On a New Approach to Cointegration – The State-Space Error Correction Model

THOMAS RIBARITS† AND BERNARD HANZON‡

† European Investment Bank
Risk Management Directorate
Financial Risk Department
100, boulevard Konrad Adenauer, L-2950 Luxembourg
E-mail: ribarits@eib.org

‡ University College Cork
School of Mathematical Sciences
Cork, Ireland
E-mail: b.hanzon@ucc.ie

Summary
In this paper we consider cointegrated $I(1)$ processes in the state-space framework. We introduce the state-space error correction model (SSECM) and provide a complete treatment of how to estimate SSECMs by (pseudo-)maximum likelihood methods, including reduced rank regression techniques. In doing so, we follow very closely the Johansen approach for the VAR case; see Johansen (1995). The remaining free parameters will be represented using a novel type of local parametrization. Simulation studies and a real world application clearly reveal the potential of the new approach.

Keywords: Cointegration, state-space models, data-driven local coordinates, maximum likelihood estimation, reduced rank regression.

1. INTRODUCTION

The main aim of this paper lies in the generalization of the classical cointegration analysis of Johansen (1995) from the VAR framework to the state-space case. The motivations for this effort are the following:

- State-space models offer a larger and more flexible model class, containing VAR models as special cases.
- State-space and VARMA models are equivalent, yet the state-space approach offers many practical advantages such as simpler parametrization and more efficient methods to compute likelihood functions and their gradients.
- In practice, the data generating process (DGP) will typically neither be of VAR nor of VARMA type. Yet, its approximation by means of a state-space model will in general require less parameters than by means of a VAR model.
- State-space models are heavily used in continuous-time modelling, especially in financial applications, and discrete-time versions of such continuous-time state-space models are most naturally presented as discrete-time state-space models; examples include the work of Dai and Singleton (2000) on affine term structure models and references mentioned there.

In our attempt to generalize VAR cointegration to cointegration analysis in the state-space framework we will proceed along the lines of classical VAR cointegration. In Johansen’s celebrated method for calculating this paper has been presented at the 62nd European Meeting of the Econometric Society, 27 – 31 August 2007, Budapest, Hungary.
the maximum likelihood estimator of a cointegrated VAR model, the mathematical core can be summarized as follows:

1. First, a reduction of the likelihood maximization problem to a so-called reduced rank regression problem is achieved by ‘concentrating out’ a number of linear parameters in the VAR representation.
2. Then, the reduced-rank regression problem is solved.
3. After this concentration step, one is left with the problem of finding the optimum of a quotient of determinants of matrices which are quadratic in terms of an unknown matrix, \( \beta \) say; this unknown matrix \( \beta \) needs to be determined in such a way that an optimum is reached. The optimum can be calculated based on the solution of a generalized eigenvalue-eigenvector problem.
4. Steps 1–3 above solve the maximum likelihood problem in the VAR case, no other parameters remain to be estimated.

In the present paper this methodology is generalized to the case of cointegrated linear state-space models. To this end, we will first – in analogy to the celebrated VECM, the (autoregressive) vector error correction model – introduce the state-space error correction model which will be abbreviated by SSECM.

Then, mimicking the estimation of autoregressive VAR error correction models, the number of parameters appearing in the original non-linear likelihood optimization problem will be reduced by sequentially ‘concentrating out’ parameters with respect to which the optimization problem can be solved analytically. More precisely, the idea is that

1. by fixing two of the three matrices in the state-space representation, the constrained likelihood maximization problem can be reduced to a reduced rank regression problem by ‘concentrating out’ the third state-space matrix.
2. Then, the reduced-rank regression problem is solved.
3. After this concentration step, one is left with a problem of finding the optimum of a quotient of determinants of matrices being quadratic in terms of an unknown matrix \( \beta \) which is similar to step (3.) in the VAR case above. In general, the optimum can now be calculated based on the solution of a reduced rank generalized eigenvalue-eigenvector problem.
4. In contrast to the VAR case, substituting the optimal parameter values obtained in steps 1–3 above into the likelihood function leaves us with a criterion function that still depends on some (but a much smaller number of) unknown parameters. Instead of trying to choose a particular canonical form for the two matrices in the state-space representation that remain to be determined, we apply the so-called DDLC approach, a simple local parametrization technique which has been used e.g. in McKelvey et al. (2004) and Ribarits et al. (2004) in the context of stationary time series models.

An important remark is that the calculation of the gradient of the concentrated likelihood function, in case it exists, can be done by calculating the gradient of the original likelihood function before ‘concentration’ at the parameter point that results from the concentration step. Moreover, even if the concentrated likelihood function shall turn out to be non-differentiable, the ‘adapted’ gradient-based search algorithm proposed in this paper, which operates only on the remaining free parameters (after ‘concentration’), is still ensured not to get stuck at any other point than a critical point of the original likelihood function!1 Hence, 

1The importance of this lies in the fact that differentiation of eigenvalues and eigenvectors, which could lead to very tedious differentiation calculations in the algorithm, can be avoided in this way and very efficient numerical differentiation techniques can be directly applied to the original likelihood function. It is also important from the mathematical view-point as eigenvalues and eigenvectors obtained in the last concentration step need not be differentiable, potentially leading to a non-differentiable concentrated likelihood function.
one can find a local optimum of the likelihood function by means of a gradient-type search method, like for instance the Gauss-Newton algorithm. This in turn implies, at least for sufficiently large number of data, that we find the maximum likelihood estimator.

Using either

- Johansen’s classical VAR approach – the cointegrated VAR model obtained can be written in state-space form, and the state dimension can easily be reduced by a ‘truncating approach’
- or so-called (regression based) subspace algorithms

the determination of integer-valued parameters (the system order and the number of common trends) is performed. Both approaches also provide the user with good starting values for the remaining unknown real-valued parameters in the likelihood optimization problem.

The paper is organized as follows: In Section 2 the concept of cointegrated processes of order one is briefly reviewed. Section 3 introduces state-space models for cointegrated processes and, most importantly, the new state-space error correction model SSECM. Section 4 gives details on how we propose to approach the estimation problem for SSECMs by means of (pseudo-)maximum likelihood methods. Here, a stepwise ‘concentration’ of the likelihood is performed, leading to a generalized eigenvalue problem in the final step. In Section 5 we apply a new parametrization method (DDLC) in order to further reduce the number of parameters in the non-linear likelihood optimization problem. Section 6 intends to give a very brief overview of how model selection can be performed in the framework of SSECMs and also treats the question of how to obtain initial estimates for the real-valued parameters. In Section 7 we apply the new approach in two ways: First, a simulation study is performed using the data generating processes presented in Saikkonen and Luukkonen (1997). Second, data of the US economy as presented in Lütkepohl (1991) and Lütkepohl and Claessen (1997) are used in order to demonstrate the practicability of the proposed procedure. We finish in Section 8 with a summary and concluding remarks.

2. MOTIVATION FOR COINTEGRATION

Econometric time series often show non-stationarities, e.g. trends in the mean and variance. Whereas each individual component of the time series might show such non-stationary behaviour, economic theory may suggest that there are static long-term equilibrium relations between the variables under consideration. Such equilibrium relations, which are lost by differencing the series but contain important economic information, will in general not be satisfied exactly at each point in time. The idea, however, is that the deviation from the equilibrium relation at each point in time is at least ‘small’, namely stationary, as compared to the ‘large’, namely non-stationary, individual variables.

One way of modeling non-stationarities is by making use of integrated processes: A p-dimensional process \((y_t|t \in \mathbb{N})\) is called integrated of order 1, or briefly \(I(1)\), if \(\Delta y_t = (1 - z)y_t\) is ‘stationary’, where \(z\) denotes the Lag operator, i.e. \(z(y_t|t \in \mathbb{N}) = (0, y_1, y_2, \ldots)\).

In order to model the equilibrium relations, one usually resorts to cointegrated processes: An integrated vector valued process is called cointegrated if there exists at least one vector \(0 \neq \beta \in \mathbb{R}^p\) such that the process \((\beta'y_t)_{t \in \mathbb{N}}\) is ‘stationary’. As is well known, the number of such linear independent cointegrating vectors \(\beta\) is called the rank of cointegration, \(r\) say. The number \(c = s - r\) is the number of common trends. The subspace of \(\mathbb{R}^p\) spanned by the cointegrating vectors is called the cointegrating space.

