DISCUSSION PAPERS
Institute of Economics
University of Copenhagen

03-05

Cointegration Analysis in the Presence of Outliers

Heino Bohn Nielsen

Studiestræde 6, DK-1455 Copenhagen K., Denmark
Tel. +45 35 32 30 82 - Fax +45 35 32 30 00
http://www.econ.ku.dk

Cointegration Analysis in the Presence of Outliers

Heino Bohn Nielsen*
Institute of Economics,
University of Copenhagen
heino.bohn.nielsen@econ.ku.dk

March 18, 2003

Abstract

The effects of innovational outliers and additive outliers in cointegrated vector autoregressions are examined and it is analyzed how outliers can be modelled with dummy variables. Using a Monte Carlo simulation it is illustrated how misspecified dummies may distort inference on the cointegration rank in finite samples. That questions the common practice in applied cointegration analyses of including unrestricted dummy variables to account for large residuals. Instead it is suggested to test the adequacy of a particular specification of dummies prior to determining the cointegration rank. The points are illustrated on a UK money demand data set.

KEYWORDS: Cointegrated VAR, Innovational outlier, Additive outlier, Dummy variables, Monte Carlo.
JEL CLASSIFICATION: C32.

*I am grateful to Hans Christian Kongsted, Bent Nielsen, participants at the Conference on Bridging Economics and Econometrics in Florence 2001, as well as participarts in seminars at the Institute of Economics, Copenhagen, and Nuffield College, Oxford, for numerous comments and suggestions.

1 Introduction

Economic time series are frequently affected by special events, for instance policy interventions, strikes or gross measurement errors. Such events often show up as large residuals, or outliers, in econometric models and that raises two issues for an applied econometrician; first the inferential consequences of outliers if they are not detected, and second how the irregularities can be modelled with dummy variables. This paper addresses these issues for the case of the cointegrated vector autoregression (VAR), which has been widely applied in many fields of empirical research.

The effects of non-modelled outliers in autoregressive models depend on their precise nature and a distinction is often made between *innovational outliers* (IOs) and *additive outliers* (AOs), see *inter alia* Fox (1972), Tsay (1986) or Muirhead (1986). An IO is produced by a shock to the innovation term of a data generating process (DGP). The shock is then propagated through the autoregressive structure of the model and the total effects on the levels of the variables depend on the autoregressive parameters. An AO, on the other hand, is superimposed on the levels of the data, i.e. independently of the autoregressive parameters.

In the case of a fixed number of outlying observations asymptotic inference in the cointegration model is unchanged, in the sense that the asymptotic distributions are unaffected. The distortionary effects could be important in finite samples, however, and in applied cointegration analyses it is commonplace to include dummy variables to *whiten* residuals. Using simulations Doornik, Hendry, and Nielsen (1998) find that ignored IOs have only minor consequences for small sample inference on the cointegration rank of a VAR process, while Franses and Haldrup (1994), Shin, Sarkar, and Lee (1996) and Vogelsang (1999) find that AOs may bias inference towards the finding of stationarity or cointegration.

The second issue, how outliers in a cointegrated VAR can be modelled with dummy variables, is less resolved. An IO is straightforward to model with an unrestricted dummy variable, whereas an AO is more difficult because it imposes non-linear restrictions on the dynamics and requires a more complicated estimation procedure. In applications of the cointegrated VAR it is common practice to identify outlying observations from the estimated residuals and to model outliers with unrestricted (innovational) dummies, see Hendry and Juselius (2001), Johansen (1996, chapter 5) and Juselius (2002, chapter 6). But to the best of our knowledge there is little justification for this practice; and it is not obvious that outliers of a general form, and AOs in particular, can be successfully modelled with unrestricted dummies.

In this paper we use a Monte Carlo simulation to analyze how IOs and AOs can be approximated with dummy variables in small samples. We find that the usual innovational model is misspecified if an outlier is additive, and in general it is very difficult to approximate AOs with unrestricted dummies. As a result the usual practice may bias the estimated parameters in the case of AOs and may distort inference on the cointegration rank. As an alternative we propose a simple algorithm for estimating the cointegrated VAR with additive dummies, and we suggest a testing procedure between the innovational model and the additive model in the detection of outlying observations.

The plan for the rest of the paper is as follows. First, Section 2 introduces the cointegrated VAR and two ways to include dummy variables. Section 3 briefly presents the estimation

algorithms and some hypotheses of interest. Section 4 then gives an empirical illustration of the role of outliers and dummy variables based on UK money demand data, and the effects of outliers and dummy variables are then analyzed in a Monte Carlo simulation in Section 5. Finally, Section 6 concludes.

2 Models for Deterministic Components

We consider the $p-$dimensional cointegrated VAR

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{t-i} + \mu_0 + \alpha\beta_0' t + \mu_t + \epsilon_t, \quad t = 1, 2, ..., T, \tag{1}$$

where ϵ_t are $i.i.d.$ Gaussian innovations, $N(0, \Omega)$. The parameters α and β are both of dimension $p \times r$ such that the rank of $\Pi = \alpha\beta'$ is $r \leq p$, and the remaining autoregressive parameters, $\Gamma_1, ..., \Gamma_{k-1}$, are each of dimension $p \times p$.

We consider a deterministic specification given by an unrestricted constant, μ_0, and a restricted linear drift term, $\mu_1 t = \alpha\beta_0' t$. This model allows for a linear trend both in the stationary and non-stationary directions of the data and is often favored in empirical applications, see Nielsen and Rahbek (2000) for a comparison with other specifications. Finally, (1) includes an additional deterministic function, μ_t, which will contain indicator variables. To characterize impulses and balanced impulses respectively, we define for a particular observation T_0 the indicators

$$D_t(T_0) = 1\{t = T_0\} \quad \text{and} \quad d_t(T_0) = \Delta D_t(T_0),$$

where $1\{\cdot\}$ is the indicator function equal to one if the expression in curly brackets is true. For ease of notation we will sometimes drop the direct reference to T_0 and denote the variables D_t and d_t. Throughout the paper we consider (1) with $\mu_t = 0$ as the baseline specification and we denote this model $H^*(r)$. In the subsections below we present two distinct ways to augment the baseline model with indicator variables, and we characterize the resulting specifications both as DGPs and estimation models. When (1) is a DGP we denote by *outliers* the effects on the data of the indicator variables in $\mu_t \neq 0$[1]. When (1) is an estimation model we use *dummy variables* to denote the indicators included in μ_t to approximate irregularities in the data.

If the $(p-r) \times (p-r)$ matrix $\alpha_\perp' \Gamma \beta_\perp$ is non-singular[2] and the roots of the characteristic polynomial

$$A(z) = (1-z) I - \alpha\beta' z - \sum_{i=1}^{k-1} \Gamma_i (1-z) z^i, \tag{2}$$

are located either outside the complex unit circle or at $z = 1$, the solution for Y_t is given by the Grangers Representation

$$Y_t = C \sum_{i=1}^{t} (\epsilon_i + \mu_t) + C_1(L)(\epsilon_t + \mu_t) + \kappa_0 + \kappa_1 t + A, \tag{3}$$

[1]Outliers are necessarily defined relative to a statistical model. The used notation implies that irregular observations are outliers relative to the baseline model, $\mu_t = 0$.

[2]For a $p \times r$ matrix α we denote by $\alpha_\perp$ the $p \times (p-r)$ dimensional orthogonal complement such that $\alpha'\alpha_\perp = 0$ and $span(\alpha : \alpha_\perp) = \mathbb{R}^p$. Further we define $\Gamma = I - \sum_{i=1}^{k-1} \Gamma_i$.

