Conditional Inference in the Cointegrated Vector Autoregressive Model

Kees Jan van Garderen and Sophocles Mavroeidis

University of Amsterdam

June 24, 2004

Abstract

A Vector Autoregressive model (VAR) with normally distributed innovations is a Curved Exponential Model (CEM). Cointegration imposes further curvature on the model and this means that in addition to the important reasons for conditioning in non-stationary time series as given by Johansen (1995, EJ), there are further reasons due to the curvature of the model. This paper investigates the effects of conditioning on the likelihood ratio test statistic for the cointegrating rank, which in this case is a natural approximate ancillary statistic. We investigate the effect of conditioning on this test statistic for inference on the long-run (beta) and also on the speed-of-adjustment (alpha) coefficients. We show that this conditioning gives virtually the same estimates of the estimator variance as using the observed information instead of the expected information. We examine the possibility of achieving asymptotic refinements for inference on alpha using a conditioning parametric bootstrap procedure.

Keywords: Conditional inference, cointegration, VAR, curved exponential models.

JEL classification: C10 C32

1 Introduction

This paper considers conditional inference in a cointegration setting and the focus is on inference on the adjustment coefficients, usually denoted α. The motivation for conditioning here is the fact that

models for cointegration are curved in a statistical sense as defined by Efron (1975). The curvature of the model induces regions in the sample space where the inference is more difficult, see van Garderen (1995) and below and this provides a very direct argument for conditioning. There are various other reasons for conditioning and some have been taken up already in the cointegration literature, most prominently by Johansen in a number of articles.

Johansen (1995b) provides a very clear discussion of the role of conditioning on ancillary statistics with non-stationary data. He shows that likelihood based asymptotic inference can be conducted the same way for ergodic as for non-ergodic processes by conditioning, subject to strong exogeneity conditions. He also shows that in a number of important cases the conditional distribution is far simpler than the unconditional (marginal) distribution. Finally, he shows that the inverse of the observed information provides a better, and more relevant measure for the uncertainty of the estimator of the long run parameter than the inverse of the expected Fisher information and that the inverse of the observed information is an appropriate measure of the variance, not of the marginal distribution but of the conditional distribution of the estimator given the available information in the sample.

Sweeting (1992) actually shows conditional asymptotic normality of the Maximum Likelihood Estimator (MLE) in a general setting when it is scaled by the random norm. There are some conditions, but these explicity allow for non-ergodic processes.

One important reason for the acceptance of conditioning in the cointegration literature is that the information matrix in the non-ergodic case is itself a random variable; the observed information scaled by T^{-2} does not go to a fixed limit but converges toa random variable. Depending on the realised sample-path there will either be very little information in the data, or a large amount of information. Inference procedures should take this into account.

This is probably one of the reasons that Johansen (1995a) focusses on the long run parameter. The adjustment coefficients α converge at the usual rate and the information matrix block corresponding to α is $O_p(T)$ and approprately scaled does converge to its asymptotic expectation.

Conditioning is also used in a number of articles on small sample corrections for tests in the cointegrating space. Johansen (2002b) considers Bartlett corrections for likelihood ration tests on the cointegrating rank and Johansen (2002a) considers Bartlett corrections for likelihood ratio tests on

the cointegrating vector. Conditioning on the common trends is actually used as a technical device to derive the correction factors.

Hansen and Rahbek (2002) in the related context of testing for unit roots, consider conditioning essentially to get rid of nuisance parameters. They use a Cox and Reid (1987) type adjustment of the likelihood ratio test based on orthogonalizing the parameters.

There are a number of similarities between the Cointegrating Vector Autoregressive (CVAR) model and the Single Structural Equation Model (SSEM). The way in which the MLE is calculated involves solving an eigenvalue problem in both cases. Moreover, cointegration in the VAR plays a similar role to overidentification in the SSEM, since both are rank restrictions on the coefficient matrices. Another feature they share is that the number of parameters is less than the number of sufficient statistics implying that they are both curved exponential models(see Hosoya, Tsukuda, and Terui 1989 and van Garderen 1997). This implies that in both models maximum likelihood estimation involves a dimensional reduction of a statistic which contains all the sample information to the parameter estimate which therefore can no longer contain all the information. This information can be recovered, in certain circumstances by conditioning on an appropriate ancillary statistic, if one exists. This idea was studied in the context of the SSEM by Hosoya, Tsukuda, and Terui (1989). They found that the distribution of the Limited Information Maximum Likelihood (LIML) estimator depends on the smallest characteristic root associated with LIML estimation. In the paper we investigate the effect of conditioning on the analogous statistic in the CVAR model, namely, the likelihood ratio test for cointegration, also known as the trace test. We actually find a much stronger effect of conditioning in the CVAR than those found for the SSEM.

One difference, however, is that when the single equation is exactly identified, the model is a full exponential model, whereas the VAR without rank restrictions is still a curved exponential model.

2 The Model

Consider a simple first order bivariate vector autoregressive (VAR) model in error correction form

$$\Delta Y_t = \Pi Y_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, T \tag{1}$$

where ε_t are zero mean independently normally distributed disturbances with contemporaneous covariance matrix Ω. For simplicity we will assume that Ω is known throughout and can therefore be set equal to the identity. The process is stable when the eigenvalues of the 2×2 matrix $(I_2 + \Pi)$ are inside the unit circle. If exactly one of the eigenvalues is unity, the matrix Π is of reduced rank and the model becomes a cointegrated VAR (CVAR). Because the rank of Π equals 1, we can write $\Pi = \alpha\beta'$ where α and β are 2-dimensional vectors. The vector β is known as the cointegrating vector with the property that $\beta' Y_t$ is a stable process which defines an equilibrium relationship between the variables in Y_t. The adjustment vector α describes the reaction of the system to last period's disequilibrium $\beta' Y_{t-1}$. The equilibrium space is a one dimensional space orthogonal to β called the attractor set which is spanned by $\beta^{\perp}$.

It is clear that any muliple of β would define the same equilibrium since the orthogonal space would be unchanged, and the only effect is that the corresponding α is reduced by the same factor, thereby leaving Π unchanged. It is clear that α and β are not identified and α cannot simply be interpreted as the speed of adjustment. We can think of α and β in terms of their angles φ and ϕ, relative to horizontal axis, for instance, and their length. Their angles are unique (modulo π) but their lengths are not. It is only the product of their length which is uniquely defined and coincides with the non-zero singular value of Π.

Let ρ be the eigenvalue of $(I_2 + \Pi)$ that is inside the unit circle. Then $\rho = 1 + \alpha'\beta$ and describes the memory of the disequilibrium, since

$$\beta' Y_t = \rho \beta' Y_{t-1} + \beta' \varepsilon_t \tag{2}$$

Another key quantity is λ_1 which relates one-to-one to the maximal canonical correlation between ΔY_t and $\beta' Y_{t-1}$ as follows

$$\lambda_1 = \frac{\rho_c^2}{1 - \rho_c^2}$$

and to the largest singular value μ_1 of Π and ρ as:

$$\lambda_1 = \frac{\mu_1^2}{1 - \rho^2}.$$

This λ_1 is the probability limit of the largest solution to an eigenvalue problem in the estimation procedure, see Equation (18) below. N.b. the second eigenvalue in this problem has probability limit equal to zero.

The *ex-ante* properties of the process are conveniently described in terms of the angle $\psi = \phi - \varphi$ between α and β and the quantities ρ and λ_1.

Suppose the process in period $t-1$ is in disequilibrium, i.e. $\beta' Y_{t-1} \neq 0$. Conditional on this value $\beta' Y_{t-1}$ the process is expected to move by $\alpha\beta' Y_{t-1}$ along α, call this $\Delta Y^e_{t|t-1} = E\left[\Delta Y_t | Y_{t-1}\right]$. The λ_1 is the expected squared distance that the process covers towards equilibrium $E\left[\left(\Delta Y^e_{t|t-1}\right)' \Delta Y^e_{t|t-1}\right]$, where the expectation is over Y_{t-1}.

From Equation (2) it is obvious that when $\rho = 0$ the disequilibrium in the next period, $\beta' Y_t$, is expected to be 0. This implies that within one period the process returns to equilibrium. More generally, let Y^*_t denote the projection of Y_{t-1} onto the equilibrium set along α (see Figure 1), thus

$$Y^*_t = \beta_\perp (\alpha'_\perp \beta_\perp)^{-1} \alpha'_\perp Y_{t-1},$$

and let $\Delta Y^*_{t|t-1} = Y^*_t - Y_{t-1}$ which is the total distance the process would need to cover along α in order to get back to equilibrium. Then,

$$1 - \rho = \frac{\left\| \Delta Y^e_{t|t-1} \right\|}{\left\| \Delta Y^*_{t|t-1} \right\|}$$

which is the proportion of the necessary total adjustment $\Delta Y^*_{t|t-1}$ that is expected to take place. So, for $0 < \rho < 1$, the adjustment to equilibrium is partial, and for $-1 < \rho < 0$ the process overcorrects in response to the disequilibrium. In the limiting case of $\rho = 1$, there is no adjustment to any equilibrium, which can arise either due to the absence of cointegration (when $\Pi = 0$), or because the process is integrated of order 2 ($\Pi \neq 0$).

The angle ψ determines the efficiency of the expected return to equilibrium, in the following sense. When ψ is equal to 180° the total distance to equilibrium, $\Delta Y^*_{t|t-1}$, is the smallest possible because the adjustment path is orthogonal to the equilibrium space, see Figure 1.