As mentioned above, the equilibrium relations are not ‘exact’. However, the stationary deviation from
an exact equilibrium relation, given by \((\beta'y_t) - 0\), is relatively 'small' compared to the non-stationary components of \((y_t)\).

Figure 2 shows an example of a bivariate cointegrated process.

Figure 1. A cointegrated process \((Y_t, C_t)'\) with a cointegrating relation \(\beta = (0.7, -1)'; \beta \in \mathbb{R}^2\) is determined up to a multiplicative constant (unequal zero).

3. THE STATE-SPACE ERROR CORRECTION MODEL

We start by briefly discussing state-space representations of \(I(1)\) processes. Consider the following linear, time invariant, discrete-time state-space system:

\[
\begin{align*}
  x_{t+1} &= Ax_t + B\varepsilon_t, \quad x_1 = 0 \quad (3.1) \\
  y_t &= Cx_t + \varepsilon_t \quad (3.2)
\end{align*}
\]

Here, \(x_t\) is the \(n\)-dimensional state vector, which is, in general, unobserved; \(A \in \mathbb{R}^{n \times n}\), \(B \in \mathbb{R}^{n \times p}\) and \(C \in \mathbb{R}^{p \times n}\) are parameter matrices; \(y_t\) is the observed \(p\)-dimensional output. In addition, \((\varepsilon_t)_{t \in \mathbb{N}}\) is a \(p\)-dimensional (weak) white noise process with \(E\varepsilon_t = 0\) and \(E\varepsilon_t\varepsilon_t' = \Sigma\).

The state-space system in (3.1, 3.2) is called minimal if \(\text{rk}[B, AB, \ldots, A^{n-1}B] = \text{rk}[C', A'C', \ldots, (A')^{n-1}C']' = n\). It is called stable if the stability condition \(|\lambda_{\text{max}}(A)| < 1\) is satisfied, where \(\lambda_{\text{max}}(A)\) denotes an eigenvalue of \(A\) of maximum modulus.

It can be shown that for \((y_t)\) to be \(I(1)\), \(A\) must have an eigenvalue of one: \(\lambda_i(A) = 1\) must hold for some \(i \in \{1, \ldots, n\}\). Moreover, the algebraic multiplicity of the eigenvalue one must coincide with its geometric multiplicity, \(c\) say; see Bauer and Wagner (2003). If this is not the case, \((y_t)\) will show higher orders of integration. As mentioned in Section 2 above, the number \(c\) is the number of common trends. When referring to the state-space model (3.1,3.2) in the sequel, we will always assume \(A\) to have an eigenvalue one of (algebraic and geometric) multiplicity \(c\).

Note that setting up a state-space model (3.1,3.2) for some observed \(I(1)\) process \((y_t)\) requires the estimation...
of both integer-valued parameters \(n\) and \(c\) and real-value parameters in \(((A, B, C), \Sigma)\). As we shall see below, there is an inherent problem of non-identifiability for the latter parameters.

A solution to (3.1,3.2) can be given by \(y_t = \sum_{j=1}^{\infty} CA^{j-1} B \varepsilon_{t-j} + \varepsilon_t\) where we put \(\varepsilon_t = 0\) for \(t \leq 0\). Using a shorthand notation, such solution can be written as

\[
y_t = k(z)\varepsilon_t = [Cz(I - zA)^{-1}B + I]\varepsilon_t = \sum_{j=1}^{\infty} CA^{j-1} B z^j + I]\varepsilon_t
\]

Here, \(z \in \mathbb{C}\) is interpreted as a complex variable, and \(k(z)\) is called the transfer function corresponding to the state-space model. The transfer function is a rational matrix-valued function and describes the input-output behavior of the state-space system, with \((\varepsilon_t)\) being viewed as input and \((y_t)\) as output.

For writing down the Gaussian likelihood function, we will make use of the one step prediction errors \(\varepsilon_t = y_t - \mathbb{E}(y_t|y_1, \ldots, y_{t-1})\). The latter can be obtained by simply rewriting equations (3.1,3.2):

\[
x_{t+1} = \begin{pmatrix} A \\ B \end{pmatrix} x_t + \begin{pmatrix} \tilde{A} \\ \tilde{B} \end{pmatrix} \varepsilon_t
\]

\[
\varepsilon_t(\tilde{A}, \tilde{B}, \tilde{C}) = -\tilde{C} \tilde{A}^{-1} \tilde{B} \varepsilon_t + \varepsilon_t
\]

Note that (3.4,3.5) represents the inverse state-space system, i.e. the system with inputs \((\varepsilon_t)\) and outputs \((\varepsilon_t)\). The transfer function corresponding to the inverse system is given by

\[
\varepsilon_t = \tilde{k}(z)\varepsilon_t = [\tilde{C}z(I - z\tilde{A})^{-1}\tilde{B} + I]\varepsilon_t = \sum_{j=1}^{\infty} \tilde{C}\tilde{A}^{j-1} \tilde{B} z^j + I]\varepsilon_t
\]

Recall that \(A\) has an eigenvalue one of multiplicity \(c\), and this implies that the inverse transfer function \(\tilde{k}(z)\) has a \(c\)-fold root at one, i.e. \(\tilde{k}(1) \in \mathbb{R}^{p \times p}\) has rank \(p - c\) only! We can thus write

\[
\tilde{k}(1) = \tilde{C}(I - \tilde{A})^{-1}\tilde{B} + I = -\alpha \beta' \quad \alpha \in \mathbb{R}^{p \times r}, \beta \in \mathbb{R}^{p \times r}
\]

where \(r = p - c\). Equation (3.7) is in fact the starting point for deriving the state-space error correction model SSECM. Note that we have

\[
\begin{align*}
\tilde{k}(0) &= 0 \quad \text{at} \quad z = 1 \\
\tilde{k}(0) &= I \quad \text{at} \quad z = 0
\end{align*}
\]

\[
\tilde{k}(z) = \tilde{k}(1)z + \left[ \tilde{k}(z) - \tilde{k}(1)z \right] = \tilde{k}(1)z + (1 - z) \left[ I - \tilde{k}(z) \right] = \tilde{k}(1)z + I(1 - z) - \tilde{k}(z)(1 - z)
\]

where \(\tilde{k}(0) = 0\) must hold true. Putting \(\tilde{k}(z) = \sum_{j=0}^{\infty} \tilde{K}_j z^j\), we therefore get \(\tilde{K}_0 = 0\). The other \(\tilde{K}_j\) are easily obtained by means of a comparison of coefficients:

\[
\tilde{K}_j = \sum_{i=j+1}^{\infty} \tilde{K}_i = \left[ \tilde{C} \tilde{A}(I - \tilde{A})^{-1} \right] \tilde{A}^{j-1} \tilde{B}, \quad j \geq 1
\]
Substituting (3.8) and (3.7) into \( \varepsilon_t = \bar{k}(z)y_t \), we obtain

\[
\varepsilon_t = \bar{k}(z)y_t = -\alpha\beta' y_{t-1} + \Delta y_t - \bar{k}(z)\Delta y_t
\]  

(3.10)

Using (3.9), \( \bar{k}(z)\Delta y_t \) can be represented in state-space form as follows:

\[
\bar{x}_{t+1} = \bar{A}\bar{x}_t + \bar{B}\Delta y_t
\]  

(3.11)

\[
\bar{k}(z)\Delta y_t = C\bar{A}(I - \bar{A})^{-1}\bar{x}_t
\]

We are now ready to formulate the state-space error correction model (SSECM):

\[
\begin{align*}
\Delta y_t &= \alpha\beta' y_{t-1} + \bar{C}\bar{A}(I - \bar{A})^{-1}\bar{x}_t + \varepsilon_t \\
\text{where} & \quad \alpha\beta' = -\left[C(I-I\bar{A})^{-1}\bar{B} + I_p\right] \\
\bar{x}_{t+1} &= \bar{A}\bar{x}_t + \bar{B}\Delta y_t
\end{align*}
\]  

(3.12)

4. MAXIMUM LIKELIHOOD ESTIMATION OF SSECM

We will assume that \( n > p \) and \( \text{rank}(\bar{B}) = p \) in the sequel. This is mainly done for pedagogical reasons as the other cases require purely technical adjustments which do not provide any value added for the clarification of the applied principles. For a unified treatment of all possible cases as well as proofs of the results presented below, the reader is referred to Ribarits and Hanzon (2005).

