}\hat{\pi}_{OLS,j}\right|^{k_\pi}} \to \infty \tag{A.10}$$

under the assumption that $\frac{\lambda_{\pi,T}\sigma_{W,\min}}{T^{1/2-k_\pi/2}} \to \infty$. The selection consistency for $\hat{\pi}_j$ then follows by the same argument used in (A.8).

The strategy for showing selection consistency in the presence of cointegration is analogous, although algebraically slightly more tedious. Let $\delta_i^s = \gamma_i^s = 0$. Then the first order condition for $\hat{\delta}_i^s \neq 0$ to be a minimum of the objective function, after pre-multiplying by $\frac{\sigma_{Z,ii}}{T}$, are again given by (A.7). Letting e_i denote the i-th column of I_{N+M}, the first term can be rewritten as

$$\begin{aligned}\frac{z_i' M_D\left(\epsilon_y - V(\hat{\gamma} - \gamma)\right)}{T} &= \frac{e_i' V' M_D\left(\epsilon_y - V(\hat{\gamma} - \gamma)\right)}{T} \\ &= \frac{e_i' Q^{-1} S_T}{T} S_T^{-1} Q V' M_D\left(\epsilon_y - VQ' S_T^{-1} S_T Q'^{-1}(\hat{\gamma} - \gamma)\right) = O_p(1),\end{aligned}$$

because $\frac{e_i' Q^{-1} S_T}{T} = O(1)$, $S_T^{-1} Q V' M_D \epsilon_y = O_p(1)$ and $S_T^{-1} Q V' M_D V Q' S_T^{-1} = O_p(1)$ by Lemma A.3, and $S_T Q'^{-1}(\hat{\gamma} - \gamma) = O_p(1)$ by Theorem 1. The second term in (A.7) again converges to zero in probability and for the third and final term we obtain

$$\frac{\lambda_{2,i} \sigma_{Z,ii}}{2T} = \frac{\lambda_{\delta,T} \sigma_{Z,ii}}{2T \left| \hat{\delta}_{OLS,i} \right|^{k_\delta}} = \frac{\lambda_{\delta,T} \sigma_{Z,ii}}{T^{1-k_\delta/2}} \frac{1}{2 \left| \sqrt{T} \hat{\delta}_{OLS,i} \right|^{k_\delta}} \to \infty,$$

under the assumption that $\frac{\lambda_{\delta,T} \sigma_{Z,ii}}{T^{1-k_\delta/2}} \to \infty$. Then, by the same argument as in (A.8) we can conclude that $\mathbb{P}(\hat{\delta}_i = 0) \to 1$.

Similarly, letting $\pi_j = 0$, the first order conditions for $\hat{\pi}_j \neq 0$ to be a minimum of (A.4) when $\pi_j = 0$ are again given by (A.9). The first term can be rewritten as

$$\frac{w_j' M_D \left(\epsilon_y - V(\hat{\gamma} - \gamma) \right)}{\sqrt{T}} = \frac{w_j' M_D \left(\epsilon_y - V Q' S_T^{-1} S_T Q'^{-1}(\hat{\gamma} - \gamma) \right)}{\sqrt{T}} = O_p(1),$$

because $\frac{w_j' M_D \epsilon_y}{\sqrt{T}} = O_p(1)$ by Lemma A.2,

$$\frac{w_j' M_D V Q' S_T^{-1}}{\sqrt{T}} = \left[T^{-1} w_j' Z_{-1} \beta \quad T^{-1} w_j' W \quad T^{-3/2} w_j' Z_{-1} \alpha_\perp \right] = O_p(1),$$

by Lemma A.3 and $S_T Q'^{-1}(\hat{\gamma} - \gamma) = O_p(1)$ by Theorem 1. Furthermore, we again have that $\frac{\lambda_{3,j} \sigma_{W,jj}}{2\sqrt{T}} \to \infty$ based on (A.10). Consequently, it follows that $\mathbb{P}(\hat{\pi}_j = 0) \to 1$ by the same argument used for (A.8), thus completing the proof. □

Proof of Theorem 3. Without loss of generality we impose an ordering on the variables such that $V = (V_{S_\gamma}, V_{S_\gamma^c}) = (Z_{S_\delta}, W_{S_\pi}, Z_{S_\delta^c}, W_{S_\pi^c})$, where the variables collected in V_{S_γ} carry non-zero coefficients in the true DGP, whereas $V_{S_\gamma^c}$ contains all irrelevant variables. The de-standardized estimate $\hat{\gamma}_{S_\gamma}$ is defined as $\sigma_{V_{S_\gamma}}^{-1} \hat{\gamma}_{S_\gamma}^s$. Since $\hat{\gamma}^s$ are the minimizers of (A.4), they must set the subgradient equations equal to zero:

$$\tilde{V}' M_D (\Delta Y - \tilde{V} \hat{\gamma}^s) - \frac{1}{2} \hat{s}\left(\hat{\gamma}^s \right) = 0,$$