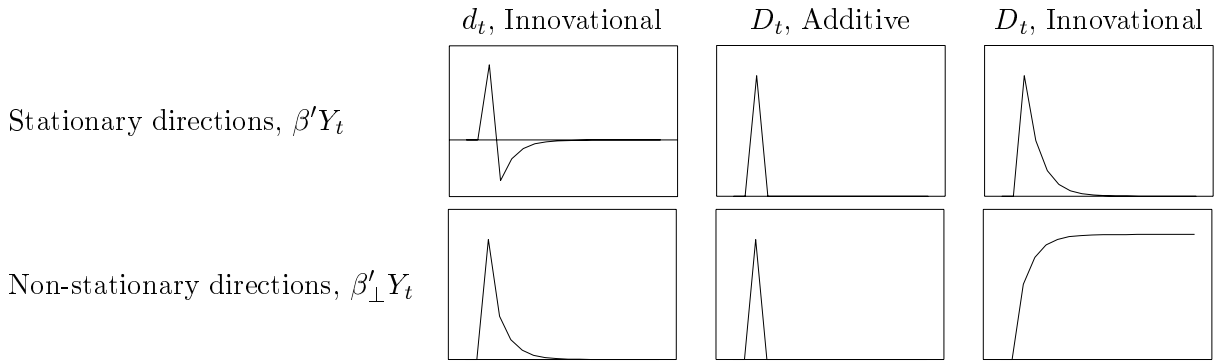

Figure 1: Examples of the effects (relative to the baseline of no outliers) of different types of outliers in the stationary and non-stationary directions. Based on a 2-dimensional VAR(1) with $r = 1$.

where $C = \beta_\perp \left(\alpha'_\perp \Gamma \beta_\perp\right)^{-1} \alpha'_\perp$ has reduced rank $p-r$, $C_1(L)$ is an infinite but convergent matrix polynomial, κ_0 and κ_1 are functions of the parameters and A involves the initial values, see Johansen (1996, theorem 4.2).

2.1 The Innovational Model, $H_I^*(r)$

The usual way to include dummy variables in cointegration analyses is to specify the deterministic function in (1) as $\mu_t = \phi \mathcal{D}_t$, where $\mathcal{D}_t$ is a $n-$dimensional vector of dummy variables and ϕ is a $p \times n$ matrix of unrestricted coefficients, see Johansen (1996, chapter 5) and Hendry and Juselius (2001). It follows from (3) that the total effect in Y_t is given by $C \sum_{i=1}^{t} \phi \mathcal{D}_t + C_1(L) \phi \mathcal{D}_t$, and unless $\alpha'_\perp \phi = 0$ the levels will contain the cumulated effect of the variables in $\mathcal{D}_t$. Since $\beta' C = 0$, the cointegrating relations, $\beta' Y_t$, annihilate both the common stochastic trends and the cumulated effect of $\mathcal{D}_t$, implying that the stationary and non-stationary directions are not balanced in terms of the deterministic specification.

The interpretation of an innovational impulse outlier at time T_0, i.e. the indicator variable $\mathcal{D}_t = D_t(T_0)$ included in a DGP, is a large shock that follows the same dynamic adjustment as the usual innovations. In columns one and three of Figure 1 we have augmented a simple $2-$dimensional DGP with $k = 1$ lag and a cointegration rank of $r = 1$ with a balanced impulse, $\mathcal{D}_t = d_t$, and a non-balanced impulse, $\mathcal{D}_t = D_t$. The effects in the stationary and non-stationary directions, relative to the baseline of no outliers, are reported in the two rows.

In an estimation model an included innovational impulse dummy, $\mathcal{D}_t = D_t(T_0)$, renders the corresponding residual, ϵ_{T_0}, equal to zero and eliminates the contribution from that residual to the likelihood function.

2.2 The Additive Model, $H_A^*(r)$

An alternative additive formulation can be written as the unobserved components model

$$Y_t = X_t + \theta \mathcal{D}_t \tag{4}$$

$$A(L) X_t - \mu_0 - \alpha \beta'_0 t = \epsilon_t, \tag{5}$$

where X_t are unobserved variables that obey a cointegrated VAR and θ contains unrestricted coefficients. In this case the effects on the levels of the observed data, Y_t, are independent of the autoregressive parameters of the model. The interpretation of an impulse AO, i.e. the indicator variable $\mathcal{D}_t = D_t$ included in the DGP, could be an isolated measurement error, for instance a typing mistake. It follows from (4) that $\theta\mathcal{D}_t$ enters the Granger Representation additively and in general, the stationary and non-stationary directions are balanced, cf. column two of Figure 1.

To characterize the properties of the additive specification as an estimation model it is useful to solve (4) and (5) for the observed variables to obtain

$$\Delta Y_t = \alpha\left(\beta' : \beta_0' : \beta_1'\right)\begin{pmatrix} Y_{t-1} \\ t \\ \mathcal{D}_{t-1} \end{pmatrix} + \sum_{i=1}^{k-1}\Gamma_i \Delta Y_{t-i} + \sum_{i=0}^{k-1}\theta_i \Delta\mathcal{D}_{t-i} + \mu_0 + \epsilon_t, \tag{6}$$

subject to the k sets of restrictions

$$\beta_1 = -\theta_0'\beta \tag{7}$$

$$\theta_i = -\Gamma_i\theta_0, \quad i = 1, ..., k-1. \tag{8}$$

Note, that the dummy variables are included with the full lag structure of the endogenous variables, $\mu_t = A(L)\theta\mathcal{D}_t$. The restrictions (7) and (8) ensures an instantaneous transition to the effects of the dummies and the effects on Y_t are qualitatively identical to the dummy variable itself. If these restrictions are not imposed the transition is approximated by the k free parameters $(\theta_0, ..., \theta_{k-1})$, which is closely related to the model proposed by Johansen, Mosconi, and Nielsen (2000) for broken levels and broken linear trends. While the approximation is often appropriate in the case of a transition to a structural break it is very costly in terms of degrees of freedom for the case of isolated outliers and there is a potential efficiency gain by imposing the restrictions[3].

It is worth noting that an additive impulse dummy, $D_t(T_0)$, eliminates the effects of the observation, Y_{T_0}, from the likelihood function rather than the effects of the residual, ϵ_{T_0}. This is equivalent to the interpretation of a dummy variable in a static model, where an entire case is cancelled; and it is closely related to the interpolation of missing values, see Gomez, Maravall, and Peña (1999).

3 Estimation and Some Hypotheses of Interest

Maximum Likelihood estimation of the basic model, $\mu_t = 0$, is based on reduced rank regression (RRR), which reduces to solving the eigenvalue problem $\left|\lambda S_{11} - S_{10}S_{00}^{-1}S_{01}\right| = 0$, where $S_{ij} = T^{-1}\sum_{t=1}^{T} R_{it}R_{jt}'$ are sample moment matrices, and R_{0t} and R_{1t} are least squares residuals of

[3]Note, that the basic model $H^*(r)$ can be expressed as the unobserved component representation, $Y_t = X_t + \widetilde{\theta}_0 + \widetilde{\theta}_1 t$, where X_t follows a cointegrated VAR with no deterministic components. In this case the non-linear restrictions (7) and (8) cancel due to coincidence between Δt and the constant. The balance of the deterministic components of this model implies that the tests for reduced rank are asymptotic similar with respect to the coefficients to the linear trend, see Nielsen and Rahbek (2000).

regressing ΔY_t and $Y_{t-1}^* = \left(Y_{t-1}' : t\right)'$ respectively on the unrestricted variables $U_t = (\Delta Y_{t-1}' : ... : \Delta Y_{t-k+1}' : 1)'$, see Johansen (1996, chapter 6). That yields $p+1$ ordered eigenvalues, $1 > \widehat{\lambda}_1 > \widehat{\lambda}_2 > ... > \widehat{\lambda}_p > \widehat{\lambda}_{p+1} = 0$, and the estimate of the cointegrating relations, $\widehat{\beta}^* = (\widehat{\beta}' : \widehat{\beta}_0')'$, is given by the eigenvectors corresponding to the r largest eigenvalues.

Different values of r define a sequence of nested models, $H^*(0) \subset ... \subset H^*(r) \subset ... \subset H^*(p)$, and a Likelihood Ratio (LR) test for a cointegration rank smaller or equal r against a rank smaller or equal p is given by the so-called Trace test statistic

$$Q_r = -2 \log Q \left(H^*(r) \mid H^*(p) \right) = -T \sum_{i=r+1}^{p} \log \left(1 - \widehat{\lambda}_i \right).$$

The asymptotic distribution of Q_r depends on the deterministic specification and involves functionals of Brownian motions. In the determination of the cointegration rank the usual strategy is to reject a model, $H^*(r)$, only if all the more restricted models, $H^*(0), ..., H^*(r-1)$, are also rejected, starting from the most restricted model $H^*(0)$, see Johansen (1996, chapter 12).