In analogy to the VAR-case in Johansen (1995), we set

\[
Z_{0t} = \Delta y_t, \quad Z_{1t} = y_{t-1}, \quad Z_{2t} = \bar{x}_t, \quad M_{ij} = \frac{1}{T} \sum_{t=1}^{T} Z_{it}Z_{jt}^T
\]

Finding the (Gaussian) maximum likelihood estimator for the parameters in (3.12) translates to solving the following non-linear constrained optimization problem:

\[
\min_{L(\alpha, \beta, \bar{A}, \bar{B}, \Sigma)} L(\alpha, \beta, \bar{A}, \bar{B}, \Sigma) = \log \det \Sigma + \frac{1}{T} \sum_{t=1}^{T} \left(Z_{0t} - \alpha\beta' Z_{1t} - \bar{C}\bar{A}(I - \bar{A})^{-1}Z_{2t}\right)^T \Sigma^{-1} \left(Z_{0t} - \alpha\beta' Z_{1t} - \bar{C}\bar{A}(I - \bar{A})^{-1}Z_{2t}\right)
\]

s. t. \( \bar{C}(I - \bar{A})^{-1}\bar{B} = -(I_p + \alpha\beta') \)

(4.13)

This problem is tackled by performing a series of ‘concentration steps’.

4.1. Step 1: Concentrating out \( \bar{C} \)

Let \( (\alpha, \beta, \bar{A}, \bar{B}, \Sigma) \) be given and assume that \( M_{22} > 0 \). Then the unique global constrained minimizer solving \( \min_{L(\alpha, \beta, \bar{A}, \bar{B}, \bar{C}, \Sigma)} L(\alpha, \beta, \bar{A}, \bar{B}, \bar{C}, \Sigma) \) is independent of \( \Sigma \) and is given by

\[
\hat{\bar{C}} = [M_{02}(I - \bar{A})^{-1}\bar{A}' - \alpha\beta' M_{12}(I - \bar{A})^{-1}\bar{A}]H_{11} - (I_p + \alpha\beta')H_{21}
\]  

(4.14)

where

\[
\begin{pmatrix}
H_{11} & H_{12} \\
H_{21} & H_{22}
\end{pmatrix} = \begin{pmatrix}
\bar{A}(I - \bar{A})^{-1}M_{22}(I - \bar{A})^{-1}\bar{A}' & (I - \bar{A})^{-1}\bar{B} \\
(I - \bar{A})^{-1}\bar{B}
\end{pmatrix}^{-1}
\]

(4.15)

62nd European Meeting of the Econometric Society, 27 – 31 August 2007, Budapest
Inserting this optimizer into the likelihood function in (4.13) yields

\[
L^1_c(\alpha, \beta, \tilde{A}, \tilde{B}, \Sigma) = L(\alpha, \beta, \tilde{A}, \tilde{B}, \tilde{C}, \Sigma)
\]

\[
= \log \det \Sigma + \frac{1}{T} \sum_{t=1}^{T} (R_{0t} - \alpha \beta' R_{1t})' \Sigma^{-1} (R_{0t} - \alpha \beta' R_{1t})
\]

(4.16)

where \( R_{0t} = Z_{0t} - \left( [M_{02}(I - \tilde{A})^{-1}' \tilde{A}] H_{11} - H_{21} \right) \tilde{A}(I - \tilde{A})^{-1} Z_{2t} \)

(4.17)

\( R_{1t} = Z_{1t} - \left( [M_{12}(I - \tilde{A})^{-1}' \tilde{A}] H_{11} + H_{21} \right) \tilde{A}(I - \tilde{A})^{-1} Z_{2t} \)

(4.18)

4.2. Step 2: Concentrating out \( \alpha \)

Let us denote by \( S_{ij} \) the (non-centered) sample covariance matrices of the residuals \( R_i \) and \( R_j \), i.e.

\[
S_{ij} = \frac{1}{T} \sum_{t=1}^{T} R_{it} R_{jt}'
\]

Furthermore, let \((\beta, \tilde{A}, \tilde{B}, \Sigma)\) be given such that \( \beta' S_{11}^{-1} \beta > 0 \). We note that this restriction is ‘innocent’ as it will typically hold true; see Section ‘Concentrating out \( \alpha_1 \)’ in Ribarits and Hanzon (2005) for a detailed discussion, showing in fact that \( S_{11} > 0 \) can be assumed in the case considered in this paper. The unique global minimizer solving \( \min_\alpha L^1_c(\alpha, \beta, \tilde{A}, \tilde{B}, \Sigma) \) is independent of \( \Sigma \) and is given by

\[
\hat{\alpha} = S_{01} \beta (\beta' S_{11}^{-1} \beta)^{-1}
\]

(4.19)

Inserting this optimizer into the likelihood function in (4.16) yields

\[
L^{2c}(\beta, \tilde{A}, \tilde{B}, \Sigma) = L^1_c(\hat{\alpha}, \beta, \tilde{A}, \tilde{B}, \Sigma) =
\]

\[
\log \det \Sigma + \frac{1}{T} \sum_{t=1}^{T} (R_{0t} - S_{01} \beta (\beta' S_{11}^{-1} \beta)^{-1} \beta' R_{1t})' \Sigma^{-1} (R_{0t} - S_{01} \beta (\beta' S_{11}^{-1} \beta)^{-1} \beta' R_{1t})
\]

(4.20)

4.3. Step 3: Concentrating out \( \Sigma \)

This step is very common: Let \((\beta, \tilde{A}, \tilde{B})\) be given such that

\[
\frac{1}{T} \sum_{t=1}^{T} (R_{0t} - S_{01} \beta (\beta' S_{11}^{-1} \beta)^{-1} \beta' R_{1t})' (R_{0t} - S_{01} \beta (\beta' S_{11}^{-1} \beta)^{-1} \beta' R_{1t}) > 0
\]

Then the unique global minimizer solving \( \min_{\Sigma} L^{2c}(\beta, \tilde{A}, \tilde{B}, \Sigma) \) is given by

\[
\hat{\Sigma} = S_{00} - S_{01} \beta (\beta' S_{11}^{-1} \beta)^{-1} \beta' S_{10}
\]

(4.21)

Inserting (4.21) into (4.20) yields

\[
L^{3c}(\beta, \tilde{A}, \tilde{B}) = L^{2c}(\beta, \tilde{A}, \tilde{B}, \hat{\Sigma}) = \log \det [S_{00} - S_{01} \beta (\beta' S_{11}^{-1} \beta)^{-1} \beta' S_{10}] + p
\]

(4.22)
As in the VAR case, the minimization of \((4.22)\) with respect to \(\beta\) leads to a generalized eigenvalue problem: Using the equalities

\[
\det \left( \begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right) = \det \Sigma_{11} \cdot \det [\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}] = \det \Sigma_{22} \cdot \det [\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}]
\]

and assuming that \(S_{11} > 0, \beta S_{11} > 0\) (see Section 4.2 above) and \(S_{00} > 0\), representing ‘innocent’ restrictions, one can rewrite \((4.22)\) as

\[
L^3c(\beta, \bar{A}, \bar{B}) = \log \det S_{00} + \log \frac{\det \beta'(S_{11} - S_{10} S_{00}^{-1} S_{01}) \beta}{\det \beta' S_{11} \beta} + p \tag{4.23}
\]

The following lemma treating the generalized eigenvalue problem is recalled for the sake of completeness:

**Lemma 4.1.** Let \(M = M', N = N', M \geq 0\) and \(N > 0\). Then \(\det(\lambda N - M) = 0\) has \(p\) solutions \(\lambda_1 \geq \cdots \geq \lambda_p \geq 0\) with eigenvectors \(V = [v_1, \ldots, v_p]\) such that \(MV = NV\Lambda\), \(V'NV = I\) and \(V'MV = \Lambda\), where \(\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)\).