or after pre-multiplication with σ_V by

$$V' M_D (\Delta Y - V \hat{\gamma}) - \frac{1}{2} \sigma_V \hat{s}\left(\hat{\gamma}^s \right) = 0, \tag{A.11}$$

where we let $\hat{s}\left(\hat{\gamma}^s \right)$ denote the sub-gradient of the penalty function $P_\lambda(\hat{\gamma}^s)$. In particular, define $\Lambda = \mathrm{diag}(\lambda_2, \lambda_3)$, then

$$\hat{s}\left(\hat{\gamma}^s \right) = \lambda_G \hat{s}_G(\hat{\delta}^s) + \Lambda \hat{s}_I\left(\hat{\gamma}^s \right),$$

where $\hat{s}_G(\hat{\delta}^s)$ is a $(N+M)$-dimensional vector with the first N elements being given by $\hat{\delta} / \left\| \hat{\delta} \right\|_2$, whenever at least one of the $\hat{\delta}_j \neq 0$, or by a N-dimensional vector x with $\|x\|_2 \leq 1$ otherwise, and the remaining M elements of $\hat{s}_G(\hat{\delta}^s)$ are equal to zero. Furthermore, $\hat{s}_I\left(\hat{\gamma}^s \right)$, has element j equal to $\mathrm{sign}(\hat{\gamma}_j^s)$ when $\hat{\gamma}_j^s \neq 0$ and can be any scalar $x \in [-1, 1]$ otherwise. Below we will additionally refer to the vector

$$\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right) = \lambda_G \hat{s}_G(\hat{\gamma}^s_{S_\gamma}) + \Lambda_{S_\gamma}\hat{s}_I\left(\hat{\gamma}^s_{S_\gamma}\right),$$

which is the sub-gradient of the penalty function for the coefficients indexed by S_γ. Important to note is that given our assumptions on the penalty terms, i.e. $\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$, $\frac{\lambda_{\delta,T}\sigma_{Z,\max}}{\sqrt{T}} \to 0$ and $\frac{\lambda_{\pi,T}\sigma_{W,\max}}{\sqrt{T}} \to 0$, it immediately follows that

$$T^{-1/2}\sigma_{V,S_\gamma}\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right) \to 0. \tag{A.12}$$

We proceed by rewriting the first order conditions (A.11) in terms of $\hat{\gamma}_{S_\gamma}$ as

$$\begin{aligned}
0 =& V'_{S_\gamma}M_D\left(\Delta Y - V_{S_\gamma}\hat{\gamma}_{S_\gamma} - V_{S^c_\gamma}\hat{\gamma}_{S^c_\gamma}\right) - \frac{1}{2}\sigma_{V,S_\gamma}\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right)\\
=& V'_{S_\gamma}M_D\left(\hat{\epsilon}_{OLS} - V_{S_\gamma}\left(\hat{\gamma}_{S_\gamma} - \hat{\gamma}_{OLS,S_\gamma}\right) - V_{S^c_\gamma}\hat{\gamma}_{S^c_\gamma}\right) - \frac{1}{2}\sigma_{V,S_\gamma}\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right)\\
=& - V'_{S_\gamma}M_D\left(V_{S_\gamma}\left(\hat{\gamma}_{S_\gamma} - \hat{\gamma}_{OLS,S_\gamma}\right) + V_{S^c_\gamma}\hat{\gamma}_{S^c_\gamma}\right) - \frac{1}{2}\sigma_{V,S_\gamma}\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right),
\end{aligned} \tag{A.13}$$

where $\hat{\epsilon}_{OLS} = M_D(\Delta y - V_{S_\gamma}\hat{\gamma}_{OLS,S_\gamma})$ such that $V'_{S_\gamma}M_D\hat{\epsilon}_{OLS} = 0$ by construction. Reordering terms in (A.13) gives

$$\hat{\gamma}_{S_\gamma} - \hat{\gamma}_{OLS,S_\gamma} = \left(V'_{S_\gamma}M_DV_{S_\gamma}\right)^{-1}V'_{S_\gamma}M_DV_{S^c_\gamma}\hat{\gamma}_{S^c_\gamma} - \frac{1}{2}\left(V'_{S_\gamma}M_DV_{S_\gamma}\right)^{-1}\sigma_{V,S_\gamma}\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right). \tag{A.14}$$

