

# Conditional Inference in the Cointegrated Vector Autoregressive Model

Kees Jan van Garderen

Sophocles Mavroeidis

University of Amsterdam

Brown University

February 15, 2006

## Abstract

A Vector Autoregressive model with normally distributed innovations is a Curved Exponential Model. 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 for inference on the speed-of-adjustment ( $\alpha$ ) coefficients. We show that for some realizations of the model the sample is far less informative on  $\alpha$  than might be expected ex-ante and this should be taken into account when making inference. Conditioning is therefore crucial. We show that conditional inference can be carried out using the observed information instead of the expected information.

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

JEL classification: C10 C32

## 1 Introduction

There are a number of economic applications where theory strongly suggests the presence of certain equilibrium relationships between non-stationary variables, but where this is not borne out by the data. A classic example based on the law of one price is the Purchasing Power Parity (PPP) hypothesis, see surveys by Rogoff (1996) or Taylor and Taylor (2004). In its weak form, PPP implies that the log real

exchange rate between two countries must be stationary even if the log prices and the log nominal exchange rate are integrated (i.e. they are stationary only after taking first differences). So, according to the PPP hypothesis there exists a cointegrating relationship and the coefficients of this relationship are known. Despite its intuitive appeal most empirical investigations have found very little evidence in favour of this hypothesis, see Taylor and Taylor (2004). There is nevertheless a strong belief amongst economists that price levels must converge in the long run. The apparent persistent deviation is arguably caused by market imperfections, but in the long run they must eventually disappear. In response to the unfavourable empirical findings, questions have been asked about the power properties of the econometric tests employed to find cointegrating relations and, on a constructive level, attention has shifted to the adjustment process and the speed of adjustment to the long run PPP relation in particular.

In this paper we investigate inference on the adjustment parameter, usually denoted  $\alpha$ , in a cointegration setting when we are not willing to abandon the implications of economic theory for the cointegrating vector even when faced with a sample that is unfavorable to the theory. In those situations, we will treat the sample as extreme in certain sense to be discussed later, but not as evidence against the theory, and address the problem of making inference on  $\alpha$  and other aspects of the model that are not specified by theory. We will consider cases where theory specifies the cointegrating vector, usually denoted  $\beta$ , and henceforth assume  $\beta$  is fixed and known. In addition to the PPP example there are a number of other prominent examples in various branches of economics where the cointegrating vector  $\beta$  is known based on theory. There are the Uncovered Interest Parity (UIP) hypothesis in international economics, present value models in finance and macro economics, the permanent income hypothesis in macro economics, to name but a few.

A prominent example is the relationship between stock prices and dividends and the implications of deviations of the dividend-price ratio from its long-run level for the stock market outlook Campbell and Shiller (2001). Following Campbell and Shiller (1987), a present value model implies that the log of real stock prices and real dividends are cointegrated with a constant cointegrating vector (1,-1), that is, the log dividend-price ratio is stationary. Therefore, if the dynamics of log prices and log dividends can be characterized by a VAR, the effects of short-term movements of the log dividend-price ratio

for the stock market outlook can be seen as hinging upon the value of the adjustment parameter  $\alpha$ . The related issue of predictability of returns is a particular hypothesis on  $\alpha$  which has been studied extensively by so-called predictive regressions, see Stambaugh (1999) and Campbell and Yogo (2004). Of particular concern in this literature is the fact that the log dividend-price ratio (or some other valuation ratio) is close to being unstable, meaning that its deviations from some constant long-run level are very persistent.

Inference in these cases requires special attention for two reasons. First, as previously mentioned, the estimated  $\alpha$  may lie very close to, or inside the non-stationarity region for the disequilibrium process. We still want to be able to construct valid confidence intervals which have correct significance levels in these circumstances. One reason why the estimated  $\alpha$  might lie in the non-stationary region is that adjustment back to equilibrium is very slow because the true value of  $\alpha$ , although still inside the stationarity region, is very close to non-stationarity. The probability of an  $\alpha$  estimate in the non-stationary region can be over 30% in these cases as we will see below.

The second aspect that requires special attention is that there can be realisations of the process with very little information on  $\alpha$ . In those cases  $\alpha$  will be very imprecisely estimated. This can occur for all values of  $\alpha$ , whether they lie inside or outside the stability region. We will show below that it occurs in cases where there is very little variation around an equilibrium set. It may be obvious that when we do not observe the system out of equilibrium very much, it is hard to estimate the return to equilibrium. These un-informative realizations can occur even in cases where *ex-ante* one might expect reasonable amounts of information. Hence there are two aspects to the amount of information, one due to the true parameter values and the second due to the sampling variation. The true parameter values determine the amount of information that one might have expected, based on averaging over all possible sample paths. One way of expressing this expected amount of information is through the expected Fisher information matrix. In standard likelihood procedures the inverse of this matrix, evaluated at the the MLE, is used as the variance of the estimator and it varies with the parameter. The sampling aspect of the information on the other hand, concerns the fact that the actual information in the actual observed sample, as measured by for instance the observed Fisher information matrix, can be rather different from that what might have been expected *ex-ante*. So after

the sample has been observed the *ex-post information* may be very different in two realizations from exactly the same model, with identical parameter values, and different from the *ex-ante information*. Put differently, the same data generating process can lead to samples of very different types.

In economic time series we naturally get one realization of the process only. Inference procedures should reflect the type of realization, which means that standard errors and confidence regions should be dependent on the type. In statistical terms this implies that the distribution of  $\hat{\alpha}$  should be conditional on the type of realization and this requires an auxiliary statistic which indicates, or measures the type.

We address the choice of such a statistic for indicating the type of realization and which is to be used in conditional procedures. There are two desirable properties that such a statistic should have: it should indicate the accuracy of the estimator, and second, it should not correlate with the estimator. In fact it is desirable for this statistic to have a distribution completely free of any parameters. Such a statistic is called ancillary and the conditionality principle from statistics would then prescribe that inference should be carried out conditionally on such a statistic. No exact ancillary statistic is known however for this problem. It is even difficult to construct a statistic which is approximately ancillary when the parameters approach the non-stationarity region. So one of the interesting results in this paper is that we suggest a new auxiliary statistic, called the signed-LM statistic, that is almost uncorrelated with the estimator irrespective of parameter values and is a good indicator of the accuracy of the estimator for the adjustment coefficient.

The signed-LM statistic is motivated by insights from the theory on Curved Exponential Families (CEMs) discussed below. Models for cointegration are curved in a statistical sense as defined by Efron (1975). In fact, since the Cointegrating Vector Autoregressive (CVAR) model is a nonlinear subset of a VAR, which are known to be curved exponential (e.g. van Garderen 1997a), the CVAR must also be a CEM. The curvature of the model induces regions in the sample space where the inference is more difficult, see van Garderen (1995) and this is discussed below. These sensitive regions provide 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.

For the cointegrating vector  $\beta$  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.

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 to a 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  $\alpha$  converge at the usual rate and the information matrix block corresponding to  $\alpha$  is  $O_p(T)$  and, appropriately scaled, does converge to its asymptotic expectation.

Basawa and Brockwell (1984) and Sweeting (1992) show conditional asymptotic normality of the Maximum Likelihood Estimator (MLE) in a general setting when it is scaled by the random norm based on the observed information. There are some conditions, but these explicitly allow for non-ergodic processes. We also find for the adjustment coefficient in the cointegrating setting,  $\hat{\alpha}$  scaled by the observed information instead of the expected information leads to standard inference where the usual  $\chi^2$  critical values lead to correct confidence intervals.

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 ratio 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.

The idea of conditioning on an ancillary statistic has also been taken up by Hosoya, Tsukuda, and Terui (1989) in the context of the Single Structural Equation Model (SSEM). There are a number of similarities between the CVAR model and the 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. One difference, however, is that when the SSEM is exactly identified, the model is a full exponential model, whereas the VAR without rank restrictions is still a curved exponential model. Another feature they share is that the number of parameters is less than the number of sufficient statistics implying that they are both CEMs (see Hosoya, Tsukuda, and Terui 1989 and van Garderen 1997a). 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 is the main motivation for conditioning by Hosoya, Tsukuda, and Terui (1989) in the context of the SSEM. They found that the distribution of the Limited Information Maximum Likelihood (LIML) estimator depends on the smallest characteristic root associated with LIML estimation. In this 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 find a stronger effect of conditioning in the CVAR than those found for the SSEM but the effect is still limited. The smallest eigenvalue has a distribution that does not depend very much on the true parameter values of the model, and is in this sense approximately ancillary, but it is not a very good indicator of the accuracy of the MLE. This goes back to the choice of auxiliary statistic and we show that the signed-LM statistic is simply preferable.

The paper is organized as follows. In the next section we describe the model and its *ex-ante* properties. Section 3 concerns post-sample inference and discusses the relevant factors in the realized

process that determine the accuracy of the estimators. Section 4 puts the CVAR in the context of curved exponential models and Section 6 discusses three alternative conditioning statistics. Section 6 presents arguments in favor of conditional inference, and gives results relevant for the correct and improved inference presented in Section 6.2 Finally, Section 7 concludes. Proofs and derivation are given in the appendix at the end.