The presence of innovational dummies does not complicate estimation. The dummies, $\mathcal{D}_t$, can be included in U_t and concentrated out prior to the RRR. The asymptotic distribution of the Trace test statistic is not changed by a fixed number of outliers or a fixed number of innovational dummies, see Doornik, Hendry, and Nielsen (1998).

It is also straightforward to modify the RRR procedure to estimate the additive model (6) without the restrictions (7) and (8). The first differences, $\Delta\mathcal{D}_t, ..., \Delta\mathcal{D}_{t-k+1}$, can be included in U_t, and Y_t^* can be augmented with the lagged levels, $\mathcal{D}_{t-1}$. If $\mathcal{D}_t$ is an impulse or a balanced impulse, the effect of the dummy *per se* is asymptotically negligible but the increased dimension of the eigenvalue problem will change the asymptotic distribution of the Trace test. To be specific, each included impulse dummy will add an independent $\chi^2(p-r)$ contribution to the distribution of Q_r, see Johansen, Mosconi, and Nielsen (2000) and Doornik, Hendry, and Nielsen (1998).

When the restrictions (7) and (8) are imposed, no closed form solution for the estimator exists. Maximum Likelihood estimates can be obtained by a standard numerical (e.g. gradient based) procedure, but due to the number of parameters this method can be relatively slow. As an alternative we suggest a simpler algorithm that switches between two conditional Maximum Likelihood estimations for which closed form solutions exist. The idea is that conditional on an estimate $\widehat{\theta}$ of θ, the parameters in (5) can be found from a RRR of the corrected variables, $X_t = Y_t - \widehat{\theta}\mathcal{D}_t$. And given the parameters of the cointegrated VAR (5), the parameters to the deterministic variable, θ, can be estimated by generalized least squares from the residuals and the structure in (4). That suggests an iterative procedure, see the Appendix for details.

3.1 Tests Between IOs and AOs

For a given dummy variable, $\mathcal{D}_t = D_t(T_0)$, both the innovational model, $H_I^*(r)$, and the additive model, $H_A^*(r)$, nest the basic model, i.e.

$$H^*(r) \subset H_I^*(r) \quad \text{and} \quad H^*(r) \subset H_A^*(r), \tag{9}$$

while $H_I^*(r)$ and $H_A^*(r)$ are not mutually nested[4]. Each of the hypotheses in (9) can be tested using e.g. a LR test statistic of the form

$$\tau_j = -2 \cdot \log Q\left(H^*(r) \left| H_j^*(r)\right.\right), \quad j = I, A. \tag{10}$$

Note, however, that if $\mathcal{D}_t$ contains dummy variables involving a fixed number of observations, like $d_t(T_0)$ and $D_t(T_0)$, then the Law of Large Numbers and the Central Limit Theorem do not apply to the corresponding columns of $\widehat{\phi}$ and $\widehat{\theta}$, see e.g. Davidson (2001, p.147). As a consequence, the estimates are not consistent, see also Doornik, Hendry, and Nielsen (1998), and the asymptotic distribution of $\tau_j(T_0)$ is only χ^2 under the (non-testable) assumption that the particular residuals in question are Gaussian under the null of no outliers.

When the basic model, $H^*(r)$, can be rejected against both outlier models, $H_I^*(r)$ and $H_A^*(r)$, a choice has to be made between the non-nested candidates. One possible test strategy is to find a model that nests both alternatives, $H_N^*(r)$ say, and test the non-nested models against the nesting alternative. In many situations, however, $H_N^*(r)$ is not a natural maintained model, and an alternative is to apply a model selection approach. One strategy is to apply the Likelihood Dominance Criterion of Pollak and Wales (1991), which states that since the two candidate models, $H_I^*(r)$ and $H_A^*(r)$, contain the same number of parameters, then irrespectively the significance level and the chosen nesting model, $H_N^*(r)$, a reduction from $H_N^*(r)$ to the candidate model with the lower likelihood would not be accepted unless the reduction to the outlier model with the higher likelihood is also accepted. That implies that if both outlier models are significant over the basic model, $H^*(r)$, and we do not allow for both an IO and an AO for the same observation, the model with the higher likelihood could be selected. A similar line of arguments could be constructed using conventional information criteria.

3.2 Outlier Detection

In the construction of the test (10) the location of a potential outlier is considered known. In practical applications, however, this is typically not the case and an important issue is how to identify also the location of outlying observations. Usually in applied cointegration analyses, outliers are identified from *a priori* information and the size of the residuals. As an example, Hendry and Juselius (2001) include unrestricted dummy variables for observations with absolute values of the standardized residuals larger that 3.3.

To determine the location as well as the type of outliers, an automatic outlier detection procedure can be used, see also Tsay (1986), Chen and Liu (1993) and Tsay, Peña, and Pankratz (2000)[5]. This corresponds to performing the test (10) for all observations, $\tau_j(T_0)$ for $T_0 = 1, 2, ..., T$, producing two series of test statistics, $\tau_j(1), ..., \tau_j(T)$, $j = I, A$. It is natural to focus

[4]For the last observation, $\mathcal{D}_t = D_t(T)$, the two models are equivalent.

[5]It should be noted that outliers are defined relative to the statistical model and the outlier detection should in principle be performed conditional on the cointegration rank. If the correct rank is imposed on the model, however, the memory structure should be very similar to the full rank case, and simulations (not reported) indicate that outlier detection based on the full rank model is roughly as effective as outlier detection based on the true reduced rank model. This facilitates a sequential procedure; First the outlying observations can be identified in the stationary VAR, and then the cointegration rank can be determined conditional on the identified outliers.

on the largest test statistic, i.e.

$$\tau_j^{\max} = \max_{1 \leq T_0 \leq T} \{\tau_j(T_0)\}. \tag{11}$$

Under the Gaussianity assumption the individual LR statistics, $\tau_j(T_0)$, are asymptotically $\chi^2(p)$. An approximation of the critical values of the maximum test statistic, $\tau_j^{\max}$, can be based on the Bonferroni inequality. In particular we can use the multi comparison significance level $\delta^{\max} = 1 - (1-\delta)^{1/T}$ in each test to ensure an overall Type I error frequency below δ.

If multiple outliers are present, we follow *inter alia* Tsay (1986) and Tsay, Peña, and Pankratz (2000) and include a dummy for the most significant outlier, and repeat the procedure conditional on previously identified outliers.

4 Empirical Illustration

To illustrate the importance of outliers and dummy variables in cointegration analyses we consider the quarterly UK money demand data $Y_t = (m_t : y_t : \Delta p_t : r_t)'$, $t = 1963:1, ..., 1989:2$, where m_t denotes the log of real money M1, y_t denotes the log of real final expenditures, Δp_t denotes the change in the log of the deflator of y_t, and r_t denotes the difference between the three month interest rate and a measure of the own interest rate of M1.[6]

The data have previously been analyzed in *inter alia* Hendry and Doornik (1994) and Doornik, Hendry, and Nielsen (1998) using a VAR with $k = 2$ lags and we set up a similar model. Hendry and Mizon (1993) analyze the data for a shorter sample and include two innovational dummy variables, $\mathcal{D}_t = (Doil_t : Dout_t)'$, where

$$\begin{aligned} Doil_t &= 1\{t = 1973:3\} + 1\{t = 1973:4\} + 1\{t = 1979:3\} \\ Dout_t &= 1\{t = 1972:4\} + 1\{t = 1973:1\} + 1\{t = 1979:2\} \end{aligned}$$

and give a detailed account for their interpretation of the dummies[7]. Hendry and Doornik (1994) and Doornik, Hendry, and Nielsen (1998) include the same dummies for the extended sample.

The Trace statistics for determining the cointegration rank are reported in Table 1. Row (A) is based on the specification with no dummies. The hypothesis of $r = 0$ can be easily rejected whereas the test for $r \leq 1$ is a borderline case with a $p-$value of 8%. Row (B) reports the results based on a model including the original innovational dummies, $Doil_t$ and $Dout_t$. This specification clearly point towards $r = 1$. The dummy variables, $Doil_t$ and $Dout_t$, identify 6 outliers in the data. The restriction that these can be modelled with two composite dummies is not important for the rank determination, and row (C) reports results based on a model with 6 unrestricted dummies.

Irrespectively that the second cointegrating relation appears to be insignificant in their preferred model, Doornik, Hendry, and Nielsen (1998) retain it and continue the analysis with

[6] All calculation have been performed using a set of procedures programmed in Ox 3.0, see Doornik (2001).