Putting \(M = S_{10} S_{00}^{-1} S_{01}\) and \(N = S_{11}\), all global minimizers solving \(\min_\beta L^3c(\beta, \bar{A}, \bar{B})\) in \((4.23)\) are then easily seen to be given by

\[
\hat{\beta} = [v_1, \ldots, v_r] \cdot \bar{T}, \quad \bar{T} \in \text{GL}(r)
\]  

where \(\bar{T} \in \text{GL}(r)\) is arbitrary and \(v_1, \ldots, v_r\) are the eigenvectors corresponding to the \(r\) largest eigenvalues \(\lambda_1, \ldots, \lambda_r\) of the generalized eigenvalue problem \(\det(\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}) = 0\). Note that we assume \(\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} \geq \cdots \geq \lambda_p > 0\). We remark that in general, e.g. if \(n < p\), this ‘concentration step’ becomes a reduced rank generalized eigenvalue problem; see Section ‘Concentrating out \(\beta_1\)’ in Ribarits and Hanzon (2005).

Inserting the optimizing arguments into \((4.23)\) finally yields

\[
L^3c(\hat{\beta}, \bar{A}, \bar{B}) = \log \det S_{00} + \log \frac{\det \beta'(S_{11} - S_{10} S_{00}^{-1} S_{01}) \beta}{\det \beta' S_{11} \beta} + p \tag{4.25}
\]

We have thus reduced the original likelihood optimization problem in \((\alpha, \beta, \bar{A}, \bar{B}, \bar{C}, \Sigma)\) to a problem in \((\bar{A}, \bar{B})\) only. The following section treats the question of whether a further reduction of the number of parameters is still possible.

5. **PARAMETRIZATION OF \((\bar{A}, \bar{B})\) USING THE DDLC APPROACH**

The parameters in \((\bar{A}, \bar{B})\) are non-identifiable: Let \(\text{GL}(n)\) denote the set of real non-singular \(n \times n\) matrices. It is not difficult to see that \((\bar{A}, \bar{B})\) and \((T \bar{T}^{-1}, \bar{T} B)\), \(T \in \text{GL}(n)\) give rise to (after carrying out in reverse order the optimization steps described in the sub-sections above) the same original transfer function \(\tilde{k}(z)\) and therefore the same likelihood value.

In order to treat the problem of the non-identifiability in \((\bar{A}, \bar{B})\), we introduce a novel local parametrization, making use of the so-called DDLC philosophy; see e.g. McKelvey et al. (2004) or Ribarits et al. (2004). DDLC stands for ‘data-driven local coordinates’: The parametrization is of local nature only, i.e. desirable properties for a parametrization can only be shown in an open neighborhood of a given point in the parameter space. Moreover, the parametrization can be viewed as a collection of (uncountably many) coordinate charts, where
On a New Approach to Cointegration – The State-Space Error Correction Model

the choice between these charts can be made in a data-driven manner. As will be seen below, we will change chart in each step of the iterative search algorithm.

The idea behind DDLC is simple: The set \( \mathcal{E}(\bar{A}, \bar{B}) = \{(T\bar{A}T^{-1}, T\bar{B}), T \in Gl(n)\} \) is called the \((\bar{A}, \bar{B})\)-equivalence class as all its elements \((T\bar{A}T^{-1}, T\bar{B})\) represent the same transfer function and thus the same likelihood value. This set has the structure of a real analytic submanifold of \(\mathbb{R}^{n^2+np}\) of dimension \(n^2\). The tangent space, \(Q(\bar{A}, \bar{B})\) say, can be easily computed. In fact, \((\bar{A}, \bar{B}) + (\bar{A}_s, \bar{B}_s) \in Q(\bar{A}, \bar{B}) \Leftrightarrow (\bar{A}_s, \bar{B}_s) = (\dot{T}\bar{A} - \bar{A}\dot{T}, \dot{T}\bar{B})\) for some \(\dot{T} \in \mathbb{R}^{n \times n}\). Using the relation \(\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)\), vectorization yields

\[
Q(\bar{A}, \bar{B}) = \left\{ \left( \begin{array}{c}
\text{vec}(\bar{A}) \\
\text{vec}(\bar{B})
\end{array} \right) \right\} + \left( \begin{array}{c}
A' \otimes I_n - I_n \otimes \bar{A} \\
B' \otimes I_n - I_n \otimes \bar{B}
\end{array} \right) \cdot \text{vec}(\dot{T}), \dot{T} \in \mathbb{R}^{n \times n} \right\}
\]

(5.26)

For DDLC, the idea is to consider as a parameter space the \(np\)-dimensional ortho-complement, \(Q^\perp\) say, to the tangent space (5.26) at the given \((\bar{A}, \bar{B})\). In this manner, we do not parametrize directions along (the tangent space to) the equivalence class (where the likelihood value stays constant), but we parametrize only directions in which the likelihood value changes:

\[
\varphi : \mathbb{R}^{np} \rightarrow \mathbb{R}^{n^2+np} \\
\tau \mapsto \left( \begin{array}{c}
\text{vec}(\bar{A}(\tau)) \\
\text{vec}(\bar{B}(\tau))
\end{array} \right) = \left( \begin{array}{c}
\text{vec}(\bar{A}) \\
\text{vec}(\bar{B})
\end{array} \right) + Q^\perp \tau
\]

(5.27)

Figure 5 illustrates again the main idea of parametrizing only directions in which the likelihood value changes (in first order).

Figure 2. The DDLC philosophy: The thick curve represents the equivalence class \(\mathcal{E}(\bar{A}, \bar{B})\). Each point in \(\mathcal{E}(\bar{A}, \bar{B})\) gives rise to the same likelihood value. The (red) dashed straight line represents the tangent space \(Q(\bar{A}, \bar{B})\). The straight line orthogonal to the dashed one thus corresponds to \(Q^\perp\), the ortho-complement of the tangent space to the equivalence class, which is the DDLC parameter space.

In the course of an iterative search algorithm, the DDLC construction is iterated in the sense that the pair \((\bar{A}(\tau), \bar{B}(\tau))\) corresponding to the updated parameter vector \(\tau\) becomes the new initial \((\bar{A}, \bar{B})\) pair (5.27) in
the next step of the iterative search procedure. For details regarding geometrical and topological properties of DDLC as well as its numerical advantages see Ribarits et al. (2004), Ribarits et al. (2005) or Ribarits et al. (2004).

Note that for the rest of this paper, \( \tau \in \mathbb{R}^{np} \) shall denote the DDLC parameter vector in (5.27). Hence, the concentrated likelihood function (4.25) can finally be rewritten as

\[
L^{4c}(\tau) = L^{4c}(\bar{A}(\tau), \bar{B}(\tau)) = \log \det S_{00} + \log \Pi_{i=1}^{r}(1 - \lambda_i) + p
\]

The remaining problem is to minimize (5.28) with respect to \( \tau \). For this purpose, we suggest to employ the following ‘adapted’ gradient-based search algorithm:

- First, compute the gradient of the original likelihood function \( L(\alpha, \beta, \hat{A}, \hat{B}, \hat{C}, \Sigma) \) in (4.13) with respect to \((\hat{A}, \hat{B})\).
- Second, evaluate it at the parameter point that results from the concentration steps, i.e. at the point \((\alpha, \beta, \hat{A}, \hat{B}, \hat{C}, \hat{\Sigma}) = (\hat{\alpha}, \hat{\beta}, \hat{A}, \hat{B}, \hat{C}, \hat{\Sigma})\), where \( \hat{C}, \hat{\alpha}, \hat{\Sigma} \) and \( \hat{\beta} \) are given in (4.14), (4.19), (4.21) and (4.24), respectively. This yields, after accounting for the derivative of \((\hat{A}, \hat{B})\) with respect to \( \tau \) – see (5.27) – the ‘gradient’ of \( L^{4c}(\tau) \).
- In fact, if \( L^{4c}(\tau) \) is differentiable, such ‘gradient’ represents the genuine gradient of \( L^{4c}(\tau) \), as is easily seen from an application of the chain rule. However, even if \( L^{4c}(\tau) \) shall turn out to be non-differentiable\(^2\), it can be shown that along this ‘gradient’ direction the original likelihood function (after carrying out in reverse order the optimization steps) is ensured to decrease if the step size is chosen small enough and a critical point of the original likelihood function has not yet been found. In other words, such ‘adapted’ gradient algorithm can only get stuck at points which coincide with critical points of the original likelihood function!