## 2 The Model

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

$$\Delta Y_t = \Pi Y_{t-1} + \varepsilon_t, \quad t = 1, \dots, T \quad (1)$$

where  $\varepsilon_t$  are zero mean independently normally distributed disturbances with contemporaneous covariance matrix  $\Omega$ . For simplicity we will assume that  $\Omega$  is known throughout and can therefore be set equal to the identity. The process is stable when the eigenvalues of the  $2 \times 2$  matrix  $(I_2 + \Pi)$  are inside the unit circle. If exactly one of the eigenvalues is unity, the matrix  $\Pi$  is of reduced rank and the model becomes a cointegrated VAR (CVAR). Because the rank of  $\Pi$  equals 1, we can write  $\Pi = \alpha\beta'$  where  $\alpha$  and  $\beta$  are 2-dimensional vectors. The vector  $\beta$  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  $\alpha$  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  $\beta$  called the attractor set which is spanned by  $\beta_\perp$ .

It is clear that any multiple of  $\beta$  would define the same equilibrium since the orthogonal space would be unchanged, and the only effect is that the corresponding  $\alpha$  is reduced by the same factor, thereby leaving  $\Pi$  unchanged. It is clear that  $\alpha$  and  $\beta$  are not identified and  $\alpha$  cannot simply be interpreted as the speed of adjustment. We can think of  $\alpha$  and  $\beta$  in terms of their angles  $\varphi$  and  $\phi$ , relative to horizontal axis, for instance, and their length. Their angles are unique (modulo  $\pi$ ) 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  $\Pi$ .

Let  $\rho$  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

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

Another key quantity is  $\lambda_1$  which is the probability limit of the largest solution to an eigenvalue problem in the Johansen likelihood estimation procedure. 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  $\alpha$  and  $\beta$  and the quantities  $\rho$  and  $\lambda_1$ . 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  $\alpha$ , call this  $\Delta Y_{t|t-1}^e = E[\Delta Y_t|Y_{t-1}]$ . The  $\lambda_1$  is the expected squared distance that the process covers towards equilibrium  $E\left[\left(\Delta Y_{t|t-1}^e\right)' \Delta Y_{t|t-1}^e\right]$ , where the expectation is over all possible  $Y_{t-1}$ .

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  $\alpha$  (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  $\alpha$  in order to get back to equilibrium. Then,

$$1 - \rho = \frac{\left\|\Delta Y_{t|t-1}^e\right\|}{\left\|\Delta Y_{t|t-1}^*\right\|}$$

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  $\psi$  determines the efficiency of the expected return to equilibrium, in the following sense. When  $\psi$  is equal to  $180^\circ$  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.

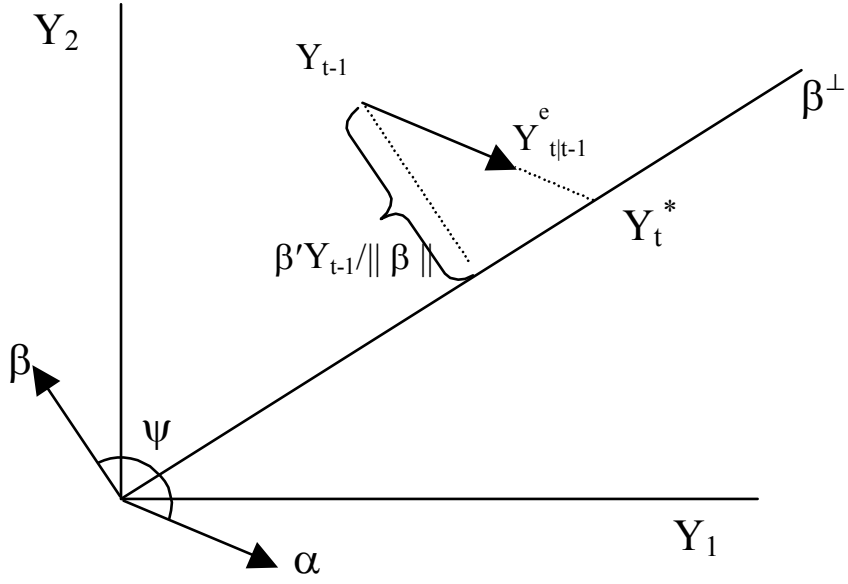


Figure 1: The dynamics of the bivariate CVAR.

The Engle and Granger (1987) representation for  $Y_t$  is

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

where  $\alpha_{\perp}$  and  $\beta_{\perp}$  are orthogonal to  $\alpha$  and  $\beta$  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

$$\sum_{i=1}^t \alpha'_{\perp} \varepsilon_i, \quad (4)$$

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

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

By the Granger representation theorem it follows that  $Y_t^* = \beta_{\perp} (\alpha'_{\perp} \beta_{\perp})^{-1} \sum_{i=1}^{t-1} \alpha'_{\perp} \varepsilon_i$ . Moreover it also follows that every shock  $\varepsilon_t$  can be decomposed into a permanent shock  $\alpha'_{\perp} \varepsilon_t$  and a transitory shock  $\alpha' \varepsilon_t$ . Further *ex-ante* properties can of course be found in many articles and books, including Engle and Granger (1987) and Johansen (1995a, Section 6).

### 3 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  $\beta$ . There are two aspects of the data that govern the accuracy with which  $\beta$  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  $\beta$  (slope of the attractor set in Figure 1). If there is very little common trend variation, for instance because the observations are evenly spread around one particular equilibrium point, then it is very difficult to determine  $\beta$ .

The other case where we can estimate  $\beta$  accurately is when the dispersion *around* the equilibrium set is small. In contrast, if the actual disequilibria are large, then  $\beta$  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  $\beta$ . 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 equilibrium relationship and the other very uninformative.

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

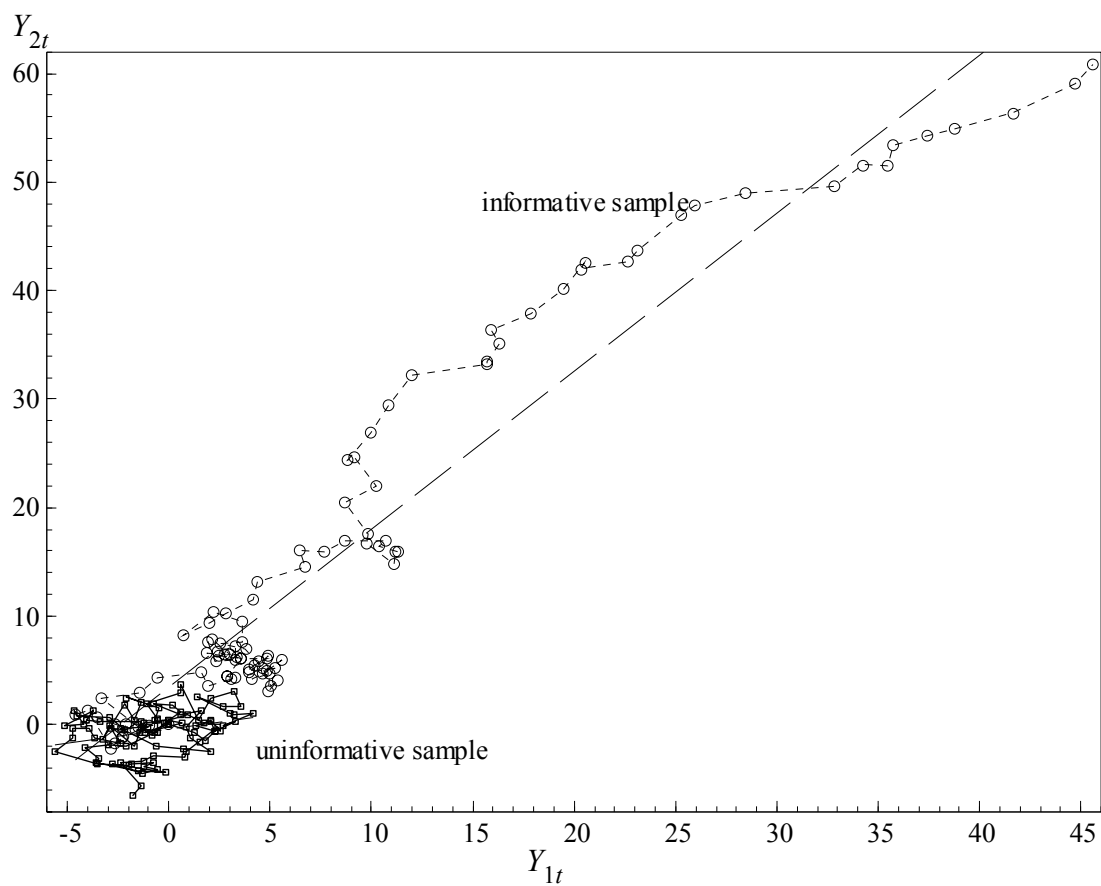


Figure 2: Scatter plot of  $Y_{2t}$  on  $Y_{1t}$  for two different samples (squares and circles) of size  $T = 50$ , drawn from the same CVAR model with parameters  $\alpha = (-.26, .16)'$  and  $\beta = (1, 1)'$  (thus,  $\rho = 0.9$  and  $\lambda_1 = 1$ ).