We now separately consider the cases without and with cointegration in the underlying DGP. Absent of cointegration we have $V_{S_\gamma} = W_{S_\pi}$ and $\gamma_{S_\gamma} = \pi_{S_\pi}$ such that after appropriately scaling (A.14) we obtain

$$\begin{aligned}
\sqrt{T}\left(\hat{\pi}_{S_\pi} - \hat{\pi}_{OLS,S_\pi}\right) =& -\left(T^{-1}W'_{S_\pi}M_DW_{S_\pi}\right)^{-1}T^{-1/2}W'_{S_\pi}M_DV_{S^c_\gamma}\hat{\gamma}_{S^c_\gamma}\\
&- \frac{1}{2}\left(T^{-1}W'_{S_\pi}M_DW_{S_\pi}\right)^{-1}T^{-1/2}\sigma_{W,S_\pi}\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right) = o_p(1),
\end{aligned}$$

where the stated convergence follows, because $\mathbb{P}(\hat{\gamma}_{S^c_\gamma} = 0) \to 1$ such that

$$\left(T^{-1}W'_{S_\pi}M_DW_{S_\pi}\right)^{-1}T^{-1/2}W'_{S_\pi}M_DV_{S^c_\gamma}\hat{\gamma}_{S^c_\gamma}$$

vanishes in probability and

$$\frac{1}{2}\left(T^{-1}W'_{S_\pi}M_DW_{S_\pi}\right)^{-1}T^{-1/2}\sigma_{W,S_\pi}\hat{s}\left(\hat{\gamma}^s_{S_\gamma}\right) = o_p(1)$$

by Lemma A.2 and (A.12). Alternatively, when cointegration is present in the data we make use of S_{T,S_γ} and Q_{S_γ} as defined in Theorem 3. Observe that

$$Q_{S_\gamma}v_{S_\gamma,t} = \begin{bmatrix}\beta'_{S_\delta}z_{S_\delta,t-1}\\ w_{S_\pi,t}\\ \beta'_{S_\delta,\perp}z_{S_\delta,t-1}\end{bmatrix} = \begin{bmatrix}v_{S_{\gamma_1},t}\\ v_{S_{\gamma_2},t}\end{bmatrix},$$

where $v_{S_{\gamma_1},t} = (z'_{S_\delta,t-1}\beta_{S_\delta}, w'_t)' \sim I(0)$ and $v_{S_{\gamma_2},t} = z_{S_\delta,t-1}\beta_{\perp,S_\delta} \sim I(1)$. In matrix form, we write

$$V_{s_\gamma} Q'_{S_\gamma} = \begin{bmatrix} V_{S_\gamma,1} & V_{S_\gamma,2} \end{bmatrix},$$

with $V_{S_\gamma,1} = \begin{bmatrix} Z_{-1,S_\delta}\beta_{S_\delta} & W_{S_\pi} \end{bmatrix}$ and $V_{S_\gamma,2} = Z_{-1,S_\delta}\beta_{S_\delta,\perp}$. By a straightforward adaptation[15] of Lemma 3, it then follows that

$$S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S_\gamma} Q'_{S_\gamma} S^{-1}_{T,S_\gamma} \xrightarrow{d} \begin{bmatrix} \Sigma_{V_{S_\gamma,1}} & 0 \\ 0 & \beta'_{S_\delta,\perp} C_{S_\delta} \left(\int_0^1 B^*_{S_\delta}(r) B^*_{S_\delta}(r)' \right) C'_{S_\delta} \beta_{S_\delta,\perp} \end{bmatrix}, \tag{A.15}$$

where

$$\Sigma_{V_{S_\gamma,1}} = \begin{bmatrix} \mathbb{E}\left(\beta'_{S_\delta} u_{S_\delta,t} u'_{S_\delta,t} \beta_{S_\delta} \right) & 0 \\ 0 & \mathbb{E}\left(w_{S_\pi,t} w'_{S_\pi,t} \right) \end{bmatrix},$$

and $C_{S_\delta} = \beta_{\perp,S_\delta} \left(\alpha'_\perp \beta_\perp \right)^{-1}$. Then, it follows that

$$\begin{aligned} S_{T,S_\gamma} Q'^{-1}_{S_\gamma} \left(\hat{\gamma}_{S_\gamma} - \hat{\gamma}_{OLS,S_\gamma} \right) = & - \left(S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S_\gamma} Q'_{S_\gamma} S^{-1}_{T,S_\gamma} \right)^{-1} S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S^c_\gamma} \hat{\gamma}_{S^c_\gamma} \\ & - \frac{1}{2} \left(S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S_\gamma} Q'_{S_\gamma} S^{-1}_{T,S_\gamma} \right)^{-1} S^{-1}_{T,S_\gamma} Q_{S_\gamma} \sigma_{V,S_\gamma} \hat{s} \left(\hat{\gamma}^s_{S_\gamma} \right) \\ = & \; o_p(1), \end{aligned}$$

where the convergence follows because

$$\left(S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S_\gamma} Q'_{S_\gamma} S^{-1}_{T,S_\gamma} \right)^{-1} S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S^c_\gamma} \hat{\gamma}_{S^c_\gamma}$$

vanishes in probability since $\mathbb{P}(\hat{\gamma}_{S^c_\gamma} = 0) \to 1$ and