The Engle and Granger (1987) representation for Y_t is

$$Y_t = \beta_\perp \left(\alpha'_\perp \beta_\perp\right)^{-1} \sum_{i=1}^{t} \alpha'_\perp \varepsilon_i + \alpha \left(\beta' \alpha\right)^{-1} \sum_{i=0}^{t-1} \rho^i \beta' \varepsilon_{t-i} \tag{3}$$

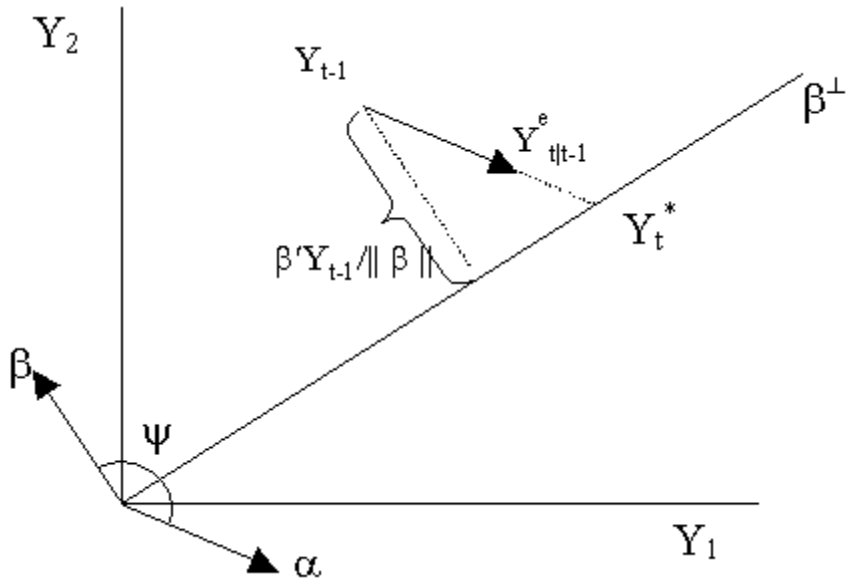


Figure 1: The dynamics of the bivariate CVAR.

where $\alpha_\perp$ and $\beta_\perp$ are orthogonal to α and β respectively, and where we have set the initial value Y_0 to 0. This representation is particularly useful in our present discussion, because it highlights the distinction between the evolution of Y_t along the equilibrium space determined by the common stochastic trend

$$\sum_{i=1}^{t} \alpha_\perp' \varepsilon_i, \tag{4}$$

and the evolution around the equilibrium space, as measured by the disequilibria

$$\beta' Y_t = \sum_{i=0}^{t-1} \rho^i \beta' \varepsilon_{t-i}. \tag{5}$$

By the Granger representation theorem it follows that $Y_t^* = \beta_\perp \left(\alpha_\perp' \beta_\perp\right)^{-1} \sum_{i=1}^{t-1} \alpha_\perp' \varepsilon_i$. Moreover it also follows that every shock ε_t can be decomposed into a permantent shock $\alpha_\perp' \varepsilon_t$ and a transitory shock $\alpha' \varepsilon_t$.

2.1 Post Sample Inference

We now turn to the problem making inference on the parameters on the basis of a sample $\{Y_t\}_{t=1}^T$. The theoretical properties of the system, as discussed in the previous section, determine the type of sample paths that are likely to be observed. *Ex-ante*, i.e. before any given sample path is observed,

the accuracy of any estimator is determined by averaging over all possible paths. *Ex-post,* however, after the sample has been realized, it may turn out that the observed sample path is more or less informative on the parameters than expected *ex-ante*, for the given parameter values and sample size. In this section we give some heuristic discussion of post sample inference and in the next section we take a more analytical approach based on the likelihood function.

First, consider estimation of β. There are two aspects of the data that govern the accuracy with which β can be estimated: the dispersion along the equilibrium set and the dispersion around it. The first one increases and the second reduces the accuracy.

If the observations are widely dispersed *along* the equilibrium set, as measured by cumulated variation of the common trends (4), then we can very accurately determine the equilibrium relationship defined by β (slope of the attractor set in Figure 1). If there is very little common trend variation, for instance because the obervations are evenly spread around one particular equilibrium point, then it is very difficult to determine β.

The other case where we can estimate β accurately is when the dispersion *around* the equilibrium set is small. In contrast, if the actual disequilibria are large, then β is less accurately estimated. This, however, is a second order effect relative to the dispersion along the equilibrium set, which relates to the superconsistency of the MLE for β. Johansen (1995b, Section 6) provides a clear discussion of these points.

Two contrasting cases are shown in Figure 2, which plots Y_1 against Y_2 for two realizations of the process with identical parameter values and sample size ($T = 50$). One sample is very informative about the equilibirum relationship and the other very uninformative.

Similarly, there are also two aspects of the sample that determine the accuracy with which α is estimated: the dispersion around the equilibrium set and the accuracy of the estimator for β.

In the extreme hypothetical situation when we only observe the economy in equilibrium i.e. $\beta' Y_{t-1} = 0$ for all t, then it is impossible to determine the disequilibrium adjustment coefficient α. Conversely, when the realized disequilibria are large, α can be estimated accurately. To demonstrate this idea consider the case where β is known and α contains the slopes of the regression of ΔY_t

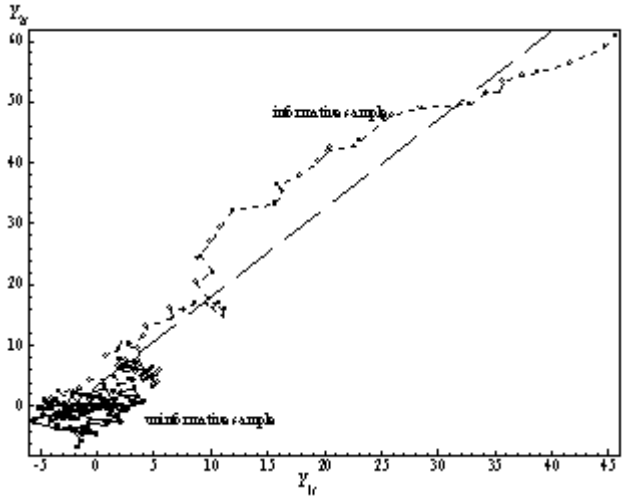


Figure 2: Scatter plot of Y_{2t} on Y_{1t} for two different samples of size $T = 50$, drawn from the same CVAR model with parameters $\alpha = (-.26, .16)'$ and $\beta = (1, 1)'$ (thus, $\rho = 0.9$ and $\lambda_1 = 1$).

on $\beta' Y_{t-1}$. Figure 3 plots the change in the first element of ΔY_t against the disequilibrium $\beta' Y_{t-1}$ and shows one sample that is very informative on the adjustment coefficient α_1 and the other not being very informative.

More generally, β is unknown, and we can think of estimating α by first estimating β, which can be done superconsistently, and then regressing ΔY_t on the generated regressor $\hat{\beta}' Y_{t-1}$. When $\hat{\beta}$ is the MLE, the OLS estimator of α in the regression just described, is also the MLE. Replacing $\beta' Y_{t-1}$ with $\hat{\beta}' Y_{t-1}$ induces a measurement error in the regression determining α and this error causes additional variation in the distribution of $\hat{\alpha}$.

So for α we see two effects of the variation in the disequilibrium terms: a direct effect which is positive and an indirect negative effect caused by the less accurate estimation of β.

It is also interesting to note the asymmetry in the estimation of α and β. It is impossible to estimate α accurately without an accurate estimate of β. It is perfectly possible, however, to estimate β very accurately without an accurate estimate of α. In fact, the most accurate estimate of β is obtained when there is no disequilibrium in the system and α cannot be estimated at all.

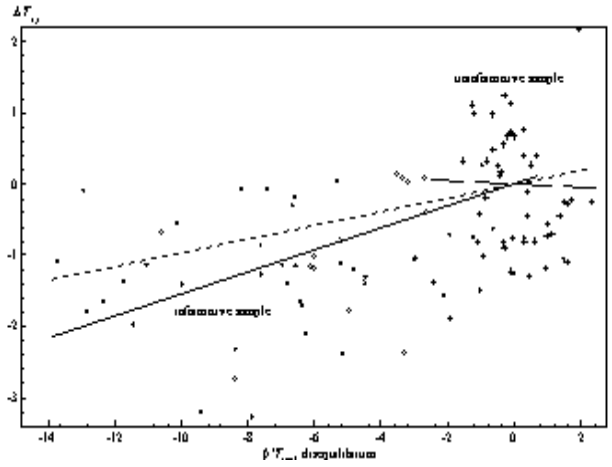


Figure 3: Scatter plot of ΔY_{1t} on $\beta' Y_{t-1}$, for two different samples of size $T = 50$, drawn from the same CVAR model with parameters $\alpha = (-.26, .16)'$ and $\beta = (1, 1)'$ (thus, $\rho = 0.9$ and $\lambda_1 = 1$).

3 Curved Models

The CVAR is embedded in a more general VAR model. The rank restrictions imposed on the Π matrix are non-linear and the CVAR is therefore a nonlinear subset of the embedding VAR. VARs are generally thought of as being linear because the conditonal mean of the process depends linearly on past values. In terms of their deeper mathematical structure VAR models are not linear because they are not linear exponential models. As was shown in van Garderen (1997) a VAR is a Curved Exponential Model (CEM) and is itself embedded in a larger linear- or Full Exponential Model (FEM). The log-likelihood function of an FEM admits the following canonical representation:

$$l(\eta) = \eta \cdot s - \kappa(\eta), \tag{6}$$

where $\eta \in \mathcal{H} \subseteq \mathcal{R}^k$ is the canonical parameter, $s \in \mathcal{S} \subseteq \mathcal{R}^k$ is the minimal sufficient statistic and $\cdot$ denotes an inner product. If η is genuinely k dimensional, then the model is an FEM. If η lies on a smooth manifold of lower dimension than d, the model is a CEM-(k, d). It will then be possible, at least locally, to write η as a differentiable function of a new parameter, θ say. In that case we have $\eta = \eta(\theta)$, but there is no dimensional reduction in the minimal sufficient statistic s, as long as $\eta(\cdot)$ is nonlinear. This is what characterizes CEMs, namely that the dimension of the minimal sufficient

statistic is larger than the number of parameters, as shown in van Garderen (1997) for dependent and non-identically distributed observations.