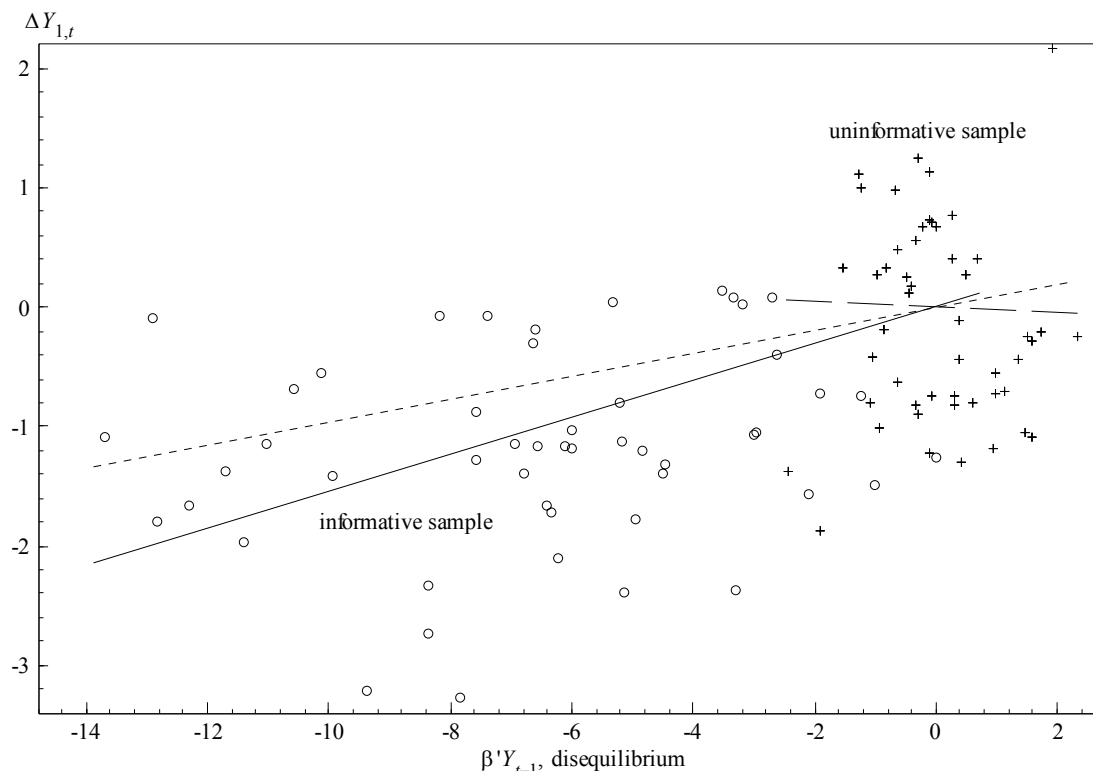


Figure 3: Scatter plot of  $\Delta Y_{1t}$  on  $\beta'Y_{t-1}$ , for two different samples (plus and circles) of size  $T = 50$ , drawn from the same CVAR model with parameters  $\alpha = (-.26, .16)'$  and  $\beta = (1, 1)'$  (thus,  $\rho = 0.9$  and  $\lambda_1 = 1$ ).

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  $\alpha$ . Conversely, when the realized disequilibria are large,  $\alpha$  can be estimated accurately. To demonstrate this idea consider the case where  $\beta$  is known and  $\alpha$  contains the slopes of the regression of  $\Delta Y_t$  on  $\beta'Y_{t-1}$ . Figure 3 plots the change in the first element of  $\Delta Y_t$  against the disequilibrium  $\beta'Y_{t-1}$  and shows one sample that is very informative on the adjustment coefficient  $\alpha_1$  and another that is not very informative.

For known  $\beta$  the total sum of squared disequilibrium terms is (proportional to) the observed information and we will show below how this plays an important role in the accuracy of the estimator, the construction of confidence intervals, and for conditional inference.

More generally,  $\beta$  is unknown, and we can think of estimating  $\alpha$  by first estimating  $\beta$ , which can

be done superconsistently, and then regressing  $\Delta Y_t$  on the generated regressor  $\hat{\beta}' Y_{t-1}$ . When  $\hat{\beta}$  is the MLE, the OLS estimator of  $\alpha$  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  $\alpha$  and this error causes additional variation in the distribution of  $\hat{\alpha}$ .

So for  $\alpha$  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  $\beta$ .

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

The question now becomes how to construct an auxiliary statistic based on these insights that can be used for conditional inference. There are various possibilities such as the Efron-Hinkley ancillary which is approximately ancillary in i.i.d. cases but does not work when  $\rho$  is in a neighbourhood of 1, in the sense that it neither can be used to indicate the variance, nor is it even close to being ancillary. An alternative is a relative distance statistic suggested by Van Garderen (1995) and discussed below, but also breaks down when  $\rho$  approaches 1. We may also construct approximate ancillary statistics based on test statistics and in order to do so it is sensible to first discuss CEMs and its embedding model against which the test is constructed.

## 4 Curved Models

The CVAR model is embedded in a more general VAR model. The rank restrictions imposed on the  $\Pi$  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 conditional 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:

$$l(\eta) = \eta \cdot s - \kappa(\eta), \quad (6)$$

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  $\eta$  is genuinely  $k$  dimensional, then the model is a FEM. If  $\eta$  lies on a smooth manifold of lower dimension  $d$ , the model is a CEM- $(k, d)$ . It will then be possible, at least locally, to write  $\eta$  as a differentiable function of a new  $d$ -dimensional parameter,  $\theta$  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 (1997a) for dependent and non-identically distributed observations.

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

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

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  $\eta = (\Omega^{-1}, \Pi'\Omega^{-1}, \Pi'\Omega^{-1}\Pi)$ ,  $s = (S_{00}, S_{01}, S_{11})$  and inner product defined by the trace. Because of the symmetries involved, there are a number of redundant elements in  $\eta$  and  $s$ . The dimension  $k$  is 10 in the bivariate VAR (and  $k = n^2 + n(n+1)$  when  $Y_t$  is  $n$ -dimensional), while the number of free parameters is 7 (and  $d = n^2 + n(n+1)/2$ , when  $Y_t$  is  $n$ -dimensional). The difference in dimension is 3 ( $n(n+1)/2$  in general). If  $\Omega$  is known, as we are assuming throughout, the dimensions reduce to 7 and 4 for  $s$  and  $\theta$  respectively.

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

$$l(\alpha, \beta) = -\frac{T}{2} \text{tr} (-2S_{01}\beta\alpha'\Omega^{-1} + S_{11}\beta\alpha'\Omega^{-1}\alpha\beta'). \quad (8)$$

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- $(7,3)$ .

If we impose the additional restriction that  $\beta$  is known, the model becomes

$$l(\alpha, \beta) = T\alpha'\Omega^{-1}S_{0\beta} + -\frac{T}{2}S_{\beta\beta}\alpha'\Omega^{-1}\alpha.$$

where  $S_{0\beta} = S_{01}\beta$  and  $S_{\beta\beta} = \beta' S_{11}\beta$ . The minimal sufficient statistic  $(S_{0\beta}, S_{\beta\beta})$  has dimension 3 and there are only 2 parameters and the model with known  $\beta$  is a CEM-(3,2). Since the model is embedded in a 3-dimensional space we can represent it graphically in full.

Note that in contrast, when  $\alpha$  is assumed known and  $\beta$  is unknown, the model becomes CEM-(5,2). But this situation is of limited empirical relevance, so we will not discuss it here.

#### 4.1 Consequences of $k - d > 0$

The difference in dimension between the minimal sufficient statistic and the number of parameters has two immediate consequences.

First, 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 auxiliary statistic  $a$  say, such that  $s \mapsto (\hat{\theta}, a)$  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 in statistics.

Second, since there is a 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 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.

$$\frac{\partial l}{\partial \theta'} = [s - \tau(\theta)]' \frac{\partial \eta}{\partial \theta'}, \quad (9)$$

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  $\theta$  for the CEM. 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. Note that if the model is full, that  $\partial \eta / \partial \theta'$  is of full rank  $k$  and the MLE is uniquely defined by  $\tau(\theta) = s$ .

Third, the observed Fisher information  $\mathcal{J}_\theta$ , defined as minus the Hessian of the likelihood function,

does not equal its expectation, the expected Fisher information  $\mathcal{I}_\theta$ :

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

$$\mathcal{I}_\theta = -\frac{\partial \eta'}{\partial \theta} \frac{\partial^2 \kappa}{\partial \eta \partial \eta'} \frac{\partial \eta}{\partial \theta'}, \quad (11)$$

In standard likelihood inference  $\mathcal{I}_\theta$  plays a crucial role in the sense that its inverse is used as the variance-covariance matrix of the MLE. It is constant for given  $\theta$  and hence results in the same estimated covariance matrix for samples that give the same value of  $\hat{\theta}$ .  $\mathcal{J}_\theta$  on the other hand varies with  $s$  for constant  $\hat{\theta}$  and plays an essential role in conditional inference. We see that the quantity  $s - \tau(\hat{\theta})$  linearly determines 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 a full- or linear exponential model. The existence of a non-empty critical set and the sensitive region is therefore 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. One of the objectives of this paper is to identify statistics that can be used for this purpose. First we will make the expression here explicit for the CVAR with  $\beta$  known.

## 4.2 CVAR with $\beta$ known

To illustrate these ideas explicitly in the cointegration setting, consider the simplest possible framework with  $\beta$  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  $s = (S'_{0\beta}, S_{\beta\beta})'$ , and we can illustrate

the ideas above graphically.

The MLE equals

$$\hat{\alpha}(\beta) = S_{0\beta} S_{\beta\beta}^{-1}, \quad (12)$$

which solves the first order conditions for the MLE based on (9),  $(s - \tau(\hat{\theta}))' \partial\eta(\hat{\theta})/\partial\theta' = 0$ , with:

$$\frac{\partial\eta}{\partial\theta'} = \begin{pmatrix} I_2 \\ -\alpha' \end{pmatrix}_{3 \times 2}, \text{ and orthogonal complement } \left(\frac{\partial\eta}{\partial\theta'}\right)^\perp \propto \begin{pmatrix} \alpha \\ 1 \end{pmatrix}.$$

It is easy to verify that  $s - \tau(\hat{\theta})$  is proportional to the orthogonal complement of  $\partial\eta(\hat{\theta})/\partial\theta'$ , since the expectation of  $s$  is (see Appendix):

$$\tau(\alpha) = E \begin{pmatrix} S_{0\beta} \\ S_{\beta\beta} \end{pmatrix} = \frac{\beta' \Omega \beta [(1 - \rho^2) - (1 - \rho^{2T})/T]}{(1 - \rho^2)^2} \begin{pmatrix} \alpha \\ 1 \end{pmatrix}, \quad (13)$$

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

$$\mathcal{J}_\alpha = T S_{\beta\beta} \Omega^{-1} \quad (14)$$

which is singular if and only if  $S_{\beta\beta}$  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 a single point, namely the origin in the 3-dimensional sample space of the sufficient statistics.