$$\frac{1}{2} \left(S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S_\gamma} Q'_{S_\gamma} S^{-1}_{T,S_\gamma} \right)^{-1} S^{-1}_{T,S_\gamma} Q_{S_\gamma} \sigma_{V,S_\gamma} \hat{s} \left(\hat{\gamma}^s_{S_\gamma} \right) = o_p(1)$$

because of (A.15) and (A.12). This completes the proof. □

Proof of Corollary 1. We first show that

$$S_{T,S_\gamma} Q'^{-1}_{S_\gamma} (\hat{\gamma}_{S_\gamma,OLS} - \gamma_{S_\gamma}) \xrightarrow{d} \begin{bmatrix} \mathcal{N}\left(0, \sigma^2_{\epsilon_y} \Sigma^{-1}_{V_{S_\gamma,1}}\right) \\ \left(\beta'_{S_\delta,\perp} C_{S_\delta} \left(\int_0^1 B^*_{S_\delta}(r) B^*_{S_\delta}(r)' \right) C'_{S_\delta} \beta_{S_\delta,\perp} \right)^{-1} \beta'_{S_\delta,\perp} C_{S_\delta} \left(\int_0^1 B^*_{S_\delta}(r) dB_\epsilon(r) \right) \end{bmatrix}, \tag{A.16}$$

where $\sigma^2_{\epsilon_y} = \mathbb{E}(\epsilon^2_{y,t})$. Note that

$$\begin{aligned} S_{T,S_\gamma} Q'^{-1}_{S_\gamma} (\hat{\gamma}_{S_\gamma,OLS} - \gamma_{S_\gamma}) &= S_{TS_\gamma} Q'^{-1}_{S_\gamma} \left(V'_{S_\gamma} M_D V_{S_\gamma} \right)^{-1} V'_{S_\gamma} M_D \epsilon_y \\ &= \left(S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D V_{S_\gamma} Q' S^{-1}_{T,S_\gamma} \right)^{-1} S^{-1}_{T,S_\gamma} Q_{S_\gamma} V'_{S_\gamma} M_D \epsilon_y. \end{aligned}$$

[15] The adaptation of Lemma A.3 follows from replacing $\alpha_\perp$, Σ_{V_1} and C with $\beta_{S_\delta,\perp}$, $\Sigma_{V_{S_\gamma,1}}$ and C_{S_δ}, respectively.

By a straightforward adaptation of Lemma A.3 it follows that

$$S_{T,S_\gamma}^{-1} Q_{S_\gamma} V_{S_\gamma}' M_D \epsilon_y \xrightarrow{d} \begin{bmatrix} N\left(0, \sigma_{\epsilon_y}^2 \Sigma_{V_{S_\gamma},1}\right) \\ \beta_{S_\delta,\perp}' C_{S_\delta} \left(\int_0^1 B_{S_\delta}^*(r) dB_\epsilon(r)\right) \end{bmatrix},$$

such that by (A.15) in combination with the continuous mapping theorem for functionals and Slutsky's theorem, the result in (A.16) follows.

As a direct consequence, we have

$$T^{1/2} Q_{S_\gamma}'^{-1} \left(\hat{\gamma}_{S_\gamma} - \gamma_{S_\gamma}\right) \xrightarrow{d} \mathscr{N}\left(0, \sigma_{\epsilon_y}^2 \begin{bmatrix} \Sigma_{V_{S_\gamma},1}^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right),$$

such that

$$\sqrt{T}\left(\hat{\gamma}_{S_\gamma} - \gamma_{S_\gamma}\right) \xrightarrow{d} \mathscr{N}\left(0, \sigma_{\epsilon_y}^2 Q_{S_\gamma}' \begin{bmatrix} \Sigma_{V_{S_\gamma},1}^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q_{S_\gamma}\right) = \mathscr{N}\left(0, \sigma_{\epsilon_y}^2 \begin{bmatrix} \beta_{S_\delta} \Sigma_U^{-1} \beta_{S_\delta}' & 0 \\ 0 & \Sigma_{W_{S_\pi}}^{-1} \end{bmatrix}\right),$$

with Σ_U and $\Sigma_{W_{S_\pi}}$ as defined in Corollary 1. This proves the part of Corollary 1 on the convergence of the estimator.

We proceed by showing that the matrix $\beta_{S_\delta} \Sigma_U^{-1} \beta_{S_\delta}'$ is uniquely defined, regardless of the choice of the basis matrix β_{S_δ}. Naturally, the basis matrix β_{S_δ} itself is not unique, as any matrix whose columns form a basis for the left nullspace of $\beta_{\perp,S_\delta}$ may be used in the construction of Q_{S_γ}. Accordingly, assume that another matrix satisfying this condition is given by $\beta_{S_\delta}^*$ with the i-th column vector given by $\beta_{S_\delta,i}^* = \beta_{S_\delta} x_i$, where x_i are the coordinates of $\beta_{S_\delta,i}^*$ with respect to the basis β_{S_δ}. Then, we can represent our new basis as