Turning to the VAR model (1), its likelihood is given by

$$l(\Pi) = -\frac{T}{2}\ln|\Omega| - \frac{T}{2}\,\mathrm{tr}\left(S_{00}\Omega^{-1} - 2S_{01}\Pi'\Omega^{-1} + S_{11}\Pi'\Omega^{-1}\Pi\right), \tag{7}$$

where $S_{00} = T^{-1}\sum_{t=1}^{T}\Delta Y_t \Delta Y_t'$, $S_{01} = T^{-1}\sum_{t=1}^{T}\Delta Y_t Y_{t-1}'$ and $S_{11} = T^{-1}\sum_{t=1}^{T} Y_{t-1}Y_{t-1}'$. It is clear from (7) that the model is a CEM, with $\eta = \left(\Omega^{-1}, \Pi'\Omega^{-1}, \Pi'\Omega^{-1}\Pi\right)$, $s = (S_{00}, S_{01}, S_{11})$ and inner product defined by the trace. It is clear that because of the symmetries involved, there are a number of redundant elements in η and s. The dimension k is 10 in the bivariate VAR ($n^2 + n(n+1)$ when Y_t is n-dimensional), while the number of free parameters is 7 ($n^2 + n(n+1)/2$). The difference in dimension is 3 ($n(n+1)/2$ in general). If Ω is known, as we are assuming throughout, the dimensions reduce to 7 and 4 for s and θ respectively.

Next, consider the CVAR, which imposes the restriction that the rank of Π is equal to 1, so that $\Pi = \alpha\beta'$. The log-likelihood (with Ω known and normalized to the identity) can be written as

$$l(\alpha, \beta) = -\frac{T}{2}\mathrm{tr}\left(-2S_{01}\beta\alpha' + S_{11}\beta\alpha'\alpha\beta'\right). \tag{8}$$

The dimension of the sufficient statistic remains the same, while the number of parameters is reduced to 3, and hence the bivariate CVAR is a CEM-(7,3).

It is worth mentioning at this stage that if impose the additional restriction that β is known, the model becomes a CEM-(3,2) and we can represent it graphically in full. In that case, the minimal sufficient statistic becomes $(S_{0\beta}, S_{\beta\beta})$, where $S_{0\beta} = S_{01}\beta$ and $S_{\beta\beta} = \beta' S_{11}\beta$. In contrast, when α is assumed to be known, the model becomes CEM-(5,2).

Consequences of $k - d > 0$ The difference in dimension between the minimal sufficient statistic and the number of parameters has two immediate consequences. Any estimator of the parameters is of lower dimension than that of any sufficient statistic s and cannot contain all the information. The mapping $s \mapsto \hat{\theta}$ is non-invertible but we could augment $\hat{\theta}$ with an auxilliary statistic a say, such that $s \mapsto \left(\hat{\theta}, a\right)$ is invertible. This statistic a could be used to recover information lost by the estimator through conditioning on a. This is one of the important classical arguments for conditioning.

Second, it is also clear that, because of the dimensional reduction in $s \mapsto \hat{\theta}$, there will be different values of s that give rise to the same value of $\hat{\theta}$. In principal we can invert the estimator and find the all the points in the sample space that result in the same value of the estimator $\hat{\theta}$. For the MLE we can easily characterize this inverted MLE by looking at the likelihood and its derivatives.

$$\underset{1\times k}{\frac{\partial l}{\partial \theta'}} = \underset{1\times k}{[s - \tau(\theta)]'} \underset{k\times d}{\frac{\partial \eta}{\partial \theta'}} \tag{9}$$

$$\mathcal{J}_\theta = -\frac{\partial^2 l}{\partial \theta \; \partial \theta'} = \mathcal{I}_\theta - \sum_{i=1}^{d} [s_i - \tau_i(\theta)] \frac{\partial^2 \eta_i}{\partial \theta \; \partial \theta'} \tag{10}$$

where $\tau = \partial\kappa/\partial\eta'$ is the expected value of s in the full embedding model and $\tau(\theta)$ is the value evaluated at θ for the CEM. $\mathcal{I}_\theta = -\frac{\partial \eta'}{\partial \theta} \frac{\partial^2 \kappa}{\partial \eta \; \partial \eta'} \frac{\partial \eta}{\partial \theta'}$ denotes the expected information w.r.t. θ and $\mathcal{J}$ is the observed information.

The MLE is found by setting the score (9) equal to zero. From this it is immediate that for fixed $\hat{\theta}$, and therefore $\tau(\hat{\theta})$ and $\partial\eta(\hat{\theta})/\partial\theta'$ also fixed, all points s in the sample space such that $s - \tau(\hat{\theta})$ is orthogonal to $\partial\eta(\hat{\theta})/\partial\theta'$ satisfy the first order conditions. This characterizes the inverted MLE.

From the second derivative we see that the quantity $s - \tau(\hat{\theta})$ determines (linearly) the difference between the observed and expected information. The expected information is always positive definite as long as the parameters are identified, but the observed information can be made singular by moving s along the inverted MLE. When the observed information, and hence the Hessian, is singular, the likelihood function is flat and has no unique maximum. The set of points in the sample space for which this happens is called the critical set. In a neighbourhood of this critical set the MLE will be more sensitive to small changes in s than for s close to its expected value $\tau(\theta)$. We refer to regions close to the critical set as the sensitive regions.

Note that $\mathcal{J}_\theta = \mathcal{I}_\theta$ for all s, if and only if the model is full. The existence of a non-empty critical set and the sensitive region is a direct consequence of the curvature of the model.

Proper inference procedures should take account of the fact that certain samples may fall in a region where the MLE is more sensitive. This is a direct argument for conditioning on statistics which indicate proximity of the observed sample to the critical set. The objective of this paper is to identify statistics that can be used for this purpose.

3.1 Inference on α when β known

To illustrate these ideas in the cointegration setting, consider the simplest possible framework where β is assumed known. As we have already shown, this is a CEM-(3,2) with canonical parameter $\eta = (\alpha_1, \alpha_2, (\alpha_1^2 + \alpha_2^2)/2)'$ and corresponding canonical statistic $s = \left(S_{0\beta}', S_{\beta\beta}\right)'$, and we can illustrate the ideas above graphically. The expectation of s is (see Appendix):

$$\tau(\alpha) = E \begin{pmatrix} S_{0\beta} \\ S_{\beta\beta} \end{pmatrix} = \frac{\beta'\beta \left[\left(1-\rho^2\right) - \left(1-\rho^{2T}\right)/T\right]}{\left(1-\rho^2\right)^2} \begin{pmatrix} \alpha \\ 1 \end{pmatrix} \tag{11}$$

where $\rho = 1 + \beta'\alpha$. The MLE equals

$$\widehat{\alpha}\left(\beta\right) = S_{0\beta} S_{\beta\beta}^{-1}. \tag{12}$$

The observed information equals

$$\mathcal{J}_\alpha = T \; S_{\beta\beta} \; I_2 \tag{13}$$

which is singular when $S_{\beta\beta}$ is zero (the probability on this event is zero, but $S_{\beta\beta}$ close to 0 are possible). This only happens when $\beta' Y_{t-1} = 0$ for all t, and hence $S_{0\beta} = 0$. The critical set is therefore the origin in the 3-dimensional sample space of the sufficient statistics. The interpretation of this event is of course that there are no deviations from equilibrium at all.

The expected information, using the expectation of $S_{\beta\beta}$ is

$$\mathcal{I}_\alpha = T \; \frac{\left[\left(1-\rho^2\right) - \left(1-\rho^{2T}\right)/T\right]}{\left(1-\rho^2\right)^2} \beta'\beta \; I_2 \tag{14}$$

The difference between the observation s and its expected value given the estimated parameter value equals

$$s - \tau\left(\hat{\theta}\right) = \begin{pmatrix} S_{0\beta} \\ S_{\beta\beta} \end{pmatrix} - \frac{\beta'\beta \left[\left(1-\hat{\rho}^2\right) - \left(1-\hat{\rho}^{2T}\right)/T\right]}{\left(1-\hat{\rho}^2\right)^2} \begin{pmatrix} \hat{\alpha} \\ 1 \end{pmatrix} = d \cdot \begin{pmatrix} \hat{\alpha} \\ 1 \end{pmatrix} \tag{15}$$

where the last equality follows from the fact that $s - \tau\left(\hat{\theta}\right)$ must be proportional to the orthogonal complement of $\partial\eta(\hat{\theta})/\partial\theta'$, since by the first order conditions of the MLE $\left(s - \tau\left(\hat{\theta}\right)\right)' \partial\eta(\hat{\theta})/\partial\theta' = 0$,

$$\underset{3\times 2}{\frac{\partial\eta}{\partial\theta'}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -\alpha_1 & -\alpha_2 \end{pmatrix}, \text{ with orthogonal complement } \left(\frac{\partial\eta}{\partial\theta'}\right)^\perp \propto \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ 1 \end{pmatrix}$$

For a fixed $\hat{\theta}$, this characterizes all the s that give the same value for the MLE.