The interpretation of this event is that there are no deviations from equilibrium at all since in every period  $\beta' Y_{t-1} = 0$ . It is then impossible to estimate the speed of return to equilibrium, since we never observe any deviations from equilibrium as we discussed previously.

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

$$\mathcal{I}_\alpha = T \frac{[(1 - \rho^2) - (1 - \rho^{2T})/T]}{(1 - \rho^2)^2} (\beta' \Omega \beta) \Omega^{-1}. \quad (15)$$

The inverse of this matrix would normally be used as the estimate of the covariance matrix of  $\hat{\alpha}$  since the asymptotic distribution of  $\hat{\alpha}$  is given by (cf. Johansen, 1995, Theorem 13.5):

$$\sqrt{T}(\hat{\alpha} - \alpha) \stackrel{a}{\sim} N \left( 0, \frac{1}{\beta' \Omega \beta} \frac{(1 - \rho^2)^2}{[(1 - \rho^2) - (1 - \rho^{2T})/T]} \Omega \right). \quad (16)$$

Note that  $\hat{\alpha}_1$  and  $\hat{\alpha}_2$  are asymptotically independent if  $\Omega$  is diagonal as assumed here, and that  $\sqrt{T}(\hat{\alpha} - \alpha)$  appropriately normalized will have approximately a standard normal distribution. In

small samples, however, this approximation is very poor indeed. Furthermore we are interested in the conditional distribution of  $\hat{\alpha}$  when the process has very little variation around the equilibrium. The approximation (16) in that case is even worse, but in order to show this we first need to derive an auxiliary statistic on which to condition.

## 5 Auxilliary Statistics

In this section we discuss two new auxiliary statistics that can be used for conditioning and to show that conditioning is relevant. In our previous discussion on *ex-post* accuracy, we argued that certain samples are less informative about the parameters than other samples, but we still need a to derive a statistic to indicate what type of sample we have and to indicate how we should adjust measures of accuracy, confidence intervals, tests, etc. Such a statistic should preferably be ancillary, i.e. have a distribution that does not depend on the parameters of interest and therefore not contain information about the parameters itself, since this information could otherwise be used to improve the estimators. In many models no exact ancillaries are known and this has lead to the development of approximate ancillaries. These asymptotic ancillaries are not satisfactory however in models for near- and non-stationary data. Either because their distribution changes rapidly with the parameters, or because they have no information on the accuracy of the estimator or inference procedures. The two statistics we introduce here do have information on accuracy and inference procedures, and of the two the signed LM statistic has a distribution that does not vary very much with  $\rho$ , even as  $\rho$  reaches 1. It should be noted that in unreported simulations we found that the unsigned version of the LM statistic shares the approximate ancillary property, but contains hardly any information on the accuracy of the estimator. First we introduce the statistics and investigate their ancillarity properties, and in the next section we show their effectiveness and use for conditional inference.

### 5.1 RD: The Relative Distance statistic

This statistic was proposed in Van Garderen (1995) based on distance of the observation  $s$  to the critical point where the observed information is singular, relative to the total distance from this

critical point to the expected value of  $s$ . The critical point in this case is where  $S_{\beta\beta} = 0$  since  $\mathcal{J}_\alpha = T S_{\beta\beta} I_2$ . The difference between the observation  $s$  and its expected value given the estimated parameter value equals

$$s - \tau(\hat{\theta}) = \begin{pmatrix} S_{0\beta} \\ S_{\beta\beta} \end{pmatrix} - \frac{\beta' \Omega \beta \left[ (1 - \hat{\rho}^2) - (1 - \hat{\rho}^{2T}) / T \right]}{(1 - \hat{\rho}^2)^2} \begin{pmatrix} \hat{\alpha} \\ 1 \end{pmatrix} \quad (17)$$

where the last equality follows from the fact that  $s - \tau(\hat{\theta})$  must be proportional to the orthogonal complement of  $\partial\eta(\hat{\theta})/\partial\theta'$ , by the first order conditions of the MLE  $(s - \tau(\hat{\theta}))' \partial\eta(\hat{\theta})/\partial\theta' = 0$ . For a fixed  $\hat{\theta}$ , this characterizes all the  $s$  that give the same value for the MLE.

The critical set has  $S_{\beta\beta} = 0$  and the corresponding  $s_{crit} = (0, 0, 0)'$ . The following statistic, denoted  $rd$ , measures the total distance of  $s$  from  $s_{crit}$  relative to what is expected ex ante (see appendix for details):

$$rd = \frac{(1 - \hat{\rho}^2)^2}{(1 - \hat{\rho}^2) - (1 - \hat{\rho}^{2T}) / T} \frac{S_{\beta\beta}}{\beta' \Omega \beta} \quad (18)$$

The sensitive region is when  $rd$  is close to zero.

The statistic  $rd$  can also be interpreted as measuring the *ex-post* variability of the disequilibrium  $S_{\beta\beta}$  relative to what would be expected *ex-ante* for the *estimated* value of  $\alpha$ , which equals  $\beta' \Omega \beta \left[ (1 - \hat{\rho}^2) - (1 - \hat{\rho}^{2T}) / T \right] / (1 - \hat{\rho}^2)^2$ . Thus when the variability in the sample is higher than expected, there is more information on  $\alpha$  and  $\alpha$  is more accurately estimated. The  $rd$  statistic seems to work well in terms of ancillarity properties and providing an indication of the accuracy when  $\rho < 0.7$ , but when  $\rho$  approaches 1 the distribution of  $rd$  changes and the ancillarity properties are lost. This is shown in the following graph. Figure 4 plots the simulated density of the  $rd$  statistic in our model for three values of  $\rho = 0.7, 0.9$  and  $0.999$  and  $T = 10$ . It is clear that the location and skewness of the distribution varies with  $\rho$

## 5.2 Signed-LM statistic

In general we can use test statistics to construct approximate ancillary statistics. In CEM models we can test the curved model against the ambient, linear embedding model. The curved model with restrictions on the canonical parameter  $\eta(\theta)$  can be tested against the model for  $s$  where  $\eta$  is not

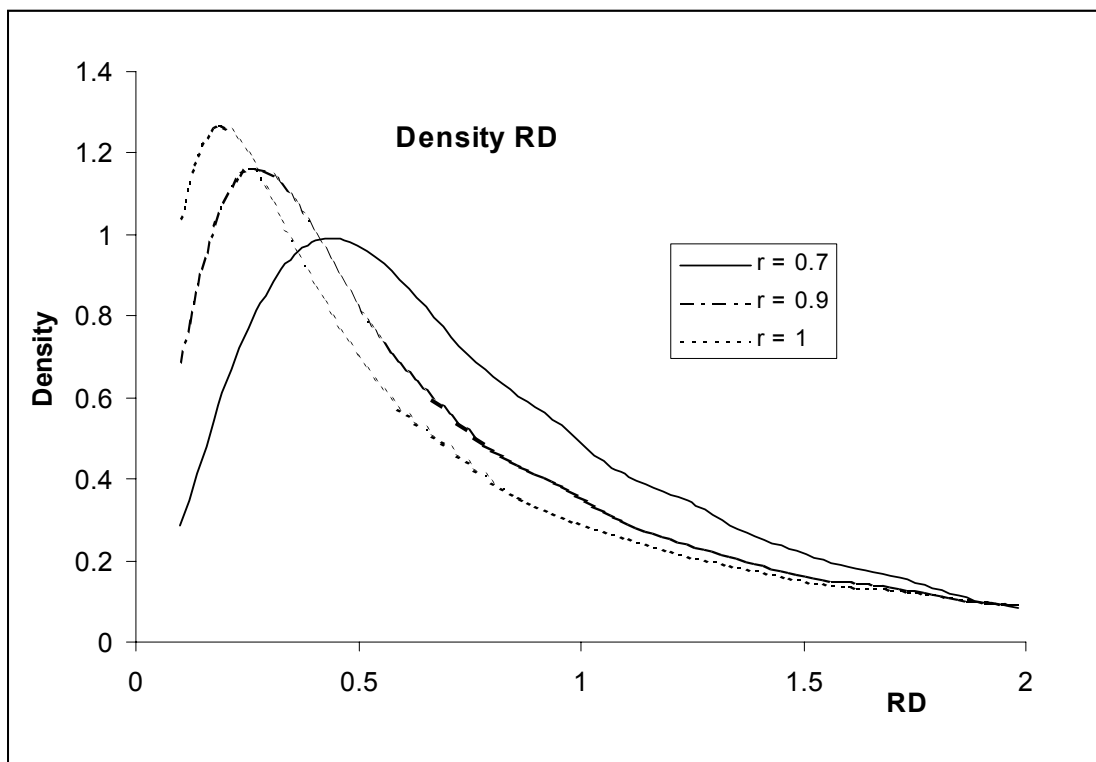


Figure 4: The density of the rd statistic for various  $\rho = 0.7, 0.9$  and  $0.999$ , at  $T = 10$ .

restricted. Since the curved model is known to be true, standard test statistics will have the usual  $\chi^2_{k-d}$  distribution with degrees of freedom equal to the number of restrictions imposed by  $\eta(\theta)$ . An obvious test statistic to use is the LR statistic for this testing problem, but requires the unrestricted estimation of  $\eta$  and this requires the cumulant function  $\kappa(\eta)$ , i.e. in termd of  $\eta$ . The cumulant function is only kown in terms of  $\theta$  and finding  $\kappa(\eta)$  is non-trivial (see Van Garderen, 2000).