$$\beta_{S_\delta}^* = \beta_{S_\delta} X,$$

where $X = \begin{bmatrix} x_1 & \dots & x_{r_2} \end{bmatrix}$. Moreover, X must be linearly independent, because otherwise there exists a $u \in R^{r_2}$ with $u \neq 0$ and

$$\beta_{S_\delta}^* u = \beta_{S_\delta} X u = 0,$$

thereby contradicting the claim that $\beta_{S_\delta}^*$ is a basis matrix. Consequently, X is an invertible linear transformation and it follows that

$$\begin{aligned} \beta_{S_\delta}^* \Sigma_U^{*-1} \beta_{S_\delta}^{*\prime} &= \beta_{S_\delta}^* \left(\mathbb{E}\left(\beta_{S_\delta}^{*\prime} u_{S_\delta,t} u_{S_\delta,t}' \beta_{S_\delta}^*\right)\right)^{-1} \beta_{S_\delta}^{*\prime} = \beta_{S_\delta} X \left(\mathbb{E}\left(X' \beta_{S_\delta}' u_{S_\delta,t} u_{S_\delta,t}' \beta_{S_\delta} X\right)\right)^{-1} X' \beta_{S_\delta}' \\ &= \beta_{S_\delta} \left(\mathbb{E}\left(\beta_{S_\delta}' u_{S_\delta,t} u_{S_\delta,t}' \beta_{S_\delta}\right)\right)^{-1} \beta_{S_\delta}' = \beta_{S_\delta} \Sigma_U^{-1} \beta_{S_\delta}', \end{aligned}$$

thereby validating the claim that $\beta_{S_\delta} \Sigma_U^{-1} \beta_{S_\delta}'$ is uniquely defined regardless of the choice of basis. □

Appendix B Supplementary Material

B.1 Proof of Corollary 2

Proof of Corollary 2. The proof of the consistency of the estimated deterministic components is straightforward, though algebraically tedious. Recall that $\theta = (\mu_0, \tau_0)'$. Based on Lemma A.4 it follows that

$$
\begin{aligned}
\hat{\theta} &= \left(D'D\right)^{-1} D' \left(\Delta y - V\hat{\gamma}\right) = \left(D'D\right)^{-1} D' \left(\hat{\epsilon}_{y,OLS} - V\left(\hat{\gamma} - \hat{\gamma}_{OLS}\right) + D\hat{\theta}_{OLS}\right) \\
&= \hat{\theta}_{OLS} - \left(D'D\right)^{-1} D'V\left(\hat{\gamma} - \hat{\gamma}_{OLS}\right),
\end{aligned}
$$

such that

$$
\hat{\theta} - \hat{\theta}_{OLS} = -\left(D'D\right)^{-1} D'V\left(\hat{\gamma} - \hat{\gamma}_{OLS}\right). \tag{B.1}
$$

Note that

$$
(D'D)^{-1} = \frac{1}{|D'D|} \begin{bmatrix} \bar{t}'\bar{t} & -\iota'\bar{t} \\ -\iota'\bar{t} & T \end{bmatrix},
$$

where

$$
\left|D'D\right| = T\bar{t}'\bar{t} - (\iota'\bar{t})^2 = O(T^4).
$$

The analytical expression for the constant can be derived from (B.1). Assuming for the moment that $\mu \neq 0$, $\tau \neq 0$ and $\delta = 0$, we obtain

$$
\begin{aligned}
\hat{\mu}_0 - \hat{\mu}_{0,OLS} &= \frac{1}{|D'D|} \left[\left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right) V\right] \left[\hat{\gamma} - \hat{\gamma}_{OLS}\right] \\
&= \frac{1}{|D'D|} \begin{bmatrix} \left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right) Z_{-1} & \left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right) W \end{bmatrix} \begin{bmatrix} \hat{\delta} - \hat{\delta}_{OLS} \\ \hat{\pi} - \hat{\pi}_{OLS} \end{bmatrix} \\
&= O(T^{-4}) \begin{bmatrix} O_p(T^{9/2}) & O_p(T^4) \end{bmatrix} \begin{bmatrix} o_p(T^{-1}) \\ o_p(T^{-1/2}) \end{bmatrix} = o_p(T^{-1/2}).
\end{aligned} \tag{B.2}
$$

This may be verified by writing out each term and applying Lemma A.2. We demonstrate this for this particular instance. Note that

$$
\begin{aligned}
\left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right) Z_{-1} &= \left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right)\left(S_{-1}C' + \iota\mu' + \bar{t}\tau' + U_{-1}\right) \\
&= \left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right) S_{-1}C' + \left(T\bar{t}'\bar{t} - (\iota'\bar{t})^2\right)\mu' + \left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right) U_{-1} \\
&= O_p(T^{9/2}) + O(T^4) + O_p(T^{7/2}).
\end{aligned}
$$