From (15) it is straightforward to see that the proportionality factor $d = S_{\beta\beta} - \beta'\beta/\left(1-\hat{\rho}^2\right)$. The total distance from $\tau\left(\hat{\theta}\right)$ to the critical set, $(0,0,0)\prime$, for that particular $\hat{\theta}$ is $\beta'\beta/\left(1-\hat{\rho}^2\right)$ and the observation is therefore a proportion

$$rd = \frac{\left(1-\hat{\rho}^2\right)^2}{\left(1-\hat{\rho}^2\right) - \left(1-\hat{\rho}^{2T}\right)/T} \frac{S_{\beta\beta}}{\beta'\beta} \tag{16}$$

of this total distance away from the critical set. The sensitive region is when rd is close to zero and the difference between $S_{\beta\beta}$ and its expectation is negative.

The statistic rd can also be interpreted as measuring the *ex-post* variability of the disequilibrium $S_{\beta\beta}/\beta'\beta$ relative to what would be expected *ex-ante* for the *estimated* value of α, which equals $\left[\left(1-\hat{\rho}^2\right) - \left(1-\hat{\rho}^{2T}\right)/T\right]/(1-\hat{\rho}^2)^2$. Thus when the variability in the sample is higher than expected, there is more information on α and α is more accurately estimated.

Figure 4 plots the ratio of the conditional over the unconditional variance of α_1 and shows that rd has a profound effect on the accuracy of the MLE for α_1.

The figure shows a number of interesting facts. The first thing to notice is that the variance of $\hat{\alpha}$ depends heavily on rd. For small rd, the variance can three times larger than for large rd and more than 30% larger than the unconditional variance. This means that if one is always reporting the marginal variance

Next, we turn to hypothesis testing on the coefficients α. We consider a point null hypothesis $H_0 : \alpha = \alpha_0$ against a two-sided alternative $H_1 : \alpha \neq \alpha_0$. This has the convenient property that there are no nuisance parameters under the null (when β and Ω are known, of course), so an exact test can be constructed by Monte Carlo simulation. We compare two alternative tests of this hypothesis: (i) a Wald test based on the expected information (14), denoted $W_{\exp}(\alpha_0;\beta)$; and (ii) a Wald test based on the observed information (13), denoted $W_{obs}(\alpha_0;\beta)$. These are derived in the appendix, where we also show that $W_{obs}(\alpha_0;\beta)$ is equal to the Likelihood ratio test in this case. It is also shown in the appendix that $W_{obs}(\alpha_0;\beta)$ is invariant w.r.t. changes the parameter λ_1, it only varies with ρ.

Starting from $W_{\exp}$, Figure 5 plots the 10% critical value of the test statistic conditional on rd, and compares this with the exact marginal (unconditional) critical value and the associated critical

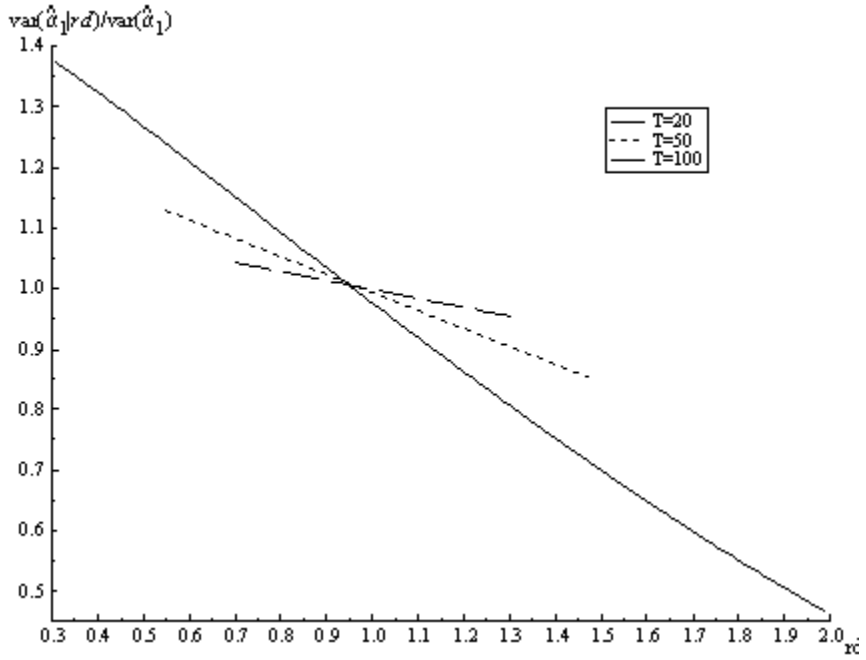


Figure 4: Variance of the MLE of α_1 in a bivariate CVAR model with known β, conditional on the value of the rd statistic, relative to its unconditional variance. The true parameters are $\alpha = (-0.26, 0.16)'$ and $\beta = (1, 1)'$. Variances computed using a nonparametric Nadaraya-Watson estimator based on 10^5 Monte Carlo replications.

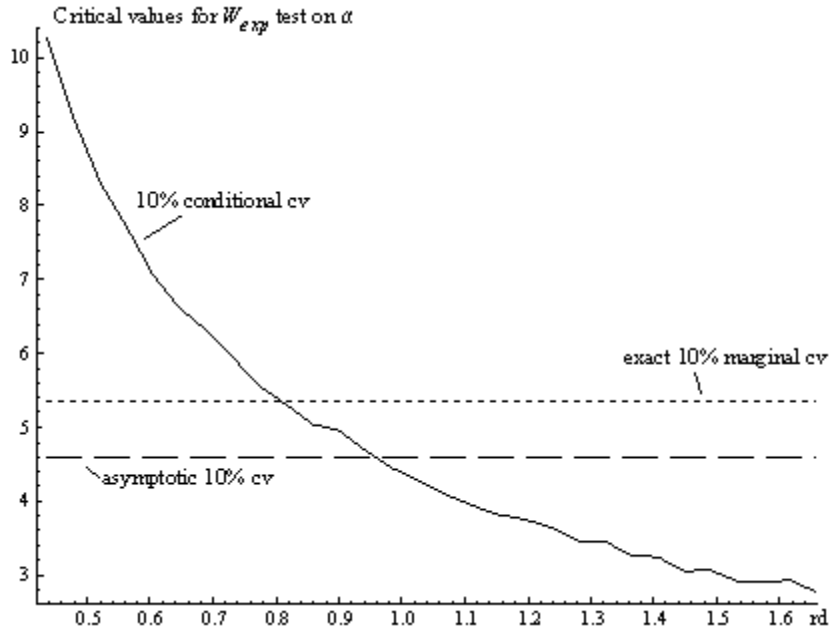


Figure 5: Three different sets of critical values for the W_{exp} test of the hypothesis $H_0 : \alpha = \alpha_0$ when β is known are plotted against the rd statistic. The parameters are $\alpha_0 = (-.19, -.11)'$, $\beta = (1, 1)'$, hence $\rho = 0.7$. The sample size is $T = 20$. The dashed line is the asymptotic critical value – the 90% quantile of the $\chi^2(2)$ distribution. The dotted line gives the (simulated) exact 90% quantile of the W_{exp} statistic under the null. The continuous line gives the estimated conditional 90% quantile of the W_{exp} statistic given the rd statistic. This is estimated by simulation using 400000 replications split over a set of 30 non-overlapping equally spaced grids.

value based on the $\chi^2(2)$ asymptotic approximation. [ADD COMMENTS]. Figure plots the *conditional* rejection frequency under the null hyphothesis (NRF) of the W_{exp} statistic when using the the asymptotic critical value.

Next, we turn to W_{obs} and plot its critical values and null rejection frequencies in Figures 7 and 8 respectively. Adjusting for scale, we see a completely different picture than for W_{exp}, namely, the effect of conditioning is almost negligible. In other words, using the observed, as opposed to the expected information results in much more reliable inference.

The above analysis relied on the assumption that β is known. This was mainly motivated by the need to simplify the exposition of the key ideas, but it admits two more justifications. One

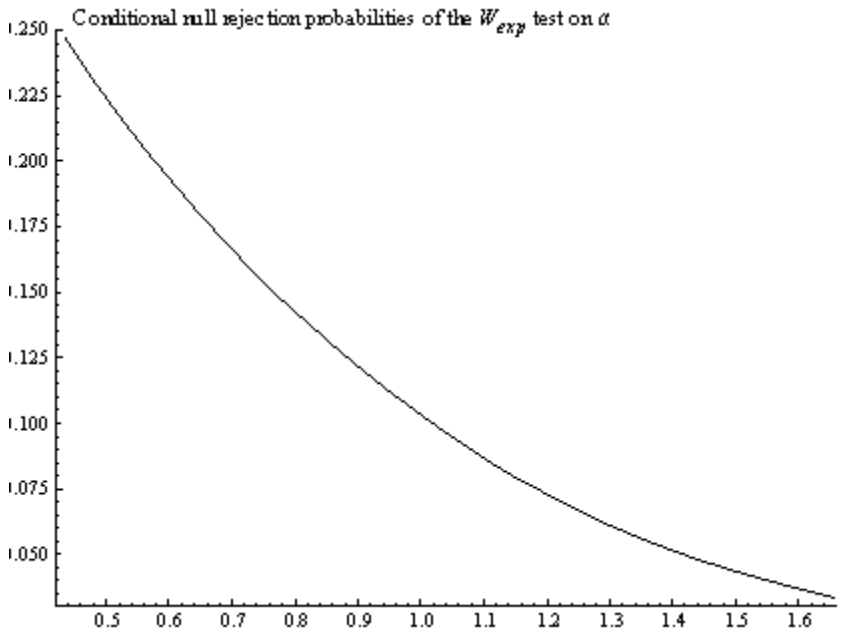


Figure 6: This graph plots the rejection frequency of the $W_{\exp}$ test of the hypothesis $H_0 : \alpha = \alpha_0$ (when β is known) under the null conditional on the rd statistic. The parameters are $\alpha_0 = (-.19, -.11)'$, $\beta = (1, 1)'$, hence $\rho = 0.7$. The sample size is $T = 20$. The conditional rejection frequency is estimated by means of a non-parametric Gaussian kernel estimator based on 400000 replications of data from the CVAR model under the null.