[7] $Dout_t$ accounts for expansionary economic policy measures attributed to the Heath-Barber boom and the first effect of the Thatcher government. $Doil_t$ accounts for the effects of the two oil crises.

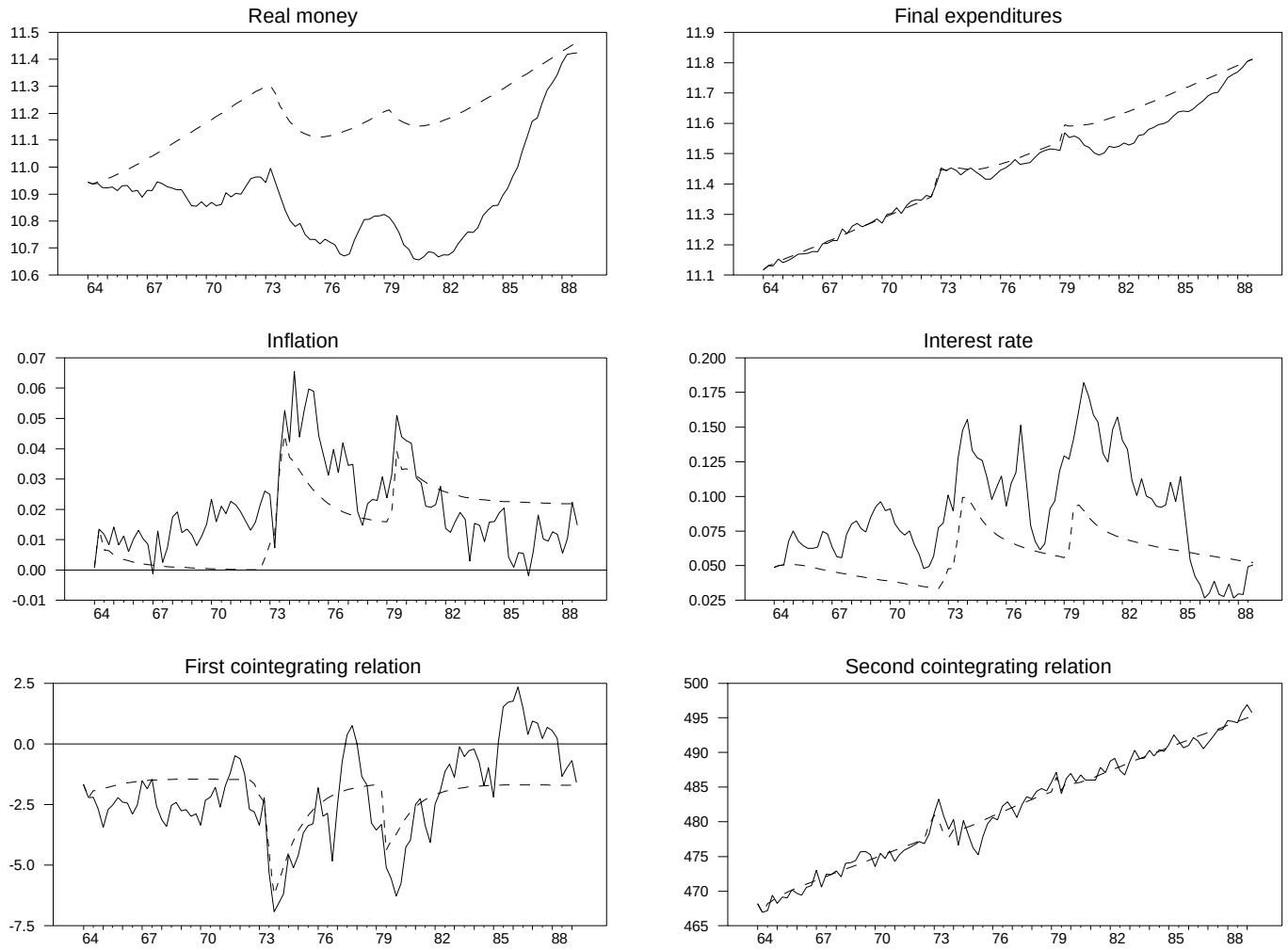

Figure 2: Original data (—) and the estimated deterministic components of the original model of Doornik, Hendry, and Nielsen (1998) (- -). Calculated for r=2.

	Dummy specification	Hypotheses			
		$r = 0$	$r \leq 1$	$r \leq 2$	$r \leq 3$
(A)	No dummies.	119.38	40.89	12.12	4.48
		[.00]	[.08]	[.80]	[.67]
(B)	Original dummies, $\mathcal{D}_t = (Doil_t : Dout_t)'$.	108.53	29.25	14.77	6.20
		[.00]	[.55]	[.60]	[.45]
(C)	The 6 unrestricted dummies from (B).	106.19	32.48	14.13	5.07
		[.00]	[.37]	[.65]	[.59]
(D)	Preferred specification.	129.37	47.67	14.05	5.56
	3 additive and 2 innovational dummies, see Table 2.	[.00]	[.01]	[.66]	[.53]

Table 1: Trace test statistics, Q_r, based on different dummy specifications. $p-$values based on the $\Gamma-$approximation of Doornik (1998) in square brackets.

$r = 2$. To illustrate the role of the dummies in the original specification of Doornik, Hendry, and Nielsen (1998), Figure 2 depicts the actual data together with the estimated deterministic components. The latter reflects the total effects of the dummies, $Doil_t$ and $Dout_t$, and the linear drift term, and is generated from the initial values by setting all innovations equal to zero in the estimated model. Furthermore the cointegrating linear combinations are given together with the estimated deterministic components. The unrestricted impulse dummies produce marked shifts in the levels of the data and the short run transitions are prolonged. The dummies have no long run effects in the cointegrating relations, but the short run effects account for a considerable proportion of the variation around the linear trend. We believe this kind of illustration could be very useful in empirical applications, because it highlights the consequences of a given deterministic specification.

As an alternative to the original model we apply the automatic outlier detection procedure, and Table 2 reports the outliers detected in each iteration. In first iteration the observation $1973 : 2$ is chosen to be an outlier. Hendry and Mizon (1993) identify 1973 as a year for potential problems due to the Heath government attempts to "go for growth" and the first oil crisis. The particular observation is associated with a large positive residual in the real money equation and a negative residual in the inflation equation, but the observation is not modelled by the original dummies. Based on the likelihood values it is difficult to distinguish a model with an innovational dummy and an additive dummy, but the innovational model is marginally preferred. In second iteration, which is conditional on the dummy for $1973 : 2$, an AO in $1974 : 2$ is detected, apparently in the inflation equation. Neither this one is picked up by the original dummies. In third iteration an outlier in $1973 : 1$ is detected. This is an outlier mainly in total expenditures and is also identified in $Dout_t$ and interpreted as an expansionist policy measure. The tests suggest to model the outlier as additive but the likelihood function is not very informative on the outlier type. Number four is an IO for $1979 : 2$ located in total expenditures, and is also included in $Dout_t$. In iteration five an AO is detected for $1971 : 1$. Hendry and Mizon (1993) mention the Competition and Credit Control Regulations in 1971 as a special event but the particular observation is not modelled.

There are still 22 observations for which a dummy is significant according to the individual critical value $\chi^2_{0.95}(4)$, but none of these are significant judged at the joint significance level $\delta^{\max}$.

Iteration	Detected outliers		Test statistics		Critical values		Standardized residuals			
	Date	Type	τ_I	τ_A	$\chi^2_{0.95}(4)$	$\chi^2_{0.9995}(4)$	m_t	y_t	Δp_t	r_t
1	1973:2	IO	29.785	28.924	9.488	19.942	4.07	0.39	-3.95	-1.90
2	1974:2	AO	24.911	29.874	9.488	19.942	-0.04	1.22	3.28	-1.30
3	1973:1	AO	26.246	27.209	9.488	19.942	-0.51	3.98	-1.36	0.86
4	1979:2	IO	24.799	8.514	9.488	19.942	0.33	4.40	1.03	-0.10
5	1971:1	AO	18.858	22.167	9.488	19.942	2.83	-2.32	-0.02	-0.18

Table 2: Outlier detection. The critical value $\chi^2_{0.95}(4)$ is based on the distribution of individual test statistics, while $\chi^2_{0.9995}(4)$ is based on the Bonferroni inequality and allows for the fact that in each iteration we focus on the maximum test statistic.