Hence, regardless of whether $\mu \neq 0$ or $\tau \neq 0$, it holds that $\left(\bar{t}'\bar{t}\iota' - \iota'\bar{t}\bar{t}'\right) Z_{-1} = O_p(T^{9/2})$. Similarly, for the term in (B.2) involving W, we note that

$$
W = \begin{bmatrix} \Delta X & \Delta Z_{-1} & \dots & \Delta Z_{-p} \end{bmatrix} = \begin{bmatrix} \Delta Z & \dots & \Delta Z_{-p} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{1\times((P+1)N-1)} \\ I_{(P+1)N-1} \end{bmatrix},
$$

where $\Delta Z_{-j} = \iota\tau' + U_{-j}$ with

$$U'_{-j} = (C + C(L)(1-L)) \begin{bmatrix} \mathbf{0}_{N\times j} & \epsilon_1 & \dots & \epsilon_{T-j} \end{bmatrix}.$$

Then, since

$$\left(\bar{t}'\bar{t}\iota' - \iota'\overline{tt}'\right) \Delta Z_{-j} = \left(\bar{t}'\bar{t}\iota' - \iota'\overline{tt}'\right) \iota\tau' + \left(\bar{t}'\bar{t}\iota' - \iota'\overline{tt}'\right) U_j = O(T^4) + O_p(T^{7/2}),$$

it follows that $W = O_p(T^4)$ when $\tau \neq 0$ and $W = O_p(T^{7/2})$ when $\tau = 0$. However, when $\tau = 0$ the rate of $\hat{\mu}_0$ will be determined by the term in (B.2) involving Z_{-1} and the convergence rate is thus invariant to the presence of a constant or deterministic trend.

In the remainder of the proof we proceed along a similar strategy by deriving the stochastic order for varying δ, τ and μ. However, for the sake of brevity, we refrain from writing out each individual term and rather refer to each term's stochastic order directly. We start by deriving a similar result to (B.2), but for the case $\delta \neq 0$. Then, (B.1) can be written as

$$\begin{aligned} \hat{\theta} - \hat{\theta}_{OLS} &= -\left(D'D\right)^{-1} D'VQ'Q'^{-1}\left(\hat{\gamma} - \hat{\gamma}_{OLS}\right) \\ &= -\left(D'D\right)^{-1} D' \begin{bmatrix} Z_{-1}\beta & W & Z_{-1}\alpha_{\perp} \end{bmatrix} \begin{bmatrix} (\alpha'\beta)^{-1}\alpha'(\hat{\delta} - \hat{\delta}_{OLS}) \\ \hat{\pi} - \hat{\pi}_{OLS} \\ (\beta'_{\perp}\alpha_{\perp})^{-1}\beta'_{\perp}(\hat{\delta} - \hat{\delta}_{OLS}) \end{bmatrix}, \end{aligned} \tag{B.3}$$

from which follows that,

$$\begin{aligned} \hat{\mu}_0 - \hat{\mu}_{0,OLS} =& \frac{1}{|D'D|} \begin{bmatrix} (\bar{t}'\bar{t}\iota' - \iota'\overline{tt}')Z_{-1}\beta & (\bar{t}'\bar{t}\iota' - \iota'\overline{tt}')W & (\bar{t}'\bar{t}\iota' - \iota'\overline{tt}')Z_{-1}\alpha_{\perp} \end{bmatrix} \\ & \times \begin{bmatrix} (\alpha'\beta)^{-1}\alpha'(\hat{\delta} - \hat{\delta}_{OLS}) \\ \hat{\pi} - \hat{\pi}_{OLS} \\ (\beta'_{\perp}\alpha_{\perp})^{-1}\beta'_{\perp}(\hat{\delta} - \hat{\delta}_{OLS}) \end{bmatrix} \\ =& O(T^{-4}) \begin{bmatrix} O_p(T^4) & O_p(T^4) & O_p(T^{9/2}) \end{bmatrix} \begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-1/2}). \end{aligned}$$

Again, one may verify that the rate of convergence holds irrespective of whether $\mu = 0$ or $\tau = 0$.

Next, we move on to the expression for the trend coefficient. For the cases with $\beta = 0$, we will rely on the expression

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = \frac{1}{|D'D|} \begin{bmatrix} (T\bar{t}' - \iota'\bar{t}\iota')Z_{-1} & (T\bar{t}' - \iota'\bar{t}\iota')W \end{bmatrix} \begin{bmatrix} \hat{\delta} - \hat{\delta}_{OLS} \\ \hat{\pi} - \hat{\pi}_{OLS} \end{bmatrix}, \tag{B.4}$$

whereas for $\beta \neq 0$ we will use the equivalent expression

$$\begin{aligned} \hat{\tau}_0 - \hat{\tau}_{0,OLS} =& \frac{1}{|D'D|} \begin{bmatrix} (T\bar{t}' - \iota'\bar{t}\iota')Z_{-1}\beta & (T\bar{t}' - \iota'\bar{t}\iota')W & (T\bar{t}' - \iota'\bar{t}\iota')Z_{-1}\alpha_{\perp} \end{bmatrix} \\ & \times \begin{bmatrix} (\alpha'\beta)^{-1}\alpha'(\hat{\delta} - \hat{\delta}_{OLS}) \\ \hat{\pi} - \hat{\pi}_{OLS} \\ (\beta'_{\perp}\alpha_{\perp})^{-1}\beta'_{\perp}(\hat{\delta} - \hat{\delta}_{OLS}) \end{bmatrix}. \end{aligned} \tag{B.5}$$