Estimation under the null hypothesis is already carried out, so the obvious alternative is to use a Lagrange Multiplier (LM) test. In this case we have  $LM = \left(s - \tau(\hat{\theta})\right)' Cov_{\hat{\theta}}(s) \left(s - \tau(\hat{\theta})\right)$ . This only requires the covariance matrix of  $s$  since we have already derived  $\tau(\hat{\theta})$ . The  $LM$  statistic is a scaled measure of how far the observed  $s$  is away from its estimated expectation. There is an obvious shortcoming though, since values of  $s$  on opposite sides of  $\tau(\hat{\theta})$  give the same value of the test statistic. The observed information, however, is decreasing when  $s$  gets closer to 0 and increasing if it moves away from the critical set. We will therefore use  $rd$  to determine whether an observation  $s$  falls inside a sensitive region closer to the critical set, or falls in a more stable region away from the critical set. Hence we use the sign of  $(rd - 1)$  to determine whether the accuracy of the estimator is positively or negatively affected.

In fact, because of the asymptotic independence of  $\hat{\alpha}_1$  and  $\hat{\alpha}_2$  and the symmetries involved, we use the distribution of  $t = (\beta' S_{0\beta}, S_{\beta\beta})'$  with parameter  $\rho$  to construct the signed-LM statistic. This implies testing the (2,1)-CEM of  $t$  against a two dimensional embedding model, instead of testing the (3,2)-CEM of  $s$  against the three dimensional embedding. Moreover, we take the square root of the LM statistic because we know from the literature that signed-LR tests have distributions that are often closer to normality (see Barndorff-Nielsen and Cox 1984). So, the auxiliary statistic becomes:

$$sign - LM = sign(rd - 1) \sqrt{\left(t - \tau_t(\hat{\theta})\right)' Cov_{\hat{\theta}}(t) \left(t - \tau_t(\hat{\theta})\right)} \quad (19)$$

Figure 5 plots the simulated density of the signed-LM statistic in our model with three values of  $\rho = 0.7, 0.9$  and  $0.999$  and a sample size of 10 observations. Apart from a noticeable negative biase and right skewness, the distribution appears to be remarkably stable as  $\rho$  approaches 1, showing its superior ancillarity property relative to the  $rd$  statistic (compare with Figure 4)

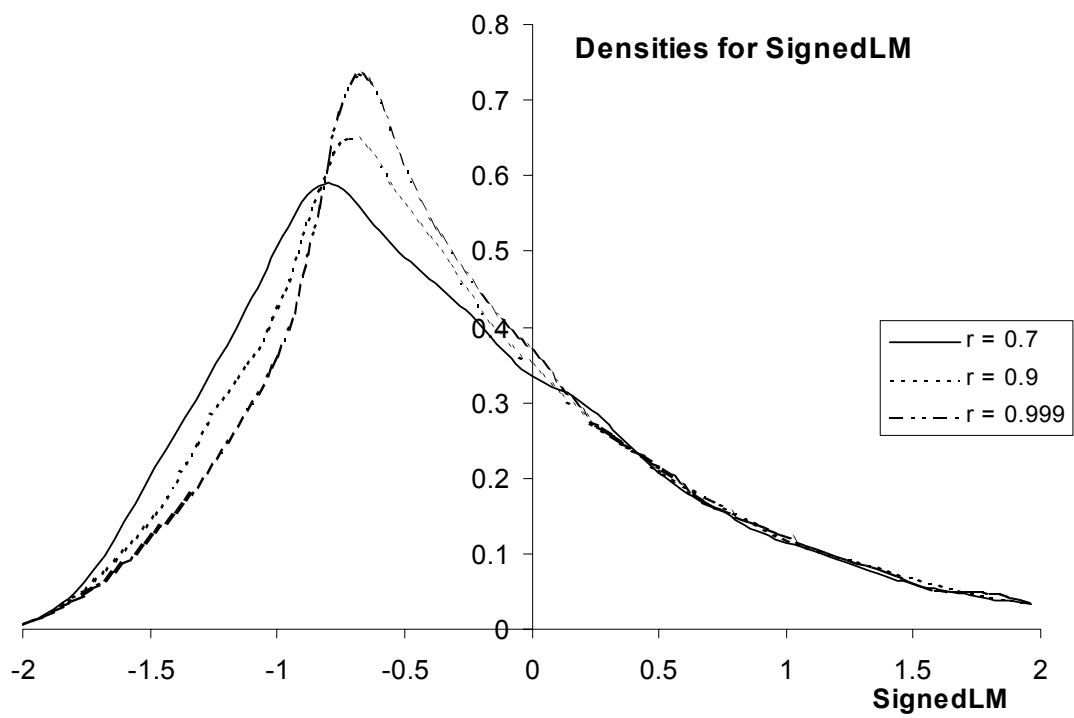


Figure 5: Density of the signed-LM statistic for different values of  $\rho = (0.7, 0.9, 0.999)$ .  $T = 10$ .

## 6 Conditional Versus Marginal Inference

### 6.1 Conditioning on the $rd$ statistic

In this section we use the statistic  $rd$  to show that the accuracy of the estimator, confidence levels, critical values, are highly dependent on the type of sample observed. If marginal inference was correct and  $rd$  were uncorrelated with  $\hat{\alpha}$ , then the conditional variance should be the same for all values of  $rd$ .

Figure 6 plots the ratio of the conditional over the unconditional variance of  $\alpha_1$  and shows that  $rd$  has a profound effect on the accuracy of the MLE for  $\alpha_1$ .

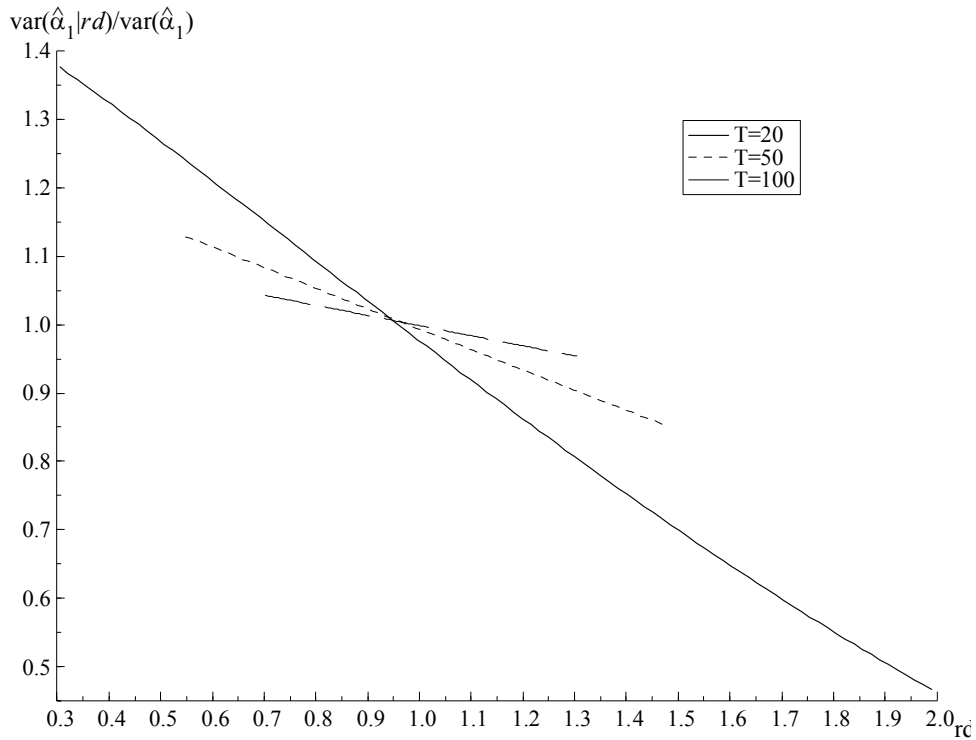


Figure 6: Variance of the MLE of  $\alpha_1$  in a bivariate CVAR model with known  $\beta$ , conditional on the value of the  $rd$  statistic, relative to its unconditional variance. The true parameters are  $\alpha = (-0.26, 0.16)'$  and  $\beta = (1, 1)'$ . Variances computed using a nonparametric Nadaraya-Watson estimator based on  $10^5$  Monte Carlo replications.

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 be three times larger than for large  $rd$  and nearly 40% larger than the unconditional variance. This means that if one is always reporting the marginal variance  $var(\hat{\alpha})$  this can give a completely false impression, depending on the value of  $rd$ . *Ex-post*, the value of  $rd$  can be calculated and reported together with the appropriate variance.

A similar picture arises if we condition on the sign-LM statistic see Figure 7. The implication is that one should not report the unconditional variance.