For the 4 observations of the original dummies not selected in Table 2, $1972:4$ is individually and close to jointly significant with a test statistic for an IO of 16.3. The remaining three are insignificant judged by the individual critical value with test statistics of 6.3 for an IO in $1973:3$, a test statistic of 7.1 for an AO in $1973:4$ and a test statistic of 2.9 for an AO in $1979:3$.

Our preferred model, based on the formal outlier detection, includes 3 additive dummies and 2 innovational dummies and thereby 20 parameters to dummy variables. The rank determination based on the preferred specification is presented in row (D) of Table 1 and suggests a cointegration rank of $r = 2$. This is in line with the priors of Doornik, Hendry, and Nielsen (1998) based on economic theory.

This empirical example illustrates that inference on the cointegration rank is sensitive to the specification of dummy variables, and care should be taken in the design of the empirical model. Furthermore, it illustrates how a formal outlier detection, combined with *a priori* knowledge on the timing of special events in the data, can be a useful tool.

5 Monte Carlo Simulation

In this section we set up a Monte Carlo simulation to analyze how outliers can be modelled with dummy variables in small samples. In particular, it is of interest to assess if the common strategy of including unrestricted dummies is robust also to the presence of additive outliers.

As the baseline DGP in the simulations we use the preferred model estimated in Section 4, where all dummy variables are excluded. This is a $4-$dimensional VAR(2) with a cointegration rank of $r = 2$.[8] We augment the basic DGP with $n = 1, 2, 4, 6$ impulse indicator variables, $\mathcal{D}_t = (D_t(T_1) : D_t(T_2) : ... : D_t(T_n))'$, where $T_i = i \cdot T \cdot (n+1)^{-1}$, to produce IOs or AOs located equidistantly in the time series. Each outlier is imposed randomly on one of the four variables, and has a magnitude of 5 times the residual standard deviation of the particular equation.

Since the set-up involves a fixed number of outliers, the asymptotic effects are negligible, and we focus on the small sample properties of different estimation strategies[9]. From the DGP

[8] The simulation has also been carried out for other DGPs. While the absolute levels of size and power vary the conclusions on outliers and dummy variables remain unchanged.

[9] An alternative set-up, where outliers are assumed to occur with a fixed probability and the (expected) number

we generate samples for $t = -101, ..., 0, 1, ..., T$ and discard 100 observations to minimize the importance of the initial values (taken from the actual data). For each case the same pseudorandom Gaussian innovations are used. The Monte Carlo results are based on 10000 replications and the asymptotic standard deviation of an estimator of the true rejection probability, γ, is given by $\sigma_\gamma = \left[\gamma \cdot (1-\gamma) \cdot 10000^{-1}\right]^{1/2}$ and for example $\sigma_{.05} \simeq .002$.

5.1 Modelling Known Outlying Observations

To illustrate the ability of different specifications to approximate outlying observations we first assume that the locations of outliers are known. We estimate the model (1) with $\mu_t = \mu_{1t} + \mu_{2t} + ... + \mu_{nt}$, where μ_{it} is given by the 6 following configurations of dummy variables:

$$\begin{aligned}
\mathcal{M}_0: &\quad \mu_{it} = 0 \\
\mathcal{M}_1: &\quad \mu_{it} = \theta_{i0} d_{it} \\
\mathcal{M}_2: &\quad \mu_{it} = \theta_{i0} D_{it} \\
\mathcal{M}_3: &\quad \mu_{it} = \alpha\theta_{i0} D_{it-1} + \theta_{i1} d_{it} + \theta_{i2} d_{it-1} \\
\mathcal{M}_4: &\quad \mu_{it} = \theta_{i0} D_{it} + \theta_{i1} D_{it-1} + \theta_{i2} D_{it-2} \\
\mathcal{M}_5: &\quad \mu_{it} = A(L)\,\theta_{i0} D_{it}.
\end{aligned}$$

Based on each configuration we calculate the Trace tests for rank determination and compare with the appropriate critical values calculated from a $\Gamma-$approximation based of the mean and variance of the asymptotic distribution estimated in a response surface regression by Doornik (1998). Furthermore, we analyze the precision of the estimated long-run parameters by imposing the correct rank, $r = 2$, and compare the estimated cointegration space, $\widehat{\beta}$, with the space spanned by the columns of β in the DGP. The comparison is based on the metric of Larsson and Villani (2001)[10]. For the case $r = 2$ this measure, $\Lambda(\widehat{\beta}, \beta)$, is bounded between $\Lambda = 0$ if $\widehat{\beta} \subset sp(\beta)$ and $\Lambda = \sqrt{2}$ if $\widehat{\beta} \subset sp(\beta_\perp)$. The distance can be interpreted as a simple measure of the efficiency of different estimators.

The simulation results for sample lengths $T = 75$ and $T = 200$ observations are reported in Table 3 and Table 4 respectively. Panel (A) and (B) of the tables report rejection frequencies of the Trace tests. The test $Q_0 = -2\log Q(r = 0 \mid r \leq 4)$ of no cointegration is (almost) always rejected and is not reported. Panel (A) reports the rejection frequencies of the test statistic $Q_1 = -2\log Q(r \leq 1 \mid r \leq 4)$ based on a nominal size of 5%. This illustrates the power of the test of a too restricted model, $r \leq 1$, against the unrestricted VAR when $r = 2$ in the DGP. Panel (B) reports the rejection frequencies of $Q_2 = -2\log Q(r \leq 2 \mid r \leq 4)$, which is the empirical size of the tests of the (true) cointegration rank $r = 2$. Finally, panel (C) reports the average distance $\Lambda(\widehat{\beta}, \beta)$ given the correct rank, $r = 2$.

---

of outliers increases with the sample length, is used in Franses and Haldrup (1994). In this case the asymptotic distributions will typically involve nuisance parameters, see also Shin, Sarkar, and Lee (1996) for a comparison of the two approaches.

[10] To compare the subspaces $span(A)$ and $span(B)$, we first find an orthonormal basis for each subspace, a and b say, and decompose $b = a\gamma_1 + a_\perp\gamma_2$, where $\gamma_1 = (a'a)^{-1}a'b = a'b$ and $\gamma_1 = (a'_\perp a_\perp)^{-1}a'_\perp b_\perp = a'_\perp b$. We now note that $\alpha_\perp$ is the as far as we can get from a and the distance is simply measured by the size of the coefficient γ_2 evaluated by the matrix norm, i.e. $\Lambda = Trace(\gamma'_2\gamma_2)^{1/2}$. The measure can also be written as a function of the angles between all pairs of columns in a and b, see Larsson and Villani (2001).

In the text we will concentrate on the results for the small sample, $T = 75$, and make reference to the medium sample, $T = 200$, only in case of differences in results.

Estimation model $\mathcal{M}_0$. The first column of the tables report results for the basic model, $\mathcal{M}_0 : \mu_{it} = 0$, which ignores potential outliers. First row in each panel illustrates the benchmark case where the DGP contains no outliers. In the small sample, $T = 75$, the Trace test is somewhat oversized with a rejection frequency of $r \leq 2$ against $r \leq 4$ of approximately 11%. The power of the test for $r \leq 1$ against $r \leq 4$ is reasonably high at around 80%. For $T = 200$ the actual size is down to 7% and the power is 100%. The size and power properties are specific to the chosen DGP and constitute the benchmark for alternative specifications.

The following rows illustrate the effects of non-modelled outliers in the data. It appears that IOs moves the actual distributions to the left, marginally decreasing the size and power. Panel (C) illustrates that the average precision of the estimated parameters increases with the number of IOs. This reflects that IOs provide events of large variations, and because IOs follow the same autoregressive adjustment as the small chocks, IOs can be helpful in revealing the autoregressive structure. The overall conclusion is that IOs do not seriously distort inference in the cointegration model, see Doornik, Hendry, and Nielsen (1998) for a similar result.

The effects of ignored AOs are more severe. The size distortion increases markedly with the number of outliers, confirming the findings in Franses and Haldrup (1994). Furthermore, the average precision in panel (C) illustrates that AOs are pure noise in the autoregressive model and they can potentially distort statistical inference.