Of course, one can in practice also choose a Gauss-Newton type search direction instead of the ‘gradient’ direction, in particular as \( L^{4c}(\tau) \) will typically be differentiable and Gauss-Newton type algorithms do then usually have better convergence properties.

Apart from the theoretical assurance to decrease the criterion value by employing the ‘adapted’ gradient algorithm, there is a second, perhaps more apparent, advantage: Differentiating \( L^{4c}(\tau) \), if possible, would amount to a multiple application of the chain rule and to differentiation of eigenvalues and eigenvectors appearing in one of the concentration steps, which in turn would lead to very tedious calculations in the algorithm. Instead, the gradient of the original likelihood function \( L(\alpha, \beta, \hat{A}, \hat{B}, \hat{C}, \Sigma) \) with respect to \((\hat{A}, \hat{B})\) can be computed in a numerically efficient way using so-called extended prediction error filters, and we can avoid computationally expensive and inaccurate numerical differencing methods.

6. MODEL SELECTION AND INITIAL PARAMETER ESTIMATES FOR SSECM

Note that up to now we have assumed to know both the state dimension \( n \), i.e. the size of the \( \hat{A} \) matrix in (3.12) and the number of common trends \( c \). In practice, these integer-valued parameters have to be estimated from data. Similarly, it is desirable to start the likelihood optimization procedure with some ‘educated guess’ for \((\hat{A}, \hat{B})\).

\(^2\)Note that from the mathematical viewpoint eigenvalues and eigenvectors, which appear in the final concentration step, need not be differentiable, and such non-differentiability can translate into non-differentiability of the final concentrated likelihood function \( L^{4c}(\tau) \)!
In principle, one can think of many ways of how to find estimates for \( n, c \) and, thereafter, initial estimates for \((\hat{A}, \hat{B})\). In the simulation study presented below, we made use of so-called subspace algorithms, whereas in the application to US macroeconomic data we used Johansen’s VAR error correction models. For the sake of completeness both approaches are briefly sketched in the sequel although the determination of estimates for \( n, c \) and initial estimates for \((\hat{A}, \hat{B})\) does not fall into the main scope of this paper.

### 6.1. Model selection based on subspace algorithms

For the estimation of \( n \), we first regress \( \{y_t, \ldots, y_{t+j}\} \) onto \( \{y_{t-1}, \ldots, y_{t-p}\} \), where we may choose e.g. \( j = \hat{p} = 2k_{AIC} \). Here, \( k_{AIC} \) denotes the estimated lag order of a VAR model of the form \( y_t = \sum_{j=1}^{k} A(j) y_{t-j} + \varepsilon_t \), using the well-known AIC criterion function. Regressing \( \{y_t, \ldots, y_{t+j}\} \) onto \( \{y_{t-1}, \ldots, y_{t-p}\} \) yields a regression matrix, and we consider the number of singular values of the regression matrix that differ significantly from zero as the estimated order, \( \tilde{n} \) say. This is done via minimization of an information-type criterion called \( SVC(n) \); see Bauer and Wagner (2002), where this procedure is shown to deliver consistent order estimates.

As far as the estimation of the number of common trends \( c = p-r \) is concerned, Bauer and Wagner (2002) provide a possible test sequence, too. The tests are based on the result that under the null hypothesis of \( c \) common trends (\( n \) correctly specified) the asymptotic distribution of the \( c \) largest eigenvalues of \( T(\hat{A} - I_n) \) is equal to the distribution of the eigenvalues of \( J_0^1 W(t) dW'(t) \left[ \int_0^1 W(t)W(t) dW(t) \right]^{-1} \), where \( W(t) \) is a \( c \)-dimensional standard Brownian motion and \( \hat{A} \) is the estimated \( A \)-matrix from the subspace algorithm we use. A simulation of the distribution can thus be used for the construction of tests to determine \( c \).

Once \( n \) and \( c \) have been estimated, an initial estimate (\( \hat{A}, \hat{B} \)) is obtained by employing the \( \text{CCA} \) subspace algorithm. Subspace algorithms, which originate from the systems and control literature, are computationally unproblematic as they only rely upon linear regressions and a singular value decomposition. For a detailed analysis of the \( \text{CCA} \) algorithm in the \( I(1) \) context, see Bauer and Wagner (2002).

### 6.2. Model selection based on Johansen’s VAR approach

An alternative way of determining \( n, c \) and an initial estimate for \( (\hat{A}, \hat{B}) \) follows along the lines of the classical Johansen approach to cointegration. This route has been taken for the second example presented in Section 7 below.

First, we estimated a conventional VAR model (in levels). Let \( \hat{k}_{BIC} \) denote the estimated lag order of such VAR model of the form \( y_t = \sum_{j=1}^{k} A(j) y_{t-j} + \varepsilon_t \), using the well-known BIC criterion function.

Given the lag order \( \hat{k}_{BIC} \), we then performed the classical trace test in Johansen’s VAR error correction model to determine \( r \) and thus \( c = p-r \). The resulting VAR error correction model is readily obtained as well, and one can then transform it back to a conventional VAR model which can, in turn, be straightforwardly written in state-space form, \( (A_{joh}, B_{joh}, C_{joh}) \) say. Note that the state-space representation of a VAR model is widely known as the ‘companion form of the VAR model’. By construction, the \( A_{joh} \) matrix has one as \( c \)-fold eigenvalue.

Applying a suitable state transformation of the form \( (\hat{A}_{joh}, \hat{B}_{joh}, \hat{C}_{joh}) = (T A_{joh} T^{-1}, T B_{joh}, C_{joh} T^{-1}) \)
yields a state-space system with block diagonal $\hat{A}_{joh}$ matrix, i.e.

$$
\begin{pmatrix}
    x_{t+1,1} \\
    x_{t+1,2}
\end{pmatrix} =
\begin{pmatrix}
    I_c & 0 \\
    0 & \hat{A}_{joh}
\end{pmatrix}
\begin{pmatrix}
    x_{t,1} \\
    x_{t,2}
\end{pmatrix} +
\begin{pmatrix}
    \hat{B}_{joh}^1 \\
    \hat{B}_{joh}^2
\end{pmatrix}
\varepsilon_t
$$

(6.29)

$$
y_t = \begin{pmatrix}
    \tilde{C}_{joh}^1 \\
    \tilde{C}_{joh}^2
\end{pmatrix}
\begin{pmatrix}
    x_{t,1} \\
    x_{t,2}
\end{pmatrix} + \varepsilon_t
$$

(6.30)

where $(\hat{A}_{joh}^1, \hat{B}_{joh}^1, \tilde{C}_{joh}^1)$ represents a stable system, i.e. $|\lambda_{\text{max}}(\hat{A}_{joh}^1)| < 1$.

In order to determine $n$, one can now apply a so-called ‘balance and truncate’ approach to the stable subsystem. Balanced model truncation is a simple and, nevertheless, efficient model reduction technique for state-space systems; see e.g. Scherrer (2002) and the references provided there for the concept of stochastic balancing, which relies upon the determination of the singular values briefly mentioned in Section 6.1 above. For given $(\hat{A}_{joh}^1, \hat{B}_{joh}^1, \tilde{C}_{joh}^1)$ the computation of these singular values boils down to the solution of two matrix equations. Using an information-type criterion such as $SVC(n)$, we again consider the number of singular values that differ significantly from zero as the estimated order of the stable sub-system. By adding $c$, we then obtain the estimated system order $\hat{n}$. Also, by simply truncating $(\hat{A}_{joh}^1, \hat{B}_{joh}^1, \tilde{C}_{joh}^1)$ in (6.29,6.30) to size $\hat{n}$, we obtain an initial state-space model which can be written in $\text{SSECM}$ form, yielding initial estimates for $(\hat{A}, \hat{B})$.

7. APPLICATIONS

In this section we present a simulation study to provide a first check on the virtues of our approach as well as an application to real world data. Both studies make use of data already published in the literature before.