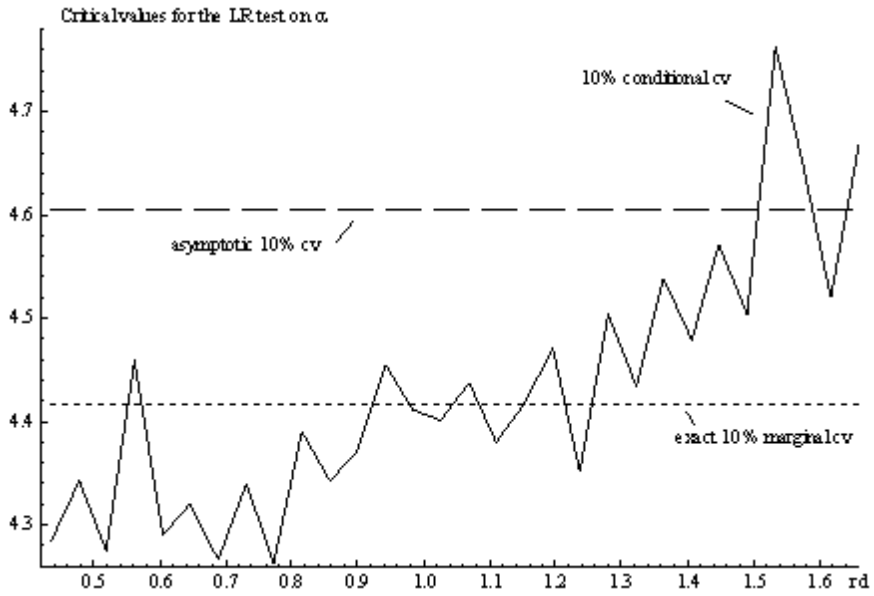


Figure 7: Three different sets of critical values for the W_{obs} test of the hypothesis $H_0 : \alpha = \alpha_0$ when β is known are plotted against the rd statistic. The parameters are $\alpha_0 = (-.19, -.11)'$, $\beta = (1, 1)'$, hence $\rho = 0.7$. The sample size is $T = 20$. The dashed line is the asymptotic critical value – the 90% quantile of the $\chi^2(2)$ distribution. The dotted line gives the (simulated) exact 90% quantile of the W_{obs} statistic under the null. The continuous line gives the estimated conditional 90% quantile of the W_{obs} statistic given the rd statistic. This is estimated by simulation using 400000 replications split over a set of 30 non-overlapping equally spaced grids.

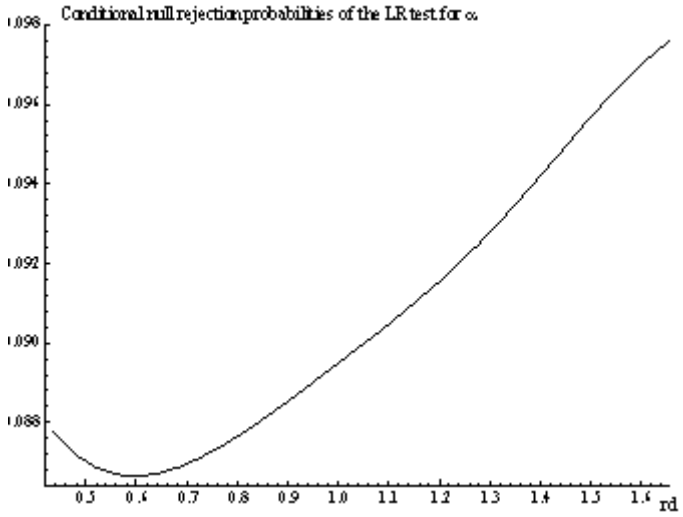


Figure 8: This graph plots the rejection frequency of the W_{obs} test of the hypothesis $H_0 : \alpha = \alpha_0$ (when β is known) under the null conditional on the rd statistic. The parameters are $\alpha_0 = (-.19, -.11)'$, $\beta = (1, 1)'$, hence $\rho = 0.7$. The sample size is $T = 20$. The conditional rejection frequency is estimated by means of a non-parametric Gaussian kernel estimator based on 400000 replications of data from the CVAR model under the null.

comes from economic theory, i.e. in situations when certain economic relationships imply known cointegrating vectors. Prominent examples are the uncovered interest parity or the purchasing power parity in international economics, see [...], the stationarity of real interest rates, see ...

Another justification comes from the superconsistency property in the estimation of β, which guarantees that the MLE $\widehat{\beta}$ converges faster to its true value than $\widehat{\alpha}$. We can argue that the rd statistic can be approximated by an estimate $\widehat{rd}$, upon substitution of $\widehat{\beta}$ for β in (16). The error in that approximation can be shown to be of order T^{-1}. However, it remains to be seen whether this property is useful in finite samples.

3.2 Inference when β unknown

When the cointegrating vector β is unknown, the CVAR is a CEM-(7,3). The difference in dimensions of the sufficient statistic and the parameters is four, implying that the inverted MLE is four-dimensional.

As mentioned, α and β are not identified and we need to impose a normalization. We consider a generic normalization by the known vector c_1 such that $c_1'\beta = 1$, and without loss of generality we may take $c_1'c_1 = 1$ and let c_2 be the orthogonormal complement of c_1 such that $c_1'c_2 = 0$ and $c_2'c_2 = 1$. We further let θ denote the identified parameters in the model.

The observed information matrix is derived in the Appendix and used here to characterize the critical set. The critical set is defined as the subset of all observations in the sample space for which the observed information matrix is singular, or equivalently its determinant is zero. In the Appendix we show that

$$\left|\mathcal{J}_{\widehat{\theta}}\right| = \left(v_2'S_{11}c_2\right)^2 \left(\widehat{\lambda}_1 - \widehat{\lambda}_2\right) \tag{17}$$

see Equation (27), where $\widehat{\lambda}_1 > \widehat{\lambda}_2 > 0$, and v_1 and v_2 solve

$$\left(\widehat{\lambda}_i S_{11} - S_{10}S_{01}\right) v_i = 0, \qquad i = 1, 2. \tag{18}$$

The $\widehat{\lambda}_i$ are the eigenvalues from the standard maximum likelihood procedure where the only difference is that here Ω is known and S_{00} does not appear in the determinantal equation. Note that because of

this, the $\widehat{\lambda}_i$ are no longer the squared sample canonical correlations between ΔY_t and Y_{t-1}, as in the standard case, and they can be larger than 1.

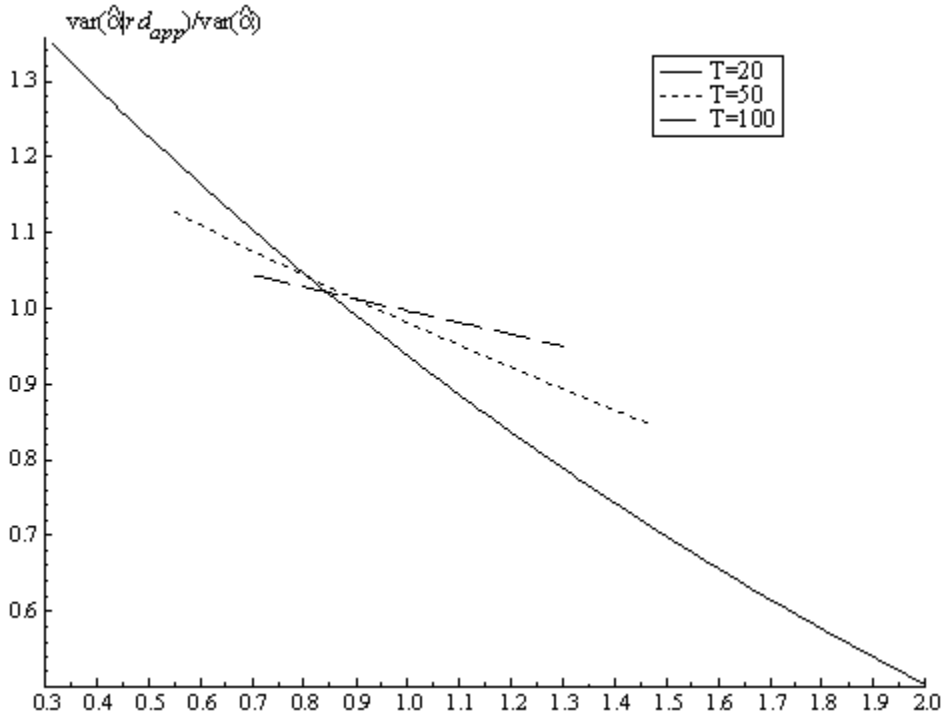
It is evident from (17) that the observed information is singular in two different cases: when $\widehat{\lambda}_1 = \widehat{\lambda}_2$, or when $v_2' S_{11} c_2 = 0$ which occurs when v_1 is proportional to c_2. The probability that either one of these single events occurs is zero, but there are cases where the observed information is close to being singular. In these sensitive regions we would expect inference to be fragile.