An important question is now if the outliers can be successfully modelled with dummy variables. And below we consider a number of different specifications.

Estimation model $\mathcal{M}_1$. In the second column outliers are modelled by balanced impulse dummies, $\mathcal{M}_1 : \mu_{it} = \theta_{i0} d_{it}$. These dummies are too 'small' in the sense that they can at most create impulses in the non-stationary directions, i.e. column 1 of Figure 1, whereas the DGP replicates column 3 and 2 respectively. The specification is not unusual, however, and Juselius (2002, chapter 6) argues that AOs (e.g. typing mistakes) can be described by this kind of balanced innovational dummies. See also Hendry and Juselius (2001).

In the case of impulse IOs in the data the test based on $\mathcal{M}_1$ is clearly under-sized, and the power is rapidly decreasing. In the case of AOs the size is only marginally distorted but power is much lower than for the baseline specification. In both cases the combination of power and size implies that the proportion of cases where the correct rank, $r = 2$, is selected using the normal test sequence is markedly lower for $\mathcal{M}_1$ than for the basic model $\mathcal{M}_0$.

The same picture appears for the average precision of the estimates of β. Both for IOs and AOs the results are clearly inferior to ignoring the outliers.

Estimation model $\mathcal{M}_2$. In the third column we consider unrestricted impulse dummies, $\mathcal{M}_2 : \mu_{it} = \theta_{i0} D_{it}$. This is probably the most common modelling strategy to control for outliers.

In the case of IOs the estimation model $\mathcal{M}_2$ is identical to the DGP and we obtain good results. The size distortion is decreased compared to the basic model $\mathcal{M}_0$, and the test of the

Outliers		Known location							Outlier detection		
in DGP		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_{4b}$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$	$\mathcal{M}_8$
					(*A*) Power, $Q_1 = -2\log Q\,(r \leq 1 \mid r \leq 4)$						
0		**80.5**	...	...	...	...	...	...	80.9	80.8	81.3
1	IO	79.8	71.0	**82.2**	87.7	79.9	82.2	80.1	82.8	82.0	83.1
2	IO	78.7	59.4	**84.2**	92.7	80.7	84.9	79.6	85.0	84.0	85.3
4	IO	77.5	40.2	**88.6**	97.7	82.0	88.4	79.1	87.7	86.2	89.6
6	IO	75.7	26.5	**92.4**	99.4	84.2	92.2	78.0	90.6	87.6	92.8
1	AO	80.3	76.0	88.2	79.5	81.5	82.8	**80.3**	87.6	82.2	81.7
2	AO	80.8	71.9	82.7	79.6	82.0	84.5	**79.8**	91.9	83.7	82.3
4	AO	83.3	63.5	97.3	82.5	83.5	87.1	**79.5**	95.8	87.1	84.5
6	AO	85.8	56.8	99.0	87.5	85.3	89.7	**79.0**	97.4	90.2	86.7
					(*B*) Size, $Q_2 = -2\log Q\,(r \leq 2 \mid r \leq 4)$						
0		**10.9**	...	...	...	...	...	...	11.3	10.8	11.6
1	IO	9.6	7.0	**9.6**	25.0	10.7	11.0	9.7	10.6	10.8	10.8
2	IO	8.8	4.2	**9.1**	38.7	10.9	11.5	8.9	10.3	10.7	10.6
4	IO	7.9	1.8	**7.5**	62.3	11.7	12.1	7.9	9.3	10.5	9.6
6	IO	7.2	.7	**6.9**	79.5	12.5	13.9	7.3	9.4	10.0	9.6
1	AO	12.0	10.3	19.7	12.1	12.5	13.3	**10.7**	19.6	13.1	12.1
2	AO	13.3	9.7	28.9	13.8	14.8	16.7	**10.7**	28.1	15.7	13.4
4	AO	17.1	8.5	46.5	19.0	17.9	21.5	**10.4**	41.4	20.0	16.3
6	AO	20.3	8.5	60.2	28.5	21.4	25.9	**10.8**	51.0	24.4	19.9
					(*C*) Average distance, given $r = 2$						
0		**.2001**	...	...	...	...	...	...	.2007	.2007	.2018
1	IO	.2014	.2102	**.1876**	.2091	.1961	.1925	.1964	.1909	.1948	.1912
2	IO	.1997	.2148	**.1736**	.2146	.1892	.1829	.1926	.1798	.1881	.1792
4	IO	.1900	.2155	**.1571**	.2246	.1812	.1717	.1820	.1684	.1790	.1639
6	IO	.1825	.2139	**.1448**	.2379	.1778	.1650	.1755	.1600	.1734	.1537
1	AO	.2118	.2139	.2258	.2099	.2080	.2103	**.2012**	.2255	.2109	.2047
2	AO	.2216	.2274	.2466	.2181	.2162	.2204	**.2020**	.2451	.2225	.2079
4	AO	.2346	.2501	.2753	.2365	.2322	.2380	**.2046**	.2688	.2391	.2165
6	AO	.2475	.2729	.2971	.2595	.2521	.2578	**.2083**	.2876	.2571	.2285

Table 3: Simulation results, T=75. Results for the size and power are the rejection frequencies at a nominal 5% level of the Trace test for the cointegration rank. Note that the tests are not performed sequentially. Bold indicates that the estimation model and the DGP coincide. Results are based on 10000 replications.

Outliers		Known location							Outlier detection		
in DGP		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_{4b}$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$	$\mathcal{M}_8$
					(*A*) Power, $Q_1 = -2\log Q\,(r \leq 1 \mid r \leq 4)$						
0		**100.0**	...	...	...	...	...	...	100.0	100.0	100.0
1	IO	100.0	100.0	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	IO	100.0	100.0	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0	100.0
4	IO	100.0	100.0	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0	100.0
6	IO	100.0	100.0	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1	AO	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	100.0	100.0	100.0
2	AO	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	100.0	100.0	100.0
4	AO	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	100.0	100.0	100.0
6	AO	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	100.0	100.0	100.0
					(*B*) Size, $Q_2 = -2\log Q\,(r \leq 2 \mid r \leq 4)$						
0		**7.0**	...	...	...	...	...	...	7.0	7.3	7.3
1	IO	7.1	6.2	**6.9**	23.9	7.3	7.2	6.9	7.2	7.2	7.2
2	IO	6.5	5.0	**6.6**	40.2	7.2	7.3	6.3	7.0	6.8	6.9
4	IO	6.4	3.6	**5.8**	65.9	7.1	6.9	5.9	6.9	6.3	6.2
6	IO	6.5	2.8	**6.0**	81.0	7.5	7.7	6.1	7.3	6.9	6.7
1	AO	7.6	7.3	10.6	7.6	7.6	7.7	**7.0**	7.7	10.4	7.3
2	AO	8.3	7.6	14.4	7.8	8.1	8.7	**7.0**	8.3	13.3	7.4
4	AO	9.8	8.0	22.8	9.2	9.1	10.1	**6.9**	10.0	20.5	7.5
6	AO	11.2	8.4	31.7	11.0	10.4	12.1	**6.9**	11.8	27.6	7.7
					(*C*) Average distance, given $r = 2$						
0		**.0571**	...	...	...	...	...	...	.0571	.0571	.0572
1	IO	.0573	.0575	**.0559**	.0566	.0563	.0561	.0567	.0563	.0562	.0561
2	IO	.0574	.0578	**.0550**	.0565	.0559	.0554	.0566	.0558	.0554	.0553
4	IO	.0565	.0573	**.0527**	.0554	.0542	.0534	.0552	.0542	.0536	.0532
6	IO	.0551	.0563	**.0510**	.0555	.0533	.0520	.0536	.0531	.0522	.0517
1	AO	.0580	.0580	.0586	.0576	.0576	.0577	**.0572**	.0577	.0586	.0573
2	AO	.0586	.0586	.0599	.0582	.0581	.0583	**.0572**	.0585	.0599	.0574
4	AO	.0598	.0600	.0623	.0591	.0591	.0596	**.0573**	.0597	.0621	.0576
6	AO	.0607	.0611	.0650	.0601	.0600	.0608	**.0574**	.0613	.0646	.0578

Table 4: Simulation results, T=200. See notes for Table 3.

too restricted model, $r \leq 1$, is very powerful. This picture is confirmed by the results on the average distance between β and $\widehat{\beta}$.