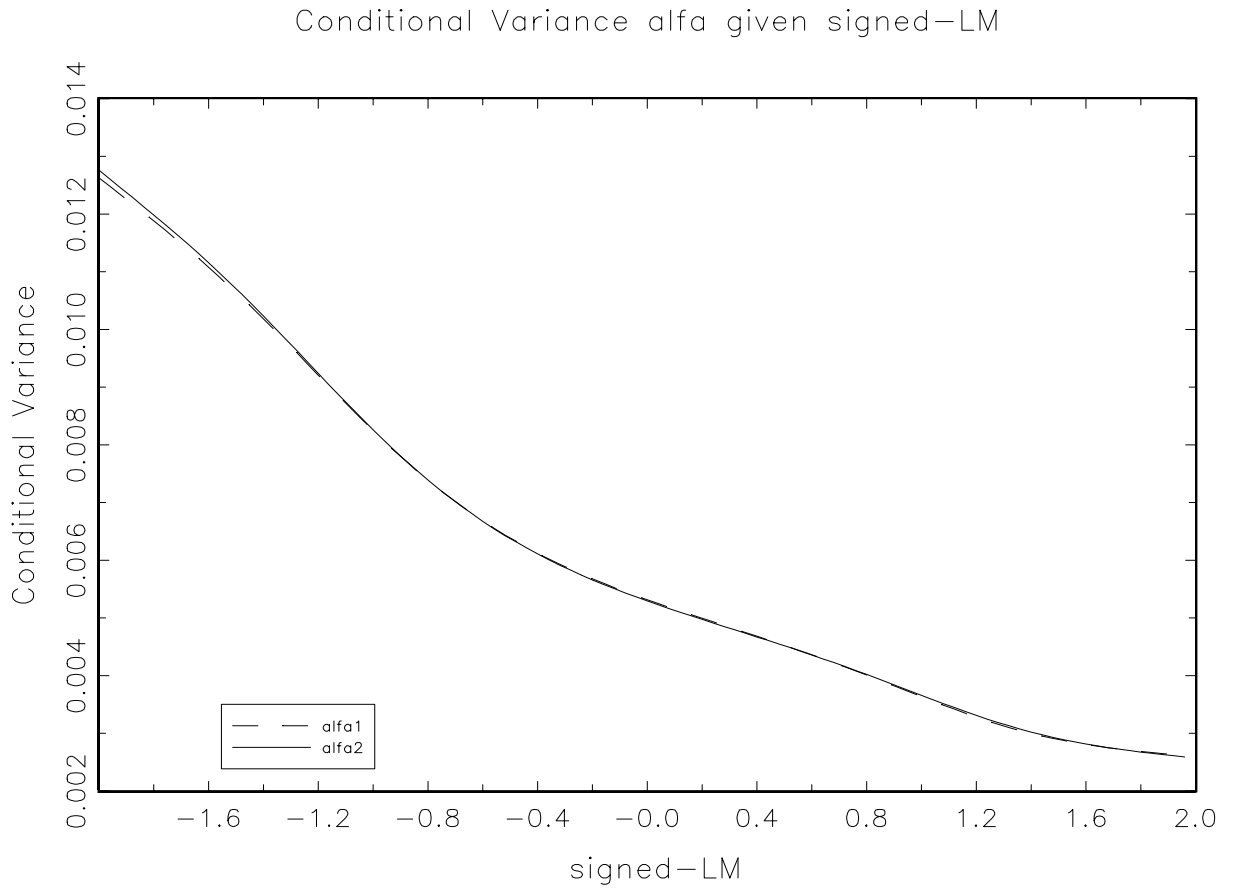


Figure 7: Conditional Variance of  $\hat{\alpha}$  given the signed-LM statistic.  $T = 25$ ;  $\alpha = (0.01, -0.01)'$ ;  $\beta = ((-1, 1)')$ ;  $\rho = 1 + \alpha'\beta = 0.98$ ; based on 100.000 simulations, estimated with Gaussian kernel estimate, bandwidth=0.4.

Next, we turn to hypothesis testing on the coefficients  $\alpha$ . 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  $\beta$  and  $\Omega$  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 (15), denoted  $W_{\text{exp}}(\alpha_0; \beta)$ ; and (ii) a Wald test based on the observed information (14), denoted  $W_{\text{obs}}(\alpha_0; \beta)$ . These are derived in the appendix, where we also show that  $W_{\text{obs}}(\alpha_0; \beta)$  is equal to the Likelihood ratio test in this case. It is also shown in the appendix that  $W_{\text{obs}}(\alpha_0; \beta)$  is invariant w.r.t. changes the parameter  $\lambda_1$ , it only varies with  $\rho$ .

Starting from  $W_{\text{exp}}$ , Figure 8 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 value based on the  $\chi^2(2)$  asymptotic approximation. Next, Figure 9 plots the *conditional* rejection frequency under the null hypothesis (NRF) of the  $W_{\text{exp}}$  statistic when using the the exact critical value and compares this with the conditional NRF of the  $W_{\text{obs}}$  test.

We see a completely different picture for in the NRFs of the two tests. The NRF for  $W_{\text{exp}}$  depend heavily on the value of  $rd$ , which implies that either the reported significance level should be adjusted or the critical value adjusted to the observed value of  $rd$ . For  $W_{\text{obs}}$  we see that the NRF is almost constant at around 10%. This means that neither the critical value nor the significance level needs adjusting. Using the observed, as opposed to the expected information results in much more reliable inference if one does not explicitly condition on one of the auxilliary variables. It seems that using the observed information implicitly does the conditioning for us. In the next section we pursue this further and show that confidence intervals based on the observed information are indeed wider for small values of the auxilliary statistic.

## 6.2 Improved Confidence Intervals

In this section we construct improved confidence intervals for  $\alpha$  based on the observed information.<sup>1</sup> In order to judge whether a confidence interval is correct we use both a marginal criterion and a conditional criterion. A proper procedure for constructing confidence intervals should on average include the true value of the parameter according to the stated confidence level. Given that sample paths can be very different and lead to very different estimation accuracy, we also want our confidence

---

<sup>1</sup>We have also applied a bias correction developed in van Garderen (2005)

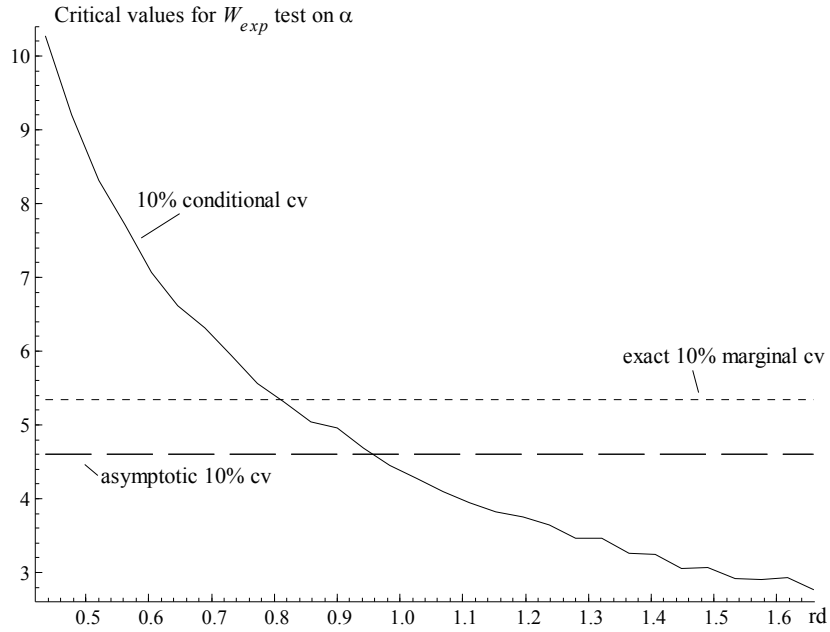


Figure 8: Three different sets of critical values for the  $W_{exp}$  test of the hypothesis  $H_0 : \alpha = \alpha_0$  when  $\beta$  is known are plotted against the  $rd$  statistic. The parameters are  $\alpha_0 = (-.19, -.11)'$ ,  $\beta = (1, 1)'$ , hence  $\rho = 0.7$ . The sample size is  $T = 20$ . 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.

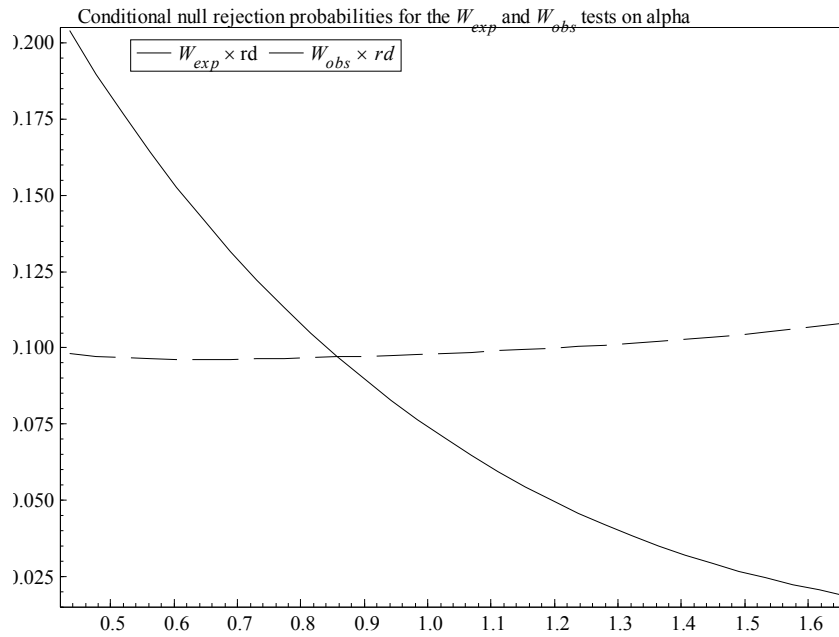


Figure 9: This graph compares the conditional null rejection frequency of the two test  $W_{exp}$  and  $W_{obs}$  of the hypothesis  $H_0 : \alpha = \alpha_0$  (when  $\beta$  is known), conditional on the  $rd$  statistic. The parameters are  $\alpha_0 = (-.19, -.11)'$ ,  $\beta = (1, 1)'$ , hence  $\rho = 0.7$ . 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. A size correction has been applied so that the unconditional rejection frequency of both tests equals their size.

intervals to reflect this, being wider when there is less information in the sample and tighter when the sample is more informative. In all cases, the confidence level should remain the same. This means that conditionally on the auxiliary statistic, the confidence intervals should change their width and have constant confidence level for all values of the conditioning statistic. It is clear that if the confidence intervals have conditionally the right level, then they will also have the correct level unconditionally, but that this does not hold the other way round. This means that the requirement that a procedure is conditionally correct puts an additional constraint on the class of procedures that we permit, but this is exactly as it should be by the arguments made in the previous sections.