7.1. A simulation study

We consider four state-space systems $(A, B, C)$, each with $p = 3$ outputs $y_t$ and with state dimension $n = 3$. The number of common trends for the first model is $c = 0$, so the first model corresponds to a stationary output process ($y_t$). The other three models correspond to $I(1)$ processes ($y_t$) with $c = 1$, $c = 2$ and $c = 3$ common trends, respectively. The fourth model is therefore fully integrated, i.e. there does not exist a cointegrating vector. The poles $(\lambda_1, \lambda_2, \lambda_3)$ of the four corresponding transfer functions (i.e. the eigenvalues of the $A$ matrix in a minimal state-space representation) are given in Table 1. Note that the zeros are the same for all models, they are located at $(0.9504, -0.6464, 0) = 3.2 \cdot (0.297, -0.202, 0)$. In fact, the systems are VARMA(1,1) systems taken from Saikkonen and Luukkonen (1997), with the only difference that the modulus of the zeros has been made 3.2 times larger:

$$
y_t + \tilde{\psi}y_{t-1} = \varepsilon_t + \tilde{\Gamma}_1\varepsilon_{t-1}
$$

$$
\tilde{\psi} = \begin{pmatrix}
    -0.29 & -0.47 & -0.57 \\
    -0.01 & -0.85 & 1.00 \\
    -0.75 & 1.39 & -0.55
\end{pmatrix}
$$

$$
\tilde{\Gamma}_1 = -3.2 \cdot C_\gamma \cdot \text{diag}(0.297, -0.202, 0) \cdot C_\gamma^{-1}
$$

(7.31)
Simulation data comprising $T = 100$ and $T = 1000$ output observations $(y_1, \ldots, y_T)$ have been generated;

<table>
<thead>
<tr>
<th>$c$</th>
<th>$(0.9, 0.8, 0.7)$</th>
<th>$(1, 0.8, 0.7)$</th>
<th>$(1, 1, 0.7)$</th>
<th>$(1, 1, 1)$</th>
</tr>
</thead>
</table>

Table 1. Poles of the four data generating models.

$\varepsilon_t$ is taken to be Gaussian white noise with

$$
E\varepsilon_t\varepsilon_t' = \Sigma = \begin{pmatrix} 0.47 & 0.20 & 0.18 \\ 0.20 & 0.32 & 0.27 \\ 0.18 & 0.27 & 0.30 \end{pmatrix}
$$

(7.32)

Both for $T = 100$ and $T = 1000$, 500 times series have been generated.

Model selection for the SSECMs, i.e. estimation of the integer valued system order $n$ and the number of common trends $c$ has been performed by the subspace based model selection procedure briefly outlined in Section 6 above. Also, initial estimates for $(\bar{A}, \bar{B})$ have been obtained in this manner.

The main purpose of the simulation study is to compare our approach using SSECMs with the classical VAR approach, abbreviated by ARjoh in the tables below; see Johansen (1995). As a ‘by-product’, we also compare the results to the subspace estimates, which are abbreviated by SSsub; see Bauer and Wagner (2002). The comparison is three-fold: model selection, in particular estimation of the number of common trends, estimation of the cointegrating space and, finally, predictive power of the estimated models.

7.1.1. Model selection

Note that model selection in the VAR setting was done as follows: The lag order $k$ of the AR polynomial has been determined using an information criterion (AIC) as it has been used in the first step of the subspace based model. Having determined a lag estimate $\hat{k}_{AIC}$ in this way, the classical Johansen trace test has been used to determine the number $r = p - c$ of cointegrating relations.

Table 2 summarizes the results of the model selection procedure, both for $T = 100$ and $T = 1000$. Notice that the results for SSsub and SSECM are the same as the model selection procedure coincides for both approaches. Therefore, SSsub is omitted in Table 2.

It becomes clear from Table 2 (a) that

- the subspace based model selection procedure used for SSECM works better than the classical trace tests in the VAR framework
- model selection becomes more and more difficult for an increasing number of common trends
- increasing the sample size clearly improves the estimation of $c$, which is what one would expect, of course.

Tables 2 (b) and (c) show that

- the system order $n$ for SSECM (and SSsub) is underestimated for $T = 100$, while for $T = 1000$ the order estimates are much better. Note that for $c = 3$ to be correctly specified the system order must be estimated to be at least 3; see the last column in the first line of Table (b) on the left.
- the lag order estimates $\hat{k}_{AIC}$ clearly increase if $T$ gets larger. This is in accordance with what one would expect, as the true systems are VARMA(1,1) and can therefore only be approximated by VAR systems.
Table 2. (a) Percentage of incorrectly specified cointegration ranks for $T = 100$ (left) and $T = 1000$ (right). (b) Average estimated system order $\hat{n}$ for SSECM (and $SS_{\text{sub}}^c$) and average estimated lag order $\hat{k}_{\text{AIC}}$ for $AR^{joh}$ in case the cointegration rank has been correctly specified for $T = 100$ (left) and $T = 1000$ (right). (c) Average estimated system order $\hat{n}$ for SSECM (and $SS_{\text{sub}}^c$) and average estimated lag order $\hat{k}_{\text{AIC}}$ for $AR^{joh}$ in case the cointegration rank has been incorrectly specified for $T = 100$ (left) and $T = 1000$ (right).

Figure 3. Estimation of the state dimension of the stable part of the SSECM.
between the estimated cointegrating space spanned by the columns of $\hat{\beta}$ and the true cointegrating space spanned by the columns of $\beta$. Denoting, somewhat sloppily, two subspaces in $\mathbb{R}^p$ by $M$ and $N$ and using the same symbols for two matrices with linearly independent columns spanning these subspaces $M$ and $N$, the gap $d(M,N)$ between $M$ and $N$ is defined by

$$d(M,N) = \max \left\{ \sup_{x \in \text{sp}(M), \|x\|=1} (I - N(N'N)^{-1}N')x, \sup_{y \in \text{sp}(N), \|y\|=1} (I - M(M'M)^{-1}M')y \right\} \tag{7.33}$$

Note that $(I - N(N'N)^{-1}N')$ is the orthogonal projection onto the ortho-complement of $N$ and $(I - M(M'M)^{-1}M')$ is the orthogonal projection onto the ortho-complement of $M$. Clearly, $0 \leq d(M,N) \leq 1$ if $M$ and $N$ are of different dimension. Thus, considering the gap between the true and the estimated cointegrating space only makes sense if $c$ (and, dually, $r = p - c$) has been correctly specified. Note, furthermore, that in case of correct specification of $c = 0$, the whole of $\mathbb{R}^3$ comprises the cointegrating space (hence, the gap will be zero) and, similarly, for a correctly specified $c = 3$, the cointegrating space only contains the trivial zero vector (hence, the gap will be zero, too). Consequently, we consider the gaps only for the cases where $c = 1$ or $c = 2$ and where the model selection procedure yielded the true $c$; see Figure 4.

For $T = 100$ observations, we can conclude the following from Figure 4:

- **SSECM** yields the best estimates for the true cointegrating spaces. In particular,
  - **SSECM** is better than $AR^{joh}$; for $c = 2$ the superiority is also significant according to the performed $t$-test.
  - **SSECM** also yields significantly lower gaps than $SS^{sub}$ both for $c = 1$ and $c = 2$.

As expected, the gaps for $T = 1000$ are extremely small for any estimation approach. This can be explained by the fact that the cointegrating space can be estimated super-consistently. In a pairwise comparison, however, **SSECM** takes the lead both relative to $AR^{joh}$ and relative to $SS^{sub}$; see Figure 5.

7.1.3. Predictive power of estimated systems

In a final step, we also evaluated both AR and **SSECM** (and $SS^{sub}$) estimates in terms of their predictive (i.e. true out-of-sample) performance. For $T = 100$ observations, we can conclude the following from Figure 6:

- The predictive power of **SSECM**s is better than the one of $VAR$ models for all $c = 0, 1, 2, 3$. However, the $t$-tests indicate a clear superiority only for $c = 3$.
- Comparing **SSECM** and $SS^{sub}$, **SSECM** is better for $c = 0$ and $c = 1$, where the superiority is significant for $c = 0$. Surprisingly, $SS^{sub}$ turns out to have lower out-of-sample mean squared errors for $c = 2$ and $c = 3$.