Then, for the case $\tau = 0$ and $\beta = 0$, (B.4) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^{7/2}) & O_p(T^{5/2}) \end{bmatrix} \begin{bmatrix} o_p(T^{-1}) \\ o_p(T^{-1/2}) \end{bmatrix} = o_p(T^{-3/2}).$$

For the case $\tau = 0$ and $\beta \neq 0$, (B.5) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^{3}) & O_p(T^{5/2}) & O_p(T^{7/2}) \end{bmatrix} \begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-3/2}).$$

Next, assuming that $\tau \neq 0$ and $\beta = 0$, it follows from (B.4) that

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^{4}) & O_p(T^{3}) \end{bmatrix} \begin{bmatrix} o_p(T^{-1}) \\ o_p(T^{-1/2}) \end{bmatrix} = o_p(T^{-1}).$$

Alternatively, if $\tau \neq 0$, $\beta \neq 0$ and $\beta'\tau = 0$, then (B.5) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^{3}) & O_p(T^{3}) & O_p(T^{4}) \end{bmatrix} \begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-1}).$$

Finally, assume that $\tau \neq 0$, and $\beta'\tau \neq 0$. Then, (B.5) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^{4}) & O_p(T^{3}) & O_p(T^{4}) \end{bmatrix} \begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-1/2}).$$

This completes the proof of Corollary 2. □

B.2 Data Description

Variable	groups	Translation	Inclusion	Differenced
vakantiebaan	Job Search	holiday job	100%	N
Unemployment	Y	Unemployment	80%	Y
uwv vacatures	Job Search	uwv vacancies	78%	Y
werkloos	Social Security	unemployed	76%	Y
ww uitkering	Social Security	ww benefits	72%	Y
Ww	Social Security	Ww	69%	Y
nationale vacaturebank	RA	nationale vacaturebank	59%	Y
cv maken	Application training	CV write	57%	Y
indeed	RA	indeed	52%	Y
jobtrack	RA	jobtrack	52%	Y
motivatiebrief	Application training	motivation letter	52%	Y
sollicitatiebrief schrijven	Application training	write application letter	50%	Y
voorbeeld cv	Application training	example cv	48%	Y
tempo team	RA	tempo team	48%	Y

ontslagvergoeding	Social Security	severance pay	46%	Y
ww uitkering aanvragen	Social Security	request unemployment benefits	46%	Y
aanvragen uitkering	Social Security	request benefits	44%	N
interin	RA	interin	44%	Y
manpower	RA	manpower	44%	Y
randstad	General	randstad (geographical area)	44%	Y
werkzoekende	Social Security	job seeker	43%	Y
job	General	job	43%	Y
uwv	Social Security	uwv	43%	Y
werk.nl	Job Search	werk.nl	41%	Y
job vacancy	Job Search	job vacancy	41%	Y
uitkering	Social Security	benefits	41%	Y
ontslag	Social Security	resignation	41%	N
vacature	Job Search	vacancy	41%	Y
sollicitatiebrief voorbeeld	Application training	application letter example	39%	Y
sollicitatie	Application training	application	39%	Y
sollicitatiebrief	Application training	application letter	39%	Y
uitzendbureau	RA	employment agency	39%	Y
vakantiewerk	Job Search	holiday job	37%	N
tence	RA	tence	37%	Y
vacaturebank	Job Search	vacaturebank	37%	Y
sollicitatiegesprek	Application training	application interview	37%	N
tempo team uitzendbureau	RA	tempo team employment agency	35%	N
motivatiebrief voorbeeld	Application training	motivation letter example	35%	Y
bijstand	Social Security	social benefits	35%	Y
open sollicitatiebrief	Application training	open application letter	35%	Y
vrijwilligerswerk	General	volunteer work	35%	N
werk nl	Job Search	werk nl	35%	N
adecco	RA	adecco	33%	N
creyfs	RA	creyfs	33%	Y
randstad uitzendbureau	Job Search	randstad employment agency	33%	Y
cv maken voorbeeld	Application training	write CV example	31%	Y
werkbedrijf	Job Search	werkbedrijf	31%	Y
tempo-team	RA	tempo-team	31%	Y
werkloosheidsuitkering	Social Security	unemployment benefits	31%	N
tempo team vacatures	RA	tempo team vacancies	31%	Y
curriculum vitae voorbeeld	Application training	CV Example	31%	Y
cv	Application training	cv	31%	N
solliciteren	Application training	applying	31%	Y
indeed jobs	RA	indeed jobs	30%	Y
motivation letter	Application training	motivation letter	30%	N
resume example	Application training	resume example	28%	N
olympia uitzendbureau	RA	olympia employment agency	28%	Y
tempoteam	RA	tempoteam	28%	Y
randstad vacatures	Job Search	randstad vacancies	26%	Y
banen	General	jobs	26%	N
vrijwilliger	General	volunteer	26%	N
baan	General	job	26%	N