When the true λ_1 is small, the *ex-ante* probability of observing a small $\widehat{\lambda}_1$ is larger. A small sample size also increases the *ex-ante* probability of observing a small $\widehat{\lambda}_1$ and a larger $\widehat{\lambda}_2$. When T increases $\widehat{\lambda}_1$ will quickly converge to its limit and also the probability of observing a $\widehat{\lambda}_2$ much larger than zero is very small. So it is in small samples with λ_1 small that we will most often find realizations of the process that lead to fragile inference. *Ex-post* we can simply calculate $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ and it becomes irrelevant what the *ex-ante* probabilities are, but it helps to understand that conditioning is more important in small samples than in large samples.

The analysis of the previous section can be repeated here using an approximate measure of the rd statistic, namely rd_{app}, upon substitution of $\widehat{\beta}$ for β in (16). The variance of the MLE $\widehat{\alpha}_1$ conditional on rd_{app} is given in Figure 3.2 below. This figure looks very similar to Figure 4 above, showing that the effect of using rd_{app} instead of rd is small.

The statistic $T\widehat{\lambda}_2$ is the likelihood ratio (trace) test for testing the cointegrating rank. So one could interpret the use of $\widehat{\lambda}_2$ as having first tested and concluded that the process has one cointegrating vector and one common trend. Under the assumption that the cointegrating rank is one, the statistic $\widehat{\lambda}_2$ is approximately ancillary (this follows from the asymptotic distribution under the null, which does not depend on parameters), and conditioning on $\widehat{\lambda}_2$ is supported by the usual asymptotic arguments for conditional inference.

Another reasoning is that having calculated and used the statistic $\widehat{\lambda}_2$ in a statistical procedure, subsequent inference should take this into account. One should expect inference to be different for cases where there is support for a second cointegrating vector ($T\widehat{\lambda}_2$ large, but smaller than the critical value) and cases where there is clearly no evidence of a second cointegrating vector ($T\widehat{\lambda}_2$ close to 0).



4 Conclusion

In this paper we have investigated the effects of conditioning for inference on the adjustment coefficients α in a simple cointegrating VAR. We have shown that the curvature of the model induces sensitive regions in the sample space where the accuracy of the estimators is much lower than expected *ex ante*. We have constructed a measure of proximity of the observed sample to this sensitive region, called the relative distance statistic, and show that the accuracy of the estimator is heavily dependent on it. We also show that testing hypotheses is adversely affected by proximity to the sensitive region. Standard tests that do not take account of this aspect of the data can be seriously oversized when the sample is close to the sensitive region. This is true for the Wald test based on the expected information.

Proper inference procedures should take into account how close an observation is to the critical set where inference breaks down. The obvious way to do this is to derive the conditional distribution of the test statistics and estimators, given the value of the relative distance statistic. An alternative approach, that we have actually pursued, is to use the observed information instead of the expected information in inference procedures. Our results show that in doing so, one is implicitly conditioning on the relative distance measure, and inference is much more reliable.

A Appendix

A.1 Standardized data in the bivariate CVAR(1)

Consider the pair of r.v.s $X_t = (X_{1t}, X_{2t})$, and the bivariate CVAR(1) model:

$$\Delta X_t = \Pi_x X_{t-1} + \epsilon_t, \quad t = 1, \ldots, T; \quad \epsilon_t \sim NID(0, \Omega). \tag{19}$$

Ω is positive definite, and therefore, it is orthogonally diagonalizable $\Omega = UDU'$, where U is the orthogonal matrix consisting of the eigenvectors of Ω, and D is a diagonal matrix, with the eigenvalues of Ω along its diagonal. Hence, where $\Omega^{1/2} = UD^{1/2}U'$, and $\Omega^{-1/2} = UD^{-1/2}U'$. Since Ω is known, we can standardize the model (19) w.r.t. it. Consider

$$Y_t = \Omega^{-1/2} X_t, \quad \varepsilon_t = \Omega^{-1/2}\epsilon_t, \quad \text{and} \quad \Pi = \Omega^{-1/2}\Pi_x\Omega^{1/2}. \tag{20}$$

A.2 Case of β known

The log-likelihood function is

$$l(\alpha) = T\alpha' S_{0\beta} - \frac{T}{2}\alpha'\alpha S_{\beta\beta}.$$

The first and second order conditions are

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= TS_{0\beta} - T\alpha S_{\beta\beta} \\ \frac{\partial^2 l}{\partial \alpha \partial \alpha'} &= -T\ I_2 S_{\beta\beta}. \end{aligned}$$

Setting the first to zero yields the MLE, see Equation (12), while the observed information is given by $J_\alpha = -\partial^2 l(\alpha)/\partial\alpha\partial\alpha'$, see Equation (13).

To derive the expected information, Equation (14), it suffices to determine $ES_{\beta\beta}$. Observe that $\beta' Y_{t-1}$ follows an AR(1) with zero mean, zero starting value $Y_0 = 0$, autoregressive coefficient $\rho =$

$\beta'\alpha + 1$, and error variance $\beta'\beta$. So $ES_{\beta\beta}$ is given by

$$\begin{aligned} ES_{\beta\beta} &= E\frac{1}{T}\sum_{t=1}^{T}\beta' Y_{t-1}Y_{t-1}'\beta = \\ &= \frac{\beta'\beta}{T}\sum_{t=1}^{T-1}\sum_{i=0}^{t-1}\rho^{2i} = \frac{\beta'\beta}{T}\sum_{t=1}^{T-1}\frac{1-\rho^{2t}}{1-\rho^2} \\ &= \frac{\beta'\beta}{T\left(1-\rho^2\right)}\left(T-1-\sum_{t=1}^{T-1}\rho^{2t}\right) \\ &= \frac{\beta'\beta}{T\left(1-\rho^2\right)}\left(T-\sum_{t=0}^{T-1}\rho^{2t}\right) = \frac{\beta'\beta\left[\left(1-\rho^2\right)-\left(1-\rho^{2T}\right)/T\right]}{\left(1-\rho^2\right)^2}. \end{aligned} \tag{21}$$

Let $W_{obs}\left(\alpha_0;\beta\right)$ and $W_{\exp}\left(\alpha_0;\beta\right)$ denote the Wald test for a point null $H_0 : \alpha = \alpha_0$ when β is known, using the observed and expected information, respectively. Then

$$W_{obs}\left(\alpha_0;\beta\right) = \left(\widehat{\alpha}-\alpha_0\right)' J_\alpha \left(\widehat{\alpha}-\alpha_0\right) = T\left(\widehat{\alpha}-\alpha_0\right)'\left(\widehat{\alpha}-\alpha_0\right) S_{\beta\beta}$$

$$\begin{aligned} W_{\exp}\left(\alpha_0;\beta\right) &= \left(\widehat{\alpha}-\alpha_0\right)' \mathcal{I}_\alpha \left(\widehat{\alpha}-\alpha_0\right) \\ &= T\left(\widehat{\alpha}-\alpha_0\right)'\left(\widehat{\alpha}-\alpha_0\right)\frac{\beta'\beta\left[\left(1-\widehat{\rho}^2\right)-\left(1-\widehat{\rho}^{2T}\right)/T\right]}{\left(1-\widehat{\rho}^2\right)^2} \end{aligned}$$

where $\widehat{\rho} = 1 + \beta'\widehat{\alpha}$.

Next, we derive the likelihood ratio test of the above null hypothesis. The maximized value of the likelihood is:

$$l_{\max}\left(\widehat{\alpha}\right) = \frac{T}{2}\frac{S_{0\beta}'S_{0\beta}}{S_{\beta\beta}}.$$

Hence, the likelihood ratio test is

$$\begin{aligned} LR\left(\alpha_0;\beta\right) &= T\left(\frac{S_{0\beta}'S_{0\beta}}{S_{\beta\beta}} - 2\alpha_0'S_{0\beta} + \alpha_0'\alpha_0 S_{\beta\beta}\right) \\ &= T\left(\widehat{\alpha}'\widehat{\alpha} - 2\alpha_0'\widehat{\alpha} + \alpha_0'\alpha_0\right) S_{\beta\beta} \\ &= T\left(\widehat{\alpha}-\alpha_0\right)'\left(\widehat{\alpha}-\alpha_0\right) S_{\beta\beta}. \end{aligned}$$

Invariance to changes in λ_1 Define the statistic

$$S_{\varepsilon\beta} = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t Y_{t-1}'\beta$$

and observe that $S_{0\beta} = \alpha S_{\beta\beta} + S_{\varepsilon\beta}$. Also, since $\beta' Y_{t-1} = \rho\beta' Y_{t-2} + \beta'\varepsilon_{t-1}$, $S_{\beta\beta}$ and $S_{\varepsilon\beta}$ are invariant to changes in the parameters that leave ρ and β unchanged. But

$$\widehat{\alpha} - \alpha_0 = S_{\varepsilon\beta}S_{\beta\beta}^{-1}$$

so W_{obs} and hence LR are invariant to changes in λ_1 that leave ρ and β constant.

Derivation of RD statistic when β is known The sufficient statistic is $s = \left(S_{0\beta}', S_{\beta\beta}\right)'$ and its expectation is $\tau = \left(ES_{0\beta}', ES_{\beta\beta}\right)$. But note that $ES_{0\beta} = \alpha ES_{\beta\beta} + ES_{\varepsilon\beta} = \alpha ES_{\beta\beta}$, since $ES_{\varepsilon\beta} = 0$. Substituting the above for $\tau = (\alpha', 1)\, ES_{\beta\beta}$ in equation (11), the result follows.