Note that the true innovation variance is given by $\Sigma$ in (7.32) and $\det \Sigma = 0.0079$. For $T = 1000$, the comparisons remain unchanged, with the only difference that the $MSE$s (in absolute terms) are already very close to the theoretical value $0.0079$. This is the reason why we have omitted to include another figure for this case.

To sum it up, we see that if the data generating model is $VARMA(1,1)$ – and can then also be represented in state-space form – then **SSECM**s offer a clear advantage over the classical VAR framework.

62nd European Meeting of the Econometric Society, 27 – 31 August 2007, Budapest
7.2. Application to US macroeconomic data

Following Chapter 11 in Lütkepohl (1991) and Lütkepohl and Claessen (1997), we consider a four dimensional system \((p = 4)\) of the US economy. The data set is comprised of 136 quarterly observations for the years 1954 to 1987 of

- the logarithm of the seasonally adjusted real money stock, \(\ln(M_1)\)
- the logarithm of the seasonally adjusted GNP in billions of 1982 USD, \(\ln(GNP)\)
- the interest rate on new issues of 91 day treasury bills, \(r_s\)
- the yield on long term (20 years) treasury bonds, \(r_l\)

Model selection for the \(\text{SSECM}\), i.e. estimation of the integer valued system order \(n\) and the number of common trends \(c\) has been performed on the basis of Johansen’s classical VAR approach. Also, initial estimates for \((\bar{A}, \bar{B})\) have been obtained in this manner; see Section 6.2 above.

7.2.1. In-sample comparison with results in Lütkepohl (1991) For a start, we have performed a VAR analysis on the whole data set: The lag order \(k\) of the AR polynomial has been determined using BIC, leading to \(k_{\text{BIC}} = 2\) and an (unrestricted) estimated model of the form \(y_t = v + A(1)y_{t-1} + A(2)y_{t-2} + \varepsilon_t\). Note that for this example we included an intercept in the VAR model, which in turn has also been included in the \(\text{SSECM}\) afterwards - the inclusion of such constant vector does not pose any severe problems.

Based on the estimation of a VAR(2) model, the classical Johansen trace test, including the intercept, yielded \(r = 1\). Using own \texttt{MATLAB} code, we reconciled the final estimated VAR(2) error correction model with the one reported in Lütkepohl (1991). Our ML estimates for \(v, A(1), A(2)\) and hence \(\alpha, \beta\) and \(\Sigma\) coincided with the ones given in Lütkepohl (1991).

In the next step, following Section 6.2, we wrote the VAR(2) error correction model in state-space form and applied a suitable state transformation to obtain a system of the form (6.29,6.30). Note that this system has state dimension 8, and \(c = 3\) (as \(r = 1\)), i.e. the first 3 states are \(I(1)\). In order to reduce the state dimension, we employed a ‘balance and truncate’ approach to the 5-dimensional stable subsystem: The singular values corresponding to the stable subsystem, sometimes also referred to as canonical correlations are given by \((0.9948, 0.7696, 0.4698, 0.0190, 0.0070)\). Using the information-type criterion \(\text{SVC}(n)\), we considered only 2 canonical correlations as being different from zero; see Figure 3. Hence, the stable subsystem was truncated taking into account only 2 states and an \(\text{SSECM}\) of state dimension \(n = 5\) was then estimated.

In order to get an idea of the importance of the choice of the initial pair \((\bar{A}, \bar{B})\) for the subsequent iterative optimization algorithm, we did not only take the \(\text{SSECM}\) model corresponding to Johansen’s VAR(2) estimate as starting point for the estimation of the \(\text{SSECM}\). Instead, we also randomly generated 99 other initial points (‘centered at’ the one obtained directly from the VAR(2) model).

Table 3 reports a number of results for the VAR(2) model (first column), the \(\text{SSECM}\) model obtained by starting the optimization algorithm at the system obtained directly from the VAR(2) model (second column) and the \(\text{SSECM}\) model showing the best in-sample likelihood fit (‘Best \(\text{SSECM}\)’ in the third column). The results can be summarized as follows:

- Starting from random initial estimates yielded slightly different \(\text{SSECM}\) models. However, their in-sample fit and the estimated cointegrating space turned out to be stable – all gaps were smaller than 0.1, and the likelihood values showed little variation.

62nd European Meeting of the Econometric Society, 27 – 31 August 2007, Budapest
On a New Approach to Cointegration – The State-Space Error Correction Model

- All SSECM models yielded slightly better in-sample likelihood values than the VAR(2) model.
- All gaps between the estimated cointegrating space using VAR and SSECMs turned out to be very small.

<table>
<thead>
<tr>
<th></th>
<th>( AR_j^{\text{estimated}} )</th>
<th>SSECM ( n = 5 )</th>
<th>Best SSECM ( n = 5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cointegrating relations</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total number of parameters</td>
<td>27</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Number of parameters in iterative optimization</td>
<td>0</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
| Normalized cointegrating vector | \[
\begin{pmatrix}
1.0000 \\
-0.4646 \\
14.5193 \\
-9.3615
\end{pmatrix}
\] | \[
\begin{pmatrix}
1.0000 \\
-0.3870 \\
14.7788 \\
-10.7014
\end{pmatrix}
\] | \[
\begin{pmatrix}
1.0000 \\
-0.4341 \\
12.4244 \\
-7.8998
\end{pmatrix}
\] |
| Likelihood Value          | -37.00                        | -37.07           | -37.11              |

Table 3. Estimation on the whole data set: Comparison of Johansen’s VAR error correction model and 2 estimated SSECMs.

7.2.2. Out of sample comparison with results in Lütkepohl and Claessen (1997)  In a second step, we have used the first 120 observations only in order to be in a position to compare forecasts obtained from Johansen’s VAR error correction model with the ones obtained from the SSECM.

As far as model selection is concerned, an analogous approach as in Section 7.2.1 can be taken - in fact, we estimated the same models, i.e. VAR(2) and SSECM with \( n = 5 \), both with one cointegrating relation (\( r = 1 \)).

Similarly, we again started the optimization for the estimation of the SSECM at 100 different \((\bar{A}, \bar{B})\) pairs, the first one corresponding to the one obtained directly from the VAR(2) model, the other ones representing randomly ‘perturbed’ matrix pairs.

Using the VAR error correction approach, we reconciled results\(^3\) with the ones reported in Lütkepohl and Claessen (1997).

Table 4 reports the results in analogy to Table 3 discussed above:

- As opposed to the estimation on the whole sample, the choice of initial \((\bar{A}, \bar{B})\) pairs can cause the iterative search algorithm to terminate at rather different systems, in particular in terms of the estimated cointegrating space - note that in the present case the estimation sample size is rather small as compared to the number of parameters to be estimated.
- This is exemplified e.g. by the fact that the gap between the cointegrating space obtained from \( AR(2) \) and SSECM equals 0.0457, whereas the gap AR(2)-’Best SSECM’ equals 0.8114.
- Almost all (98%) estimated SSECMs show clearly better in-sample likelihood values than the VAR(2) model.
- The majority (61%) of estimated SSECMs exhibits also better out-of-sample likelihood values.

\(^3\) We compared the estimated \( \beta \) and the out-of-sample mean squared errors of the VAR(2) model with Lütkepohl and Claessen (1997), and they perfectly agreed.

62nd European Meeting of the Econometric Society, 27 – 31 August 2007, Budapest
The gap between the estimated cointegrating space using VAR and SSECM turned out to be very small for the majority of estimated SSECMs. However, the ones with particularly good in-sample (and out-of-sample) fits, exhibited cointegrating relations very different from the one obtained from the AR(2) model.

We would like to add two remarks: First, comparing the VAR(2) estimate obtained on the whole data set in Section 7.2.1 with the one obtained now on the first 120 data points yields already a considerable difference in terms of the cointegrating space: Using the VAR(2) model in both cases, the gap turns out to be 0.3. Second, we have observed also surprisingly big changes in the estimation of $\beta$ in VAR($k$) estimates when $k$ was further increased. In the light of these observations it then did not turn out to be too surprising that comparing the $\beta$ estimates for the VAR(2) and the best SSECM estimate also yielded big differences.

We want to stress that state-space modeling, in particular in terms of SSECMs, offers a number of advantages over VARMA models; see the conclusions presented in Section 8 below.