start uitzendbureau	RA	start employment agency	24%	Y
jobnet	RA	jobnet	24%	N
monsterboard	Job Search	monsterboard	24%	Y
baan zoeken	Job Search	job search	20%	N
functieomschrijving	General	job position description	20%	N
resume template	Application training	resume template	19%	N
omscholen	Application training	retraining	19%	Y
job interview	Application training	job interview	19%	N
werken bij	General	working at	19%	Y
vacatures	Job Search	vacancies	19%	Y
uwv uitkering	Social Security	uwv benefits	17%	Y
job description	General	job description	17%	Y
werk zoeken	General	job search	17%	Y
jobs	General	jobs	17%	Y
resum	Application training	resume	15%	Y
bijscholen	Application training	retraining	15%	N
curriculum vitae template	Application training	CV Template	13%	N
curriculum vitae	Application training	CV	11%	Y
sollicitaties	Application training	applications	9%	Y
werkeloos	Social Security	unemployed	9%	N
werkloosheid	Social Security	unemployment	4%	N
resume	Application training	resume	2%	N
arbeidsbureau	RA	employment office	2%	N
uitzendbureaus	RA	employment agencies	2%	Y
werkloosheidswet	Social Security	unemployment law	0%	N

References

Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra.* Cambridge University Press.

Ahn, S. K. and Reinsel, G. C. (1990). Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association*, 85:813–823.

Banerjee, A., Dolado, J., and Mestre, R. (1998). Error-correction mechanism tests for cointegration in a single-equation framework. *Journal of Time Series Analysis*, 19:267–283.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41:802–837.

Boswijk, H. P. (1994). Testing for an unstable root in conditional and structural error correction models. *Journal of Econometrics*, 63:37–60.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer-Verlag, New York, 2nd edition.

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

Choi, H. and Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88:2–9.

Chou, W., Denis, K. F., and Lee, C. F. (1996). Hedging with the nikkei index futures: The convential model versus the error correction model. *The Quarterly Review of Economics and Finance*, 36:495–505.

Davidson, J. (2000). *Econometric Theory.* Blackwell Publishers, Oxford, 2nd edition.

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

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

Fan, J. and Li, R. (2001). Variable selection via nonconcace penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55:665–676.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements Of Statistical Learning.* Springer.

Hyndman, R. J. (2016). *forecast: Forecasting functions for time series and linear models.* R package version 7.2.

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

Johansen, S. (1995). *Likelihood-based Inference In Cointegrated Vector Autoregressive Models.* Oxford University Press.

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

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44:907–927.

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

Liang, C. and Schienle, M. (2015). Determination of vector error correction models in higher dimensions. Working paper, Leibniz Universität Hannover.

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

Lütkepohl, H. (2005). *New Introduction To Multiple Time Teries Analysis.* Springer Science & Business Media.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34:574–589.

Medeiros, M. C. and Mendes, E. F. (2016). ℓ_1-regularization of high-dimensional time series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191:255–271.

Palm, F. C., Smeekes, S., and Urbain, J.-P. (2010). A sieve bootstrap test for cointegration in a conditional error correction model. *Econometric Theory*, 26:647–681.

Palm, F. C., Smeekes, S., and Urbain, J.-P. (2011). Cross-sectional dependence robust block bootstrap panel unit root tests. *Journal of Econometrics*, 163:85–104.

Park, J. Y. and Phillips, P. C. (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 4(3):468–497.

Park, J. Y. and Phillips, P. C. B. (1989). Statistical inference in regressions with integrated processes: part 2. *Econometric Theory*, 5:95–131.

Phillips, P. C. and Hansen, B. E. (1990). Statistical inference in instrumental variables regression with I(1) processes. *The Review of Economic Studies*, 57:99–125.

Phillips, P. C. and Ouliaris, S. (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica*, 58:165–193.

Phillips, P. C. B. and Solo, V. (1992). Asymptotics for linear processes. *Annals of Statistics*, 20:971–1001.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Smeekes, S. and Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting*, 34(3):408–430.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202.

Wilms, I. and Croux, C. (2016). Forecasting using sparse cointegration. *International Journal of Forecasting*, 32:1256–1267.

Yamada, H. (2017). The Frisch–Waugh–Lovell theorem for the lasso and the ridge regression. *Communications in Statistics-Theory and Methods*, 46:10897–10902.