In the case of AOs the estimation model with an unrestricted dummy is not well suited. To remove the isolated outlying observations the estimation model introduces a shift in the levels of the non-stationary directions. The tests are highly over-sized, with a rejection frequency of 60% for 6 AOs, and the average precision is extremely low. With $T = 200$ the size is only down to around 30%. This clearly illustrates that it is not recommendable in practice just to insert innovational impulse dummies for observations with large residuals.

Estimation model $\mathcal{M}_3$. One possibility for modelling the additive outliers is the specification (6), where the non-linear restrictions (7) and (8) are not imposed; $\mathcal{M}_3 : \mu_{it} = \alpha\theta_{i0}D_{it-1}+\theta_{i1}d_{it}+\theta_{i2}d_{it-1}$.[11] This model can still be estimated by standard RRR.

When the DGP contains IOs results are poor and for an increasing number of outliers, the size approaches 100%. This reflects an additional contributions to the Trace test statistic from imposing the reduced rank restrictions also on the dummy terms when they are unrestricted in the DGP, see also Doornik, Hendry, and Nielsen (1998). Note that the distortion does not disappear for $T \to \infty$, see also the results for $T = 200$.

When AOs are present, $\mathcal{M}_3$ is the correct model apart from the restrictions (7) and (8) not imposed, but the model is costly in terms of degrees of freedom. In longer samples or with a limited number of outliers the size and power are reasonable and there is a small gain compared to the basic model $\mathcal{M}_0$. But in small samples and with several outliers the loss in degrees of freedom is prohibitive[12].

Estimation model $\mathcal{M}_4$. In column five we introduce a general model where the dummies are included with the full lag structure of the autoregressive model and unrestricted coefficients; $\mathcal{M}_4 : \mu_{it} = \theta_{i0}D_{it} + \theta_{i1}D_{it-1} + \theta_{i2}D_{it-2}$. A similar model is applied in Vogelsang (1999) for univariate Dickey-Fuller unit root tests. This model is interesting because it is (in principle) robust to both IOs and AOs and can be estimated using RRR. The model is, however, potentially costly in terms of degrees of freedom and this inefficiency is important in small samples. In the case of IOs the properties are marginally better than for the basic specification, $\mathcal{M}_0$, but is clearly inferior to the correct innovational specification, $\mathcal{M}_2$. For the case of AOs the specification is comparable to ignoring the outliers in small samples and implies a small gain in larger samples[13].

A simple refinement of $\mathcal{M}_4$ is to estimate the general model and then to delete insignificant dummies to save degrees of freedom. This is applied in model $\mathcal{M}_{4b}$ in columns six. More specifically we delete a dummy if all four $t-$test statistics (for the hypotheses of dummy coefficients

[11] For this model we base the critical values on the mean and variance for the basic model augmented with the moments of n independent $\chi^2(p-r)$ distributions, cf. Section 3.

[12] An interesting feature of this model is that the same test statistics are obtained independent of the magnitude of the AOs. The important thing is not the magnitude of the actual outliers, but the presence of the dummies in the estimation model.

[13] Again we can note, that the same test statistics would have been obtained irrespectively the magnitude of the AOs, because all contributions to the likelihood function involving the outlying observations are set to zero by the included unrestricted dummy variables.

equal zero) are lower than the conventional 1.96. For the case of IOs $\mathcal{M}_{4b}$ is closer to the true specification, $\mathcal{M}_2$, in terms of the average distance. For the case of AOs the results are the opposite, because the additional zero restrictions imposed in $\mathcal{M}_{4b}$ do not bring the model closer to the correct additive specification.

The broader conclusion of the results from model $\mathcal{M}_1 - \mathcal{M}_4$ seems to be that it is not really possible in small samples to approximate AOs with simple combinations of unrestricted dummies.

Estimation model $\mathcal{M}_5$. In columns seven, we finally consider the exact additive specification; $\mathcal{M}_5 : \mu_{it} = A(L)\theta_{i0}D_{it}$, estimated with the switching algorithm.

In the presence of IOs results are comparable to ignoring the outlier. This illustrates that IOs can not be approximated by additive dummies.

In the case of AOs the estimation model is identical to the DGP. Because the outlying observations are completely removed from the likelihood function, the size and the power are by and large similar to the benchmark case with no outliers - and they are independent of the magnitude of the outliers. There is no potential gain from additive outliers, but using the correct model it is possible to conduct inference conditional on additive outlying observations.

5.2 Feasible Strategies

In the specifications $\mathcal{M}_0 - \mathcal{M}_5$ we assumed that the locations of outliers were known. In this subsection we extend the analysis to strategies for modelling outliers, which are feasible in empirical applications. This introduces the additional cost of detecting the locations of outliers.

First, we consider for comparison a stylized version of the *usual practice* in applied cointegration analyses; i.e. the use of unrestricted impulse dummies for observations with large residuals. In particular, we define an outlier as an observation with an absolute value of the standardized residual larger than 3.39.[14] For these observations we insert an unrestricted (innovational) dummy, $D_t(T_i)$. The results are reported under $\mathcal{M}_6$ in column eight and are not far from the results of $\mathcal{M}_2$. The additional cost of searching for the location of outliers thus seems to be small, but the strategy is disastrous if the outliers are additive.

Next we consider a parallel to $\mathcal{M}_{4b}$ in the case of unknown location. Again we define outliers as absolute residuals larger than 3.39 and insert unrestricted dummies with the full lag structure, $\mathcal{D}_t = (D_t(T_i) : D_{t-1}(T_i) : D_{t-2}(T_i))'$. From this model we delete insignificant dummies based on standard $t-$tests. The results are reported under $\mathcal{M}_7$. Again the search *per se* is not problematic, but the unrestricted dummies are still misspecified in the case of AOs and the overall gain compared to ignoring the outliers is questionable in small samples, cf. also panel (C).

Finally, we consider under $\mathcal{M}_8$ in column ten the results from applying the formal outlier detection from Section 3.2.[15] In the case of IO the size is relatively constant around 10% and

[14] The critical value corresponds to a multi-comparison significance level $\delta^{\max}$ for a nominal 5% and $T = 75$ observations in a standard normal distribution.

[15] To minimize the computational burden we only calculate the test for an outliers at a given point in time if the numerical value of the standardized residual is larger than 2.

the power for the test of $r \leq 1$ is high. The results are not far from the correct model $\mathcal{M}_2$, cf. also the average distance in panel (C). In case of AOs the size is increasing with the number of outliers. This reflects that in small samples AOs are sometimes mistaken for IOs and that introduces a distortion similar to one driving the results the for $\mathcal{M}_2$. In larger samples, e.g. $T = 200$, this effect is small. In terms of the average distance the results are between the optimal result of $\mathcal{M}_5$ and the basic model $\mathcal{M}_0$. For $T = 200$ the results are very close to the results of the correct models both in case of IOs and AOs.

6 Summary and Concluding Remarks

In this paper we have considered the role of outliers and dummy variables in the cointegrated VAR. First, it was illustrated that inference can be distorted by additive outliers while innovational outliers are less harmful and can actually be helpful in revealing the autoregressive structure. Second, it was shown that inference based on a misspecified specification of dummy variables can be seriously misleading. And third, it was very difficult to approximate additive outliers with innovational dummies. In larger samples it is possible to use a general specification which nests both alternatives. But in small samples, none of the tried specifications was able to successfully approximate both IOs and AOs.

Taken together, these results seriously question the usual practice in applied cointegration analyses of including unrestricted dummy variables to *whiten* residuals – without concern for the adequacy of the implied innovational model. That led us to suggest an estimator for the additive model and to test for the location and types of outliers prior to inference in the cointegration model. The simulation results based on this procedure are clearly superior to the alternative feasible strategies.

An additional interesting result of the simulations was that it is clearly preferable to ignore possible outliers than to model them using a misspecified configuration of dummies. That suggests that the basic model with no dummies could always be a benchmark specification in empirical applications.