We are going to use the signed-LM statistic as auxiliary statistic because of its superior ancillarity properties and because it is a good indicator of the accuracy of the estimator.

Figure 10 plots the confidence intervals as a function of signed-LM. The first confidence interval is based on the expected Fisher information. It also varies a little with signed LM because the parameter  $\rho$  in the covariance matrix is estimated. The second confidence interval is based on the observed information and we see that for negative values of signed-LM the confidence interval is much wider, because in those cases the observed sample is much more concentrated round the equilibrium than might be expected based on the estimated parameter.

The associated probabilities are shown in the Figure 11.

The graph shows that the confidence level based on the expected information varies highly with the auxiliary statistic. When signed-LM is near -2, the confidence level has dropped below 55% when it should be 95%. The confidence level of the confidence interval based on the observed information is reasonable for all values of the auxiliary. The marginal levels are 81% and 93% for the expected information- and observed information based confidence interval respectively. The marginal confidence level can be adjusted by a simple simulation methods. This however only shifts the level of the curve, but does not make it horizontal.

Finally, due to the nonlinear nature of the confidence width as a function of sign-LM, it might be that on average the confidence interval based on the observed information, or another conditional confidence interval, might be wider when averaged over all possible sign-LM, than the marginal con-

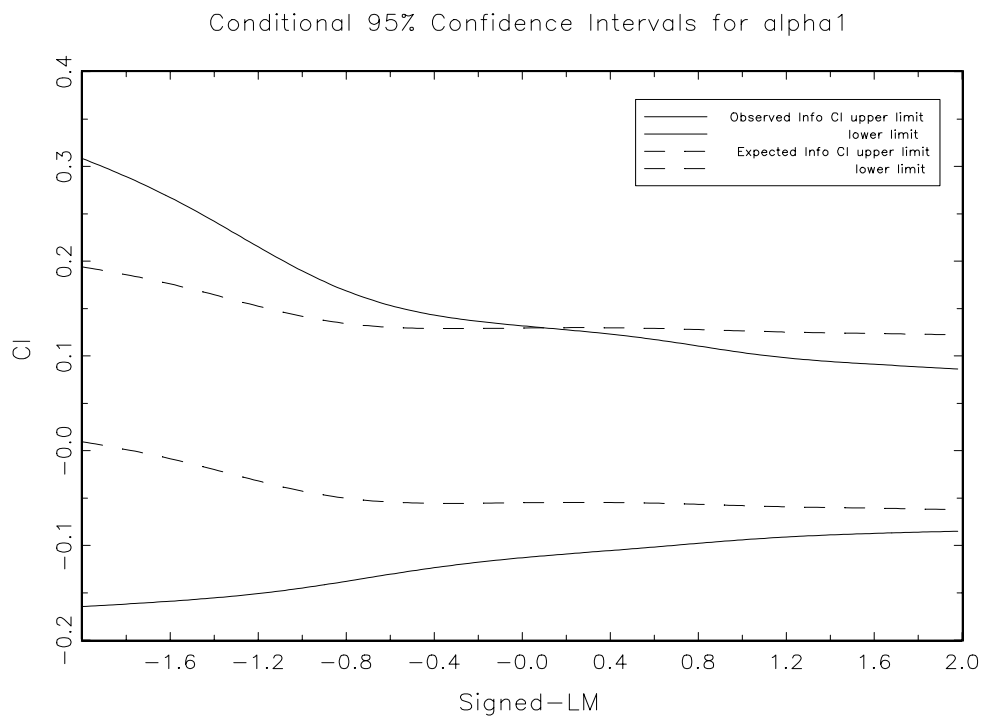


Figure 10: CI constructed using Expected and Observed Information  $T = 25$ ;  $\alpha = (0.01, -0.01)'$ ;  $\beta = (-1, 1)'$ ;  $\rho = 1 + \alpha'\beta = 0.98$ ; based on 100,000 simulations, estimated with Gaussian kernel estimate, bandwidth=0.4.

Conditional Coverage Probabilities of 95% nominal Confidence Intervals

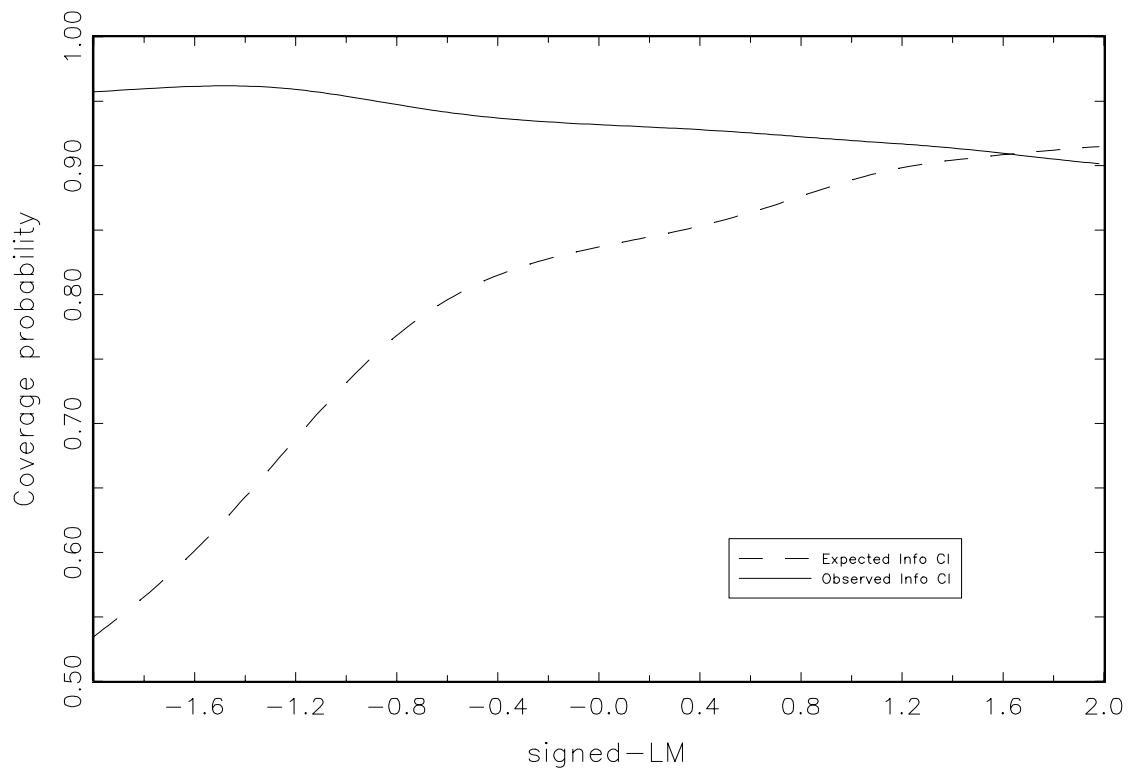


Figure 11: Probability that CI interval constructed in manner described includes true value.  $T = 25$ ;  $\alpha = (0.01, -0.01)'$ ;  $\beta = ((-1, 1)'$ ;  $\rho = 1 + \alpha'\beta = 0.98$ ; based on 100,000 simulations, estimated with Gaussian kernel estimate, bandwidth=0.4. Marginal coverage probabilities are: 0.8060 Expected info CI, 0.9288 Observed info CI

fidence interval. This, however, is not but a consequence of the marginal confidence interval getting it completely wrong for small values of sign-LM.

## 7 Conclusion

In this paper we have investigated inference on the adjustment coefficient  $\alpha$  in a simple cointegrating VAR and the effects of conditioning for inference. We have highlighted the difference between the *ex-ante* properties of the model and the *ex-post* accuracy of the estimators. We showed two aspects that influence the accuracy with which  $\beta$  can be estimated: the dispersion along the equilibrium set and the dispersion around it. Similarly, there are two aspects of the sample that determine the accuracy with which  $\alpha$  is estimated: the dispersion around the equilibrium set and the accuracy of the estimator for  $\beta$ . We have shown that the model is a curved subset in of a curved exponential model and 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*.

The starting point of the paper was that the economic theory under consideration was correct, even if the data did not necessarily support it, and that inference was still required for the adjustment parameters  $\alpha$ . It should be noted in regards to lack of support, or evidence against a theory, that if the true disequilibrium process is stationary, but with considerable persistence, there is a large probability that the estimates indicate non-stationary deviations from the equilibrium relationship. Table 1 illustrates this.