A.2.1 Score, Hessian and the MLE

Consider the log-likelihood of the CVAR model given in Equation (8):

$$l(\alpha, \beta) = T\alpha' S_{01}\beta - \frac{T}{2}\alpha'\alpha\beta' S_{11}\beta$$

The first order conditions are

$$\frac{\partial l}{\partial \alpha} = TS_{01}\beta - T\alpha\left(\beta' S_{11}\beta\right)$$
$$\frac{\partial l}{\partial \beta} = T\alpha' S_{01} - T\alpha'\alpha\beta' S_{11} = T\alpha'\left(S_{01} - \alpha\beta' S_{11}\right)$$

Solving the FOC for α and β yields the equations

$$\alpha(\beta) = S_{01}\beta\left(\beta' S_{11}\beta\right)^{-1},$$
$$\beta(\alpha) = S_{11}^{-1} S_{10}\alpha\left(\alpha'\alpha\right)^{-1}$$

So, the concentrated log-likelihoods either in terms of α or β, are

$$l_c(\beta) = \frac{T}{2}\beta' S_{10} S_{01}\beta\left(\beta' S_{11}\beta\right)^{-1} \tag{22}$$
$$l_c(\alpha) = \frac{T}{2}\alpha' S_{01} S_{11}^{-1} S_{10}\alpha\left(\alpha'\alpha\right)^{-1}. \tag{23}$$

Since α or β are not identified without normalization, the MLE is not uniquely defined by the above, but the maximum of the likelihood is.

Consider the two eigenvalue problems associated with the respective concentrated likelihood functions,

$$\left|\lambda S_{11} - S_{10} S_{01}\right| = 0$$
$$\left|\lambda I_2 - S_{01} S_{11}^{-1} S_{10}\right| = 0.$$

where $|A|$ denotes the determinant of A. Denote the eigenvalues $\widehat{\lambda}_1 > \widehat{\lambda}_2 > 0$ and let v_1, v_2 be the unique vectors that satisfy $\left(\widehat{\lambda}_i S_{11} - S_{01}' S_{01}\right) v_i = 0$, and $v_i' S_{11} v_j = 1$ if $i = j$ and 0 otherwise. Further, let u_1, u_2 denote the eigenvectors of the second problem, i.e., $\left(\widehat{\lambda}_i I_2 - S_{01} S_{11}^{-1} S_{10}\right) u_i = 0$ and $u_i' u_j = 1$ if $i = j$ and 0 otherwise.

The maximum of the likelihood function of the CVAR model is:

$$\max \; l_{CVAR} = \frac{T}{2} \widehat{\lambda}_1, \tag{24}$$

(see e.g. Johansen (1995a, Lemma A.8)). The maximum of the unrestricted VAR likelihood (7) is

$$\max l_{VAR} = \frac{T}{2} \left(\widehat{\lambda}_1 + \widehat{\lambda}_2\right), \tag{25}$$

Thus the standard likelihood ratio, trace test statistic for cointegration is $T\widehat{\lambda}_2$ (see Johansen '89).

Note that the MLE for β (when normalized) is proportional to v_1.

The second order conditions are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha \partial \alpha'} &= -T \; I_2 \left(\beta' S_{11} \beta\right) \\
\frac{\partial^2 l}{\partial \alpha \partial \beta'} &= T S_{01} - 2T \alpha \beta' S_{11} = T \left(S_{01} - 2\alpha \beta' S_{11}\right) \\
\frac{\partial^2 l}{\partial \beta \partial \beta'} &= -T \alpha' \alpha S_{11}.
\end{aligned}$$

Hence, the Hessian matrix is

$$\mathcal{H}_\theta = -T \begin{pmatrix} I_2 \left(\beta' S_{11} \beta\right) & 2\alpha \beta' S_{11} - S_{01} \\ 2 S_{11} \beta \alpha' - S_{01}' & \alpha' \alpha S_{11} \end{pmatrix}.$$

The expected value of the Hessian is useful in deriving the expected information. Let $\Sigma_{ij} = E S_{ij}$ denote the expected values of the samples second moments, given the initial condition $Y_0 = 0$. We have dropped the dependence of the Σ_{ij} on T for simplicity. Letting $C = \beta_\perp \left(\alpha_\perp' \beta_\perp\right)^{-1} \alpha_\perp'$, $B = \alpha \left(\beta' \alpha\right)^{-1} \beta'$, the Engle and Granger (1987) representation is

$$Y_t = C \sum_{i=0}^{t-1} \varepsilon_{t-i} + B \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i}$$

so that

$$var\left(Y_t\right) = CC' t + \left(CB' + BC'\right) \frac{1 - \rho^t}{1 - \rho} + BB' \frac{1 - \rho^{2t}}{1 - \rho^2}.$$

Hence,

$$\begin{aligned}\Sigma_{11} = ES_{11} &= \frac{1}{T}\sum_{t=1}^{T} var\left(Y_{t-1}\right) \\ &= CC'\frac{T-1}{2} + \left(CB' + BC'\right)\frac{T\rho\left(1-\rho\right) - \left(1-\rho^{T+1}\right)}{T\left(1-\rho\right)^2} + \\ &BB'\frac{T\rho^2\left(1-\rho^2\right) - \left(1-\rho^{2T+2}\right)}{T\left(1-\rho^2\right)^2}.\end{aligned}$$

Also, $S_{01} = \alpha\beta' S_{11} + S_{\varepsilon 1}$ and $ES_{\varepsilon 1} = 0$ imply that $\Sigma_{01} = \alpha\beta'\Sigma_{11}$, which simplifies to

$$\Sigma_{01} = \alpha\beta' C'\frac{T\rho\left(1-\rho\right) - \left(1-\rho^{T+1}\right)}{T\left(1-\rho\right)^2} + \alpha\beta' B'\frac{T\rho^2\left(1-\rho^2\right) - \left(1-\rho^{2T+2}\right)}{T\left(1-\rho^2\right)^2}$$

because $\beta' C = 0$ and $\beta' B = \beta'$. The expected Hessian is

$$E\mathcal{H}_\theta = -T\begin{pmatrix} I_2\beta'\Sigma_{11}\beta & \alpha\beta'\Sigma_{11} \\ \Sigma_{11}\beta\alpha' & \alpha'\alpha\Sigma_{11}\end{pmatrix} \tag{26}$$

Information for normalized parameters Consider the generic normalization by the (known) vector c_1 such that $c_1'\beta = 1$. Without loss of generality we may take $c_1'c_1 = 1$ and define the orthogonal complement c_2 s.t. $c_2'c_1 = 0$. We can decompose β as $\beta = c_1 + bc_2$. Note that $c_2 = \partial\beta/\partial b$ and that the orthogonal complement of β is $\beta_\perp = c_2 - bc_1$.

The Jacobian matrix of the normalizing transformation from $(\alpha', \beta')'$ to the identified parameters $\theta = (\alpha', b)'$ is

$$J = \frac{\partial\left(\alpha', \beta'\right)'}{\partial\theta'} = \begin{pmatrix} I_2 & 0 \\ 0 & c_2\end{pmatrix}$$

So, the information w.r.t. the identified parameters is

$$\mathcal{J}_\theta = -J'\frac{1}{T}\mathcal{H}J = T\begin{pmatrix} \beta' S_{11}\beta\ I_2 & 2\alpha\beta' S_{11}c_2 - S_{01}c_2 \\ 2c_2' S_{11}\beta\alpha' - c_2' S_{10} & \alpha'\alpha\ c_2' S_{11}c_2\end{pmatrix}.$$

The expected information is found using (26)

$$\mathcal{I}_\theta = T\begin{pmatrix} \beta'\Sigma_{11}\beta\ I_2 & \alpha\beta'\Sigma_{11}c_2 \\ c_2'\Sigma_{11}\beta\alpha' & \alpha'\alpha\ c_2'\Sigma_{11}c_2\end{pmatrix}.$$

The MLE can be written as: $\widehat{\beta} = v_1\left(c_1'v_1\right)^{-1}$ and $\widehat{\alpha} = S_{01}\widehat{\beta}\left(\widehat{\beta}' S_{11}\widehat{\beta}\right)^{-1} = \widehat{\lambda}_1^{1/2}u_1\left(c_1'v_1\right)$. The last expression follows from $\widehat{\alpha}\widehat{\beta}' = \widehat{\lambda}_1^{1/2}u_1 v_1'$. Also

$$\left(S_{01} - \widehat{\alpha}\widehat{\beta}' S_{11}\right)c_2 = \left(S_{01}S_{11}^{-1} - \widehat{\alpha}\widehat{\beta}'\right)S_{11}c_2 = \widehat{\lambda}_2^{1/2}u_2 v_2' S_{11}c_2$$

To simplify the derivations, define the quantities $w = \widehat{\beta}' S_{11} \widehat{\beta}$, $w_i = v_i' S_{11} c_2$ and $d_i = \widehat{\lambda}_i^{1/2} w_i$, for $i = 1, 2$.Next, note that the identity $I_2 = v_1 v_1' S_{11} + v_2 v_2' S_{11}$ implies $c_2 = v_1 w_1 + v_2 w_2$ and hence

$$c_2' S_{11} c_2 = w_1^2 + w_2^2.$$

Thus, $\mathcal{J}_\theta$ evaluated at the MLE can be written as

$$\mathcal{J}_{\widehat{\theta}} = \begin{pmatrix} I_2 w & d_1 u_1 - d_2 u_2 \\ d_1 u_1' - d_2 u_2' & \widehat{\lambda}_1 \left(w_1^2 + w_2^2 \right) / w \end{pmatrix}.$$