Appendix: A Switching Algorithm for the Additive Model

The log likelihood function of the additive model (6) subject to the non-linear restrictions (7) and (8) is (apart from a constant) given by

$$\begin{aligned} & \log L\left(\alpha,\beta,\Gamma_1,...,\Gamma_{k-1},\beta_0,\mu_0,\Omega,\theta\right) \\ = & -\frac{T}{2}\log\left|\Omega\right| - \frac{1}{2}\sum_{t=1}^{T}\left[\left(A\left(L\right)\left(Y_t-\theta\mathcal{D}_t\right)-\alpha\beta_0't-\mu_0\right)'\Omega^{-1}\left(A\left(L\right)\left(Y_t-\theta\mathcal{D}_t\right)-\alpha\beta_0't-\mu_0\right)\right], \end{aligned}$$

and that can be maximized by switching between two conditional maximum likelihood estimations. Consider iteration j. First, conditional on the estimate $\widehat{\theta}_{j-1}$ of θ from the previous iteration, the maximum likelihood estimates $\widehat{\varphi}_j$ of $\varphi=(\alpha,\beta,\Gamma_1,...,\Gamma_{k-1},\beta_0,\mu_0,\Omega)$ can be found from RRR of the cointegrated VAR for the corrected data $X_t=Y_t-\widehat{\theta}_{j-1}\mathcal{D}_t$. This enables us to construct the estimated polynomial $\widehat{A}\left(L\right)$. Secondly, we find the maximum likelihood estimates $\widehat{\theta}_j$ of θ conditional on $\widehat{\varphi}_j$ from the estimated residuals, $\widehat{e}_t=\widehat{A}\left(L\right)Y_t-\widehat{\alpha}\widehat{\beta}_0't-\widehat{\mu}_0$, which under the model (4) and (5) are given by

$$\widehat{e}_t=\widehat{A}\left(L\right)\theta\mathcal{D}_t+\epsilon_t.$$

For each dummy, $\mathcal{D}_{it}$, $i=1,2,...,n$, we can define $\widehat{H}_{it}\equiv\widehat{A}\left(L\right)\mathcal{D}_{it}$ and use the matrix $\widehat{H}_t=(\widehat{H}_{1t}:\widehat{H}_{2t}:...:\widehat{H}_{nt})$ to rewrite the equation for the estimated residuals as

$$\widehat{e}_t=\widehat{H}_t vec\left(\theta\right)+\epsilon_t,$$

where $vec\left(\theta\right)$ stacks the columns of θ. The varying part of the conditional log likelihood function is now given by

$$\log L\left(\theta\left|\widehat{\alpha},\widehat{\beta},\widehat{\Gamma}_1,...,\widehat{\Gamma}_{k-1},\widehat{\beta}_0,\widehat{\mu}_0,\widehat{\Omega}\right.\right)=-\frac{1}{2}\sum_{t=1}^{T}\left[\left(\widehat{e}_t-\widehat{H}_t vec\left(\theta\right)\right)'\widehat{\Omega}^{-1}\left(\widehat{e}_t-\widehat{H}_t vec\left(\theta\right)\right)\right],$$

which is maximized over θ by the GLS type estimator

$$vec\left(\widehat{\theta}_j\right)=\left(\sum_{i=1}^{T}\left(\widehat{H}_t'\widehat{\Omega}^{-1}\widehat{H}_t\right)\right)^{-1}\left(\sum_{i=1}^{T}\left(\widehat{H}_t'\widehat{\Omega}^{-1}\widehat{e}_t\right)\right), \tag{12}$$

see also Tsay, Peña, and Pankratz (2000) and Saikkonen and Lütkepohl (2000) for an application of a similar two step strategy. The ML estimates are obtained by iterating between the two steps until convergence.

For the case of dummy variables involving few observations the switching algorithm normally converges very fast and we can use $\widehat{\theta}_0=0$ as a starting value. In more complicated situations an initial estimate of θ ignoring the non-linear restrictions (7) and (8) can be used. A comparison between the switching algorithm and the direct maximization indicates that the switching algorithm is normally faster and more reliable[16] and the closed form solution to the maximization in each step makes the procedure easy to implement.

[16]The switching algorithm has been compared with direct maximization using the BFGS algorithm (with numerical derivatives) in Ox 3.0, see Doornik (2001). More information on the comparison is available from the author on request.

It is straightforward to impose restrictions on θ in the GLS step (12). Since θ enters vectorized, restrictions of the form

$$vec\,(\theta) = M\kappa$$

can be considered allowing cross-equation restrictions. Here M is a $pn \times f$ dimensional design matrix, and κ contains the f free parameters. Often it is of interest to test if additive outliers are present in a subset of the variables only, implying zero restrictions on θ. Another hypothesis of interest could be if the outliers cancel in the stationary relations, corresponding to $M = I_n \otimes \beta_\perp$, with β replaced by the estimate $\widehat{\beta}_j$ in each iteration. Without complications, restrictions can also be imposed on the RRR part of the algorithm.

References

Chen, C., and L.-M. Liu (1993): "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, 88(841), 284–297.

Davidson, J. (2001): *Econometric Theory*. Blackwell, Oxford.

Doornik, J. A. (1998): "Approximations to the Asymptotic Distribution of Cointegration Tests," *Journal of Economic Surveys*, 12(5), 573–593.

——— (2001): *Object-Oriented Matrix Programming Using Ox*. Timberlake Consultants Press, London, 4rd edn.

Doornik, J. A., D. F. Hendry, and B. Nielsen (1998): "Inference in Cointegrating Models: UK M1 Revisited," *Journal of Economic Surveys*, 12(5), 533–572.

Fox, A. (1972): "Outliers in Time Series," *Journal of the Royal Statistical Society, Series B*, 34(3), 350–363.

Franses, P. H., and N. Haldrup (1994): "The Effect of Additive Outliers on Tests for Unit Roots and Cointegration," *Journal of Business and Economic Statistics*, 12(4), 471–478.

Gomez, V., A. Maravall, and D. Peña (1999): "Missing observations in ARIMA models: Skipping approach versus additive outlier approach," *Journal of Econometrics*, 88, 341–363.

Hendry, D. F., and J. A. Doornik (1994): "Modelling Linear Dynamic Econometric Models," *Scottish Journal of Political Economy*, 41(1), 1–33.

Hendry, D. F., and K. Juselius (2001): "Explaining Cointegration Analysis: Part II," *Energy Journal*, 22(1), 75–120.

Hendry, D. F., and G. Mizon (1993): "Evaluating Dynamic Econometric Models by Encompassing the VAR," in *Models, Methods and Applications of Econometrics*, ed. by P. Phillips, pp. 272–300. Blackwell, Oxford.

Johansen, S. (1996): *Likelihood-based Inference in Cointegrated Autoregressive Models*. Oxford University Press, Oxford, 2nd edn.

Johansen, S., R. Mosconi, and B. Nielsen (2000): "Cointegration in the presence of structural breaks in the deterministic trend," *Econometrics Journal*, 1(3), 216–249.

Juselius, K. (2002): *The Cointegrated VAR model: Econometric Methodology and Macroeconomic Applications*. Lecture notes, Institute of Economics, University of Copenhagen.

Larsson, R., and M. Villani (2001): "A distance measure between cointegration spaces," *Economics Letters*, 70, 21–27.

Muirhead, C. (1986): "Distinguishing Outlier Types in Time Series," *Journal of the Royal Statistical Society, Series B*, 48(1), 39–47.

Nielsen, B., and A. Rahbek (2000): "Similarity Issues in Cointegration Analysis," *Oxford Bulletin of Economics and Statistics*, 62(1), 5–22.

Pollak, R. A., and T. J. Wales (1991): "The likelihood dominance criterion A new approach to model selection," *Journal of Econometrics*, 47, 227–242.

Saikkonen, P., and H. Lütkepohl (2000): "Trend Adjustment Prior to Testing for the Cointegration Rank of a VAR Process," *Journal of Time Series Analysis*, 21(4), 435–456.

Shin, D. W., S. Sarkar, and J. H. Lee (1996): "Unit root tests for time series with outliers," *Statistics & Probability Letters*, 30, 189–197.

Tsay, R. S. (1986): "Time Series Model Specification in the Presence of Outliers," *Journal of the American Statistical Association*, 81(393), 132–141.

Tsay, R. S., D. Peña, and A. E. Pankratz (2000): "Outliers in multivariate time series," *Biometrika*, 87(4), 789–804.

Vogelsang, T. J. (1999): "Two simple procedures for testing for a unit root when there are additive outliers," *Journal of Time Series Analysis*, 20(2), 237–252.