### 8. SUMMARY AND CONCLUSIONS

This paper has introduced a new approach towards $I(1)$ cointegration analysis. The basic idea was to generalize classical VAR cointegration to the state-space framework, leading to the introduction of the state-space error correction model SSECM. This generalization mimics the classical VAR case as much as possible. In fact, VAR models are embedded as a special case within the new framework – the suggested estimation algorithm does not encounter any (numerical) difficulties if the model is of VAR-type.

The main contributions of this paper are the following:

- We have introduced the state-space error correction model SSECM.

---

**Table 4.** Estimation on the first 120 data points: Comparison of Johansen’s VAR error correction model and 2 estimated SSECMs.

<table>
<thead>
<tr>
<th></th>
<th>$AR_j$ &amp; $k = 2$</th>
<th>SSECM $n = 5$</th>
<th>Best SSECM $n = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized cointegrating vector</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1.0000$</td>
<td>$-0.0430$</td>
<td>$-0.3889$</td>
<td>$1.0000$</td>
</tr>
<tr>
<td>$-16.7200$</td>
<td>$-12.5537$</td>
<td>$15.8996$</td>
<td>$0.3294$</td>
</tr>
<tr>
<td>$19.3500$</td>
<td>$-16.7200$</td>
<td>$19.3500$</td>
<td></td>
</tr>
</tbody>
</table>

- In all but 3 SSECM models $\ln(M_1)$, $r_s$ and $r_l$ are forecasted more accurately than in the VARMA model in Lütkepohl and Claessen (1997).
- The VARMA model in Lütkepohl and Claessen (1997) yields better forecasts for $\ln(GNP)$ than almost all SSECMs.

Finally, in Lütkepohl and Claessen (1997) the authors extend the VAR(2) analysis to the VARMA case and present also out-of-sample mean squared errors for each component. Comparing their estimated VARMA model with the SSECMs presented in this paper yields the following results:

- In all but 3 SSECM models $\ln(M_1)$, $r_s$ and $r_l$ are forecasted more accurately than in the VARMA model in Lütkepohl and Claessen (1997).
- The VARMA model in Lütkepohl and Claessen (1997) yields better forecasts for $\ln(GNP)$ than almost all SSECMs.
A complete treatment of how to estimate SSECMs by (pseudo-)maximum likelihood methods has been given. We have ‘concentrated out’ as many parameters as possible in order to reduce the dimensionality in the non-linear likelihood optimization problem.

- We have used a new data-driven local parametrization (DDLC) for the estimation of SSECMs.
- The gradient of the concentrated likelihood function, if existing, can be efficiently computed by using the original (non-concentrated) likelihood function. In any case, our suggested ‘adapted’ gradient algorithm can only get stuck at critical points which are also critical points of the original likelihood function.
- A simulation study has been conducted, showing the potential of the new approach. The true model being VARMA(1,1), SSECMs provided better estimates of the cointegrating spaces and better forecasts than classical VAR models.
- An application to US macroeconomic data has been presented. The estimated SSECMs outperformed the VAR estimates both in terms of in-sample and out-of-sample properties. A brief comparison to VARMA estimates shows advantages for SSECMs.

Finally, it should be stressed that SSECMs, despite representing the same model class as VARMA models, offer a number of advantages over the latter:

- Available estimation techniques for state-space models are much simpler from the point of view of computational implementation. This is mainly due to the fact that the innovations can be computed by a simple inversion of the state-space system (or, in general, the Kalman filter), whereas one has to resort to some sort of multi-step procedure to obtain the innovations in the VARMA framework.
- Using SSECMs in combination with a data-driven local parametrization, one can clearly circumvent the usage of canonical forms which often lead to considerable drawbacks:
  - The usage of canonical forms often leads to numerical difficulties when employing a gradient-based search algorithm.
  - Also, they may be very complex in nature; see e.g. the Echelon canonical form for the VARMA model in Lütkepohl and Claessen (1997).
  - Their application may require a great effort already in the model selection stage: The authors in Lütkepohl and Claessen (1997), for instance, estimated 2401 (!) initial models (corresponding to different Kronecker indices) in the model selection stage when we only estimated a single one (for determining the order $n$).
  - Finally, VARMA model selection using the Echelon canonical form might end up with a chosen Kronecker index which is incompatible with the estimated cointegrating rank $r$.
- Using SSECMs, one concentrations out as many parameters as possible and will, in general, therefore end up with a significantly lower number of parameters appearing in the final non-linear optimization problem as compared to the VARMA case.
- For SSECMs, the remaining free parameters in the non-linear optimization problem all correspond to a stable system. Note that parameters in the non-linear optimization problem corresponding to both stable and non-stable systems (as in the VARMA case) may cause convergence problems for an algorithm in practice as level sets may become very distorted.
- Finally, cointegration analysis using SSECMs is conceptually very close to the well known and classical VAR approach. One can therefore also use classical VAR cointegration techniques to directly obtain initial estimates\(^4\), and this fact should also allow practitioners to quickly adapt themselves to the SSECMs.

\(^4\)As a side remark, we do not rely upon a technical condition called ‘strict minimum phase assumption’ for an initial estimate of $(A, B, C)$ in our approach.
REFERENCES

On a New Approach to Cointegration – The State-Space Error Correction Model

Figure 4. Estimation of the cointegrating space for $T = 100$, where $c = 1$ (first row) and $c = 2$ (second row). The left column compares the gap between estimated and true cointegrating space, where estimation is done by SSECM and $SS^{sub}$. The right column provides the same information, comparing ARjoh and SSECM. The ‘winner’ of each comparison, i.e. the estimation approach yielding a lower gap in the majority (> 50%) of the estimation runs, is indicated by the usage of a larger font size for the corresponding axis label; $m_1$ and $m_2$ denote the arithmetic mean of the gaps for the estimation approach indicated on the $x$-axis and the $y$-axis, respectively; $N$ corresponds to the number of estimation runs (out of 500) where both estimation approaches of the comparison led to a correct specification of $c$; $H = 0$ indicates that a $t$-test (performed on the logarithms of the gaps) does not reject the null hypothesis of equal gap means for the two estimation approaches in the comparison, whereas $H = 1$ indicates a rejection in favour of the alternative hypothesis that the ‘winner’ of the comparison (as defined above) has indeed a lower mean gap (left-sided $t$-tests have been performed).

62nd European Meeting of the Econometric Society, 27 – 31 August 2007, Budapest
Figure 5. Estimation of the cointegrating space for $T = 1000$, where $c = 1$ (first row) and $c = 2$ (second row). For a detailed explanation see Figure 4.
On a New Approach to Cointegration – The State-Space Error Correction Model

Figure 6. Out-of-sample mean squared errors for $T = 100$, where $c = 0$ (first row), $c = 1$ (second row), $c = 2$ (third row) and $c = 3$ (last row). The estimated models have been used for one-step-ahead forecasts over 200 periods, yielding out-of-sample forecast errors $\hat{\varepsilon}_{501}, \ldots, \hat{\varepsilon}_{700}$, say. In this figure, $MSE = \det \frac{1}{200} \sum_{t=501}^{700} \hat{\varepsilon}_t \hat{\varepsilon}_t'$ are compared for estimated models where $c$ has been correctly specified. The left column compares the $MSE$ for SSECM and SS$^{sub}$, the right column for AR$^{joh}$ and SSECM. The ‘winner’ of each comparison, i.e. the estimation approach yielding a lower $MSE$ in the majority (> 50%) of the estimation runs, is indicated by the usage of a larger font size for the corresponding axis label; $m_1$ and $m_2$ denote the arithmetic mean of the $MSE$s for the estimation approach indicated on the $x$-axis and the $y$-axis, respectively; $N$ corresponds to the number of estimation runs (out of 500) where both estimation approaches of the comparison led to a correct specification of $c$; $H = 0$ indicates that a $t$-test does not reject the null hypothesis of equal $MSE$ means for the two estimation approaches in the comparison, whereas $H = 1$ indicates a rejection in favour of the alternative hypothesis that the ‘winner’ of the comparison (as defined above) has indeed a lower mean $MSE$ (left-sided $t$-tests have been performed).

62nd European Meeting of the Econometric Society, 27 – 31 August 2007, Budapest