The determinant of the observed information is (using Magnus and Neudecker (1999, p. 12))

$$\begin{aligned} \left| \mathcal{J}_{\widehat{\theta}} \right| &= |I_2 w| \left| \widehat{\lambda}_1 \frac{w_1^2 + w_2^2}{w} - \frac{1}{w} \left(d_1 u_1 - d_2 u_2 \right)' \left(d_1 u_1 - d_2 u_2 \right) \right| \\ &= w_2^2 \left(\widehat{\lambda}_1 - \widehat{\lambda}_2 \right) = \left(v_2' S_{11} c_2 \right)^2 \left(\widehat{\lambda}_1 - \widehat{\lambda}_2 \right). \end{aligned} \tag{27}$$

The intermediate calculations in the second subdeterminant are

$$\begin{aligned} \widehat{\lambda}_1 \frac{w_1^2 + w_2^2}{w} - \frac{1}{w} \left(d_1 u_1 - d_2 u_2 \right)' \left(d_1 u_1 - d_2 u_2 \right) &= \frac{1}{w} \left(\widehat{\lambda}_1 w_1^2 + \widehat{\lambda}_1 w_2^2 - d_1^2 - d_2^2 \right) \\ = \frac{1}{w} \left(\widehat{\lambda}_1 w_1^2 + \widehat{\lambda}_1 w_2^2 - \widehat{\lambda}_1 w_1^2 - \widehat{\lambda}_2 w_2^2 \right) &= \frac{w_2^2}{w} \left(\widehat{\lambda}_1 - \widehat{\lambda}_2 \right). \end{aligned}$$

Finally, using the partitioned inverse formula (see Magnus and Neudecker (1999, p.11)), the inverse of the observed information is

$$\mathcal{J}_{\widehat{\theta}}^{-1} = \begin{pmatrix} \frac{1}{w} \left(I + \frac{(d_1 u_1 - d_2 u_2)(d_1 u_1 - d_2 u_2)'}{w_2^2 \left(\widehat{\lambda}_1 - \widehat{\lambda}_2 \right)} \right) & -\frac{d_1 u_1 - d_2 u_2}{w_2^2 \left(\widehat{\lambda}_1 - \widehat{\lambda}_2 \right)} \\ -\frac{d_1 u_1' - d_2 u_2'}{w_2^2 \left(\widehat{\lambda}_1 - \widehat{\lambda}_2 \right)} & \frac{w_1}{w_2^2 \left(\widehat{\lambda}_1 - \widehat{\lambda}_2 \right)} \end{pmatrix}$$

Tests

LR test of point null, given cointegration

$$\begin{aligned} LR &= 2 \left(l \left(\widehat{\theta} \right) - l \left(\theta_0 \right) \right) = T \left(\widehat{\lambda}_1 - 2 \alpha' S_{01} \beta + \alpha' \alpha \beta' S_{11} \beta \right) \\ &= T \left(\widehat{\alpha}' S_{01} \widehat{\beta} - 2 \alpha' S_{01} \beta + \alpha' \alpha \beta' S_{11} \beta \right) \end{aligned}$$

Test of weak exogeneity This hypothesis corresponds to $\alpha = \alpha_1 c_1$, so that $c_2'\alpha = 0$, where $c_2 = (0,1)'$. To find the maximum of the restricted likelihood, substitute in the concentrated likelihood (23)

$$\begin{aligned} l(\alpha_1 c_1) &= \frac{T}{2}\alpha_1^2 c_1' S_{01} S_{11}^{-1} S_{10} c_1 \left(\alpha_1^2 c_1' c_1\right)^{-1} \\ &= \frac{T}{2} c_1' S_{01} S_{11}^{-1} S_{10} c_1. \end{aligned}$$

Clearly, α_1 is not identified without normalization, but the maximized value of the restricted likelihood is easily found. So, the LR test is

$$LR_{we} = T\left(\widehat{\lambda}_1 - c_1' S_{01} S_{11}^{-1} S_{10} c_1\right).$$

Test of stationarity This hypothesis corresponds to $\beta = \beta_1 c_1$, so that $c_2'\beta = 0$, where $c_2 = (0,1)'$. The maximum of the concentrated likelihood is given by

$$\begin{aligned} l_c(\beta_1 c_1) &= \frac{T}{2}\beta_1^2 c_1' S_{10} S_{01} c_1 \left(\beta_1^2 c_1' S_{11} c_1\right)^{-1} \\ &= \frac{T}{2} c_1' S_{10} S_{01} c_1 \left(c_1' S_{11} c_1\right)^{-1} \end{aligned}$$

which is consistent with the lack of identification of β_1, but it is sufficient for testing the hypothesis of interest.

A.3 Monte Carlo experiments

We consider a number of LR test statistics and conduct different Monte Carlo experiments to investigate their finite-sample marginal and conditional distributions, relative to their unconditional asymptotic distributions.

Let Q, S denote two statistics, and let $F_Q(q;\theta,T)$ and $F_{Q|S}(q|\ s;\theta,T)$ denote the marginal and conditional finite-sample distribution functions of Q. When Q is asymptotically pivotal, as it is the case for the LR statistics we consider, let $F_Q^{asy}(q)$ denote its asymptotic distribution (usually χ^2). An asymptotic a-level critical value is given by q_a such that $F_Q^{asy}(q_a) = 1 - a$. The exact finite-sample a-level critical value will be denoted by $q_a(\theta, T)$. Clearly, $q_a = q_a(\cdot, \infty)$.

The statistic Q is not pivotal if $F_Q(\cdot;\theta_1,T) \neq F_Q(\cdot;\theta_2,T)$ for some $\theta_1 \neq \theta_2$.

We do four sets of experiments.

I. We compare the finite sample to the asymptotic distribution. We use three different statistics:

1. The Kolmogorov-Smirnof statistic of equality of two distributions. Let F_1 and F_2 be two cdfs evaluated at n points $x_1 \dots x_n$. The KS is given by

$$KS = \max_x |F_1(x) - F_2(x)|.$$

   The critical values are $1.36n^{-1/2}$ (5%) and $1.63n^{-1/2}$ (1%).

2. The difference in the quantiles $q_a(\theta, T) - q_a$.

3. The difference in tail probabilities (dual of the above), namely $1 - F_Q(q_a; \theta, T) - a$. This is the difference in the null rejection probability (NRP) of the test from its asymptotic level a, i.e.

$$NRP(\theta, T) = 1 - F_Q(q_a; \theta, T).$$

These statistics can be simulated for different values of θ and T. It must be true that the differences go to zero as T grows, but it is interesting to see how they vary with θ.

II. The conditional distribution of Q given S.

1. Conditional quantile $q_a(s; \theta, T)$ such that $F_{Q|S}(q_a(s; \theta, T)|s; \theta, T) = a$.

2. Conditional p-value of unconditional test $p_a(s; \theta, T) = F_{Q|S}(q_a(\theta, T)|s; \theta, T)$.

3. Conditional variance of $\widehat{\alpha}, \widehat{\beta}$ (economic normalization), relative to "unconditional" variance: simulate the "variance inflation coefficient" (VIC)

   (a) For $\widehat{\alpha}$, compute $\sigma_{\widehat{\alpha}}(\theta, T)$ by simulation, report $VIC_{\widehat{\alpha}} = \sigma_{\widehat{\alpha}}(s; \theta, T) / \sigma_{\widehat{\alpha}}(\theta, T)$.

   (b) For $\widehat{\beta}$, standardize first, using random asymptotic variance (inverse of observed info): $z = \widehat{\beta}/\widehat{\sigma}_{\widehat{\beta}}$. Then, report $VIC_{\widehat{\beta}} = \sigma_z(s; \theta, T) / \sigma_z(\theta, T)$.

Summarize the results by response surfaces.

III. Bootstraping the conditional LR tests and conditional variances.

1. For point null, there are no nuisance parameters, so it is straightforward. Estimate $q_a\left(\widehat{\lambda}_2; \theta_0, T\right)$ by simulation. Check the size of the bootstrap conditional LR (BCLR) test by MC.

2. For composite nulls, estimate $q_a\left(\widehat{\lambda}_2; \widehat{\theta}_0, T\right)$, where $\widehat{\theta}_0$ is the restricted MLE estimate of θ. Check the size of the BCLR over different values of the nuisance parameters.

3. Conditional variances: bootstrap the VIC. Do a simulation to check how accurately BVIC approximates true VIC. (e.g. check MSE).

IV. Power comparisons Focus on the hypotheses H_{we} and H_{β},

References

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B 49*(1), 1–39.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist. 3*(6), 1189–1242.

Engle, R. F. and C. W. J. Granger (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica 55*(2), 251–276.

Hansen, H. and A. Rahbek (2002). Approximate conditional unit root inference. *J. Time Ser. Anal. 23*(1), 1–28.

Hosoya, Y., Y. Tsukuda, and N. Terui (1989). Ancillarity and the limited information maximum-likelihood estimation of a structural equation in a simultaneous equation system. *Econometric Theory 5*(3), 385–404.

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

Johansen, S. (1995b). The role of ancillarity in inference for non-stationary variables. *Economic Journal 105*, 302–320.

Johansen, S. (2002a). A small sample correction for tests of hypotheses on the cointegrating vectors. *J. Econometrics 111*(2), 195–221.

Johansen, S. (2002b). A small sample correction for the test of cointegrating rank in the vector autoregressive model. *Econometrica 70*(5), 1929–1961.

Magnus, J. R. and H. Neudecker (1999). *Matrix differential calculus with applications in statistics and econometrics*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd.

van Garderen, K. J. (1995). Variance inflation in Curved Exponential Models. Working Paper 9522, University of Southampton, Southampton.

van Garderen, K. J. (1997). Curved exponential models in econometrics. *Econometric Theory 13*(6), 771–790.