Table 1: Probability of obtaining  $1 + \beta'\alpha > 1$  in % based on 100.000 replications

		$1 + \beta'\alpha =$					
		0.75	0.85	0.9	0.95	0.99	1
T=10		6	12	18	25	33	35
	25		1	4	12	27	33
	50				3	22	32
	100					13	32

This is of course well known from the unit root literature (Dickey and Fuller 1979, Phillips 1986), but the implications here are different. Our starting point is that the economic theory is correct and that one would only reject this theory in light of very strong evidence to the contrary. Simply obtaining an estimate in the unstable parameter region is not sufficient evidence in itself to reject the theory. If in reality the equilibrium relation is stationary, there can be a very large probability of estimating a non-stationary equilibrium relation. Based on such a non-stationary estimate one would not reject a null hypothesis of non-stationarity, but this is not evidence that the theory is incorrect. Given very strong beliefs about the underlying economic theory and concerns about power of tests in this setting, one may feel justified in ignoring such a test and proceed to make inference on  $\alpha$ . One might simply infer that  $\rho$  is large and that it takes a long time for the process to return to equilibrium. This is precisely the case were conditioning matters and the paper focusses on proper inference for  $\alpha$  in such and other cases where inference is difficult. In the simulations we have therefore chosen  $\rho = 0.98$ .

We constructed a measure of proximity of the observed sample to this sensitive region, called the relative distance statistic, and showed that the accuracy of the estimator is heavily dependent on it. We also introduced a new statistic, called the signed-LM statistic that has superior ancillarity properties than the relative distance statistic. We showed 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 we pursued here 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 and inference is much more reliable.

## A Appendix

The log-likelihood function for the CVAR with known  $\beta$  and  $\Omega$  is given by

$$l(\alpha; \beta, \Omega) = -\frac{T}{2} (-2\alpha' \Omega^{-1} S_{0\beta} + S_{\beta\beta} \alpha' \Omega^{-1} \alpha) \quad (20)$$

where  $S_{\beta\beta} = \beta' S_{11} \beta$ ,  $S_{0\beta} = S_{01} \beta$ ,  $S_{01} = T^{-1} \sum_{t=1}^T R_{0t} R'_{1t}$ ,  $S_{11} = T^{-1} \sum_{t=1}^T R_{1t} R'_{1t}$ ,  $R_{0t} = \Delta Y_t - T^{-1} \sum_{t=2}^T \Delta Y_t$ ,  $R_{1t} = Y_{t-1} - T^{-1} \sum_{t=2}^T Y_{t-1}$ . The MLE for  $\alpha$  (and hence  $\rho$ ) is given by

$$\begin{aligned} \hat{\alpha} &= S_{0\beta} S_{\beta\beta}^{-1} \\ \hat{\rho} &= 1 + \beta' \hat{\alpha}. \end{aligned}$$

The observed and expected information matrices are, respectively,

$$\begin{aligned} \mathcal{J}_\alpha &= -T S_{\beta\beta} \Omega^{-1} \\ \mathcal{I}_\alpha &= -T \frac{1 - \hat{\rho}^2 - (1 - \hat{\rho}^{2T})/T}{(1 - \hat{\rho}^2)^2} (\beta' \Omega \beta) \Omega^{-1}. \end{aligned}$$

To derive the expected information, Equation (15), 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  $\beta' \Omega \beta$ . So  $ES_{\beta\beta}$  is given by

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

We will use the notation  $E_{\hat{\rho}}(S_{\beta\beta})$  below to denote that  $ES_{\beta\beta}$  is evaluated at  $\hat{\rho}$  in the above formula (21).

**Test statistics** Wald statistics for the restriction  $R' \alpha = r$  are given by

$$W = (R' \hat{\alpha} - r)' (R' V_{\hat{\alpha}} R)^{-1} (R' \hat{\alpha} - r)$$

where  $V_{\hat{\alpha}}$  is the asymptotic variance of  $\hat{\alpha}$ . We compute two versions of that statistic, based on either the observed or the expected information  $V_{\hat{\alpha}} = \mathcal{J}_{\hat{\alpha}}^{-1}$  or  $\mathcal{I}_{\hat{\alpha}}^{-1}$ , denoted  $W_{obs}$  and  $W_{exp}$  respectively. For example, for a point null  $H_0 : \alpha = \alpha_0$  these are given by

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

where  $\hat{\rho} = 1 + \beta' \hat{\alpha}$ . It can be shown that  $W_{obs}$  is also the likelihood ratio test in this case.

**Invariance to changes in  $\lambda_1$**  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  $\rho$  and  $\beta$  unchanged. But

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

so  $W_{obs}$  and hence  $LR$  are invariant to changes in  $\lambda_1$  that leave  $\rho$  and  $\beta$  constant.

**Derivation of the rd statistic** The sufficient statistic is  $s = (S'_{0\beta}, S_{\beta\beta})'$  and its expectation is  $\tau = (ES'_{0\beta}, ES_{\beta\beta})'$ . But note that  $ES_{0\beta} = \alpha ES_{\beta\beta} + ES_{\varepsilon\beta} = \alpha ES_{\beta\beta}$ , since  $ES_{\varepsilon\beta} = 0$ , and  $ES_{\beta\beta}$  is given in Equation (21). Thus,  $s - \tau(\hat{\theta}) = d (\hat{\alpha}', 1)'$  where  $d = S_{\beta\beta} - E_{\hat{\rho}} S_{\beta\beta}$ . The rd statistic measure the distance of  $s$  from  $s_{crit} = (0, 0, 0)'$  relative to the distance of  $E_{\hat{\rho}} S_{\beta\beta}$  from  $s_{crit}$ , i.e.  $rd = S_{\beta\beta} / (E_{\hat{\rho}} S_{\beta\beta})$  and the result in Equation (18) follows.

**The signed-LM statistic** The signed-LM statistic is defined as the LM test of the hypothesis that the statistic  $(\beta' S_{0\beta}, S_{\beta\beta})'$  is equal to its expected value evaluated at the MLE  $\hat{\alpha}$ . By simple affine transformation, we can re-express this in terms of the equivalent hypothesis in the AR(1) model, studied by van Garderen (1997). Let  $t = T S_{\beta\beta} (1/2, \hat{\rho})'$  so that  $\hat{\tau}_t = \tau_t(\hat{\rho}) = E_{\hat{\rho}} S_{\beta\beta} (1/2, \hat{\rho})'$ , so that  $t - \hat{\tau}_t = T (S_{\beta\beta} - E_{\hat{\rho}} S_{\beta\beta}) (1/2, \hat{\rho})' = T (rd - 1) E_{\hat{\rho}} S_{\beta\beta} (1/2, \hat{\rho})'$ . In other words, the sign-LM test is a

test of the hypothesis that  $E(rd) = 1$ , which is a restriction on the embedding model. The covariance matrix of  $s$  has been derived by van Garderen (1997b) and is adapted here as

$$\text{cov}_\rho(t) = (\beta' \Omega \beta)^2 \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

where, for  $\rho \neq 1$

$$\begin{aligned} \sigma_{11} &= \frac{\rho^{4T} + 4\rho^{2T+2} - 4\rho^2 + [4(1 - \rho^2)\rho^{2T} - \rho^4 + 1]T - 1}{2(\rho^2 - 1)^4} \\ \sigma_{12} &= (\rho^2 - 1)^{-4} \rho \{2(\rho^2 + 1)\rho^{2T} + \rho^{4T} - 2\rho^2 + (1 - \rho^2)[(3\rho^2 + 1)\rho^{2T-2} + 2]T - 3\} \\ \sigma_{22} &= (\rho^2 - 1)^{-4} \{(\rho^4 + 6\rho^2 + 1)\rho^{2T} + 2\rho^{4T+2} - \rho^4 - 8\rho^2 + (1 - \rho^2)[4(\rho^2 + 1)\rho^{2T} - \rho^4 + 4\rho^2 + 1]T - 1\} \end{aligned}$$

while for  $\rho = 1$

$$\begin{aligned} \sigma_{11} &= T(-1 + 2T - 2T^2 + T^3)/12 \\ \sigma_{12} &= (-1 + T)^2 T(1 + T)/6 \\ \sigma_{22} &= T(3 - 5T + 2T^3)/6. \end{aligned}$$

(For numerical stability, we use the formula for  $\rho = 1$  whenever  $|\rho - 1| < 10^{-3}$ ). The sign-LM statistic (19) is identical to:

$$\text{sign-LM} = \text{sign}(rd - 1) \sqrt{(t - \hat{\tau}_t)' \text{Cov}_{\hat{\rho}}(t) (t - \hat{\tau}_t)}.$$

## References

- Barndorff-Nielsen, O. E. and D. R. Cox (1984). The effect of sampling rules on likelihood statistics. *International Statistical Review* 52(3), 309–326.
- Basawa, I. V. and P. J. Brockwell (1984). Asymptotic conditional inference for regular nonergodic models with an application to autoregressive processes. *Annals of Statistics* 12(1), 161–171.
- Campbell, J. Y. and R. J. Shiller (1987). Cointegration and tests of present value models. *Journal of Political Economy* 95, 1062–1088.
- Campbell, J. Y. and R. J. Shiller (2001). Valuation ratios and the long-run stock market outlook: an update. Working Paper 8221, NBER, Cambridge, USA.

- Campbell, J. Y. and M. Yogo (2004). Efficient tests of stock return predictability. mimeo, Harvard University.
- Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* 49(1), 1–39.
- Dickey, D. A. and W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *JASA* 74, 427–31.
- Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics* 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. *Journal of Time Series Analysis* 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. *Journal of 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.
- Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–340.
- Rogoff, K. S. (1996). The purchasing power parity puzzle. *Journal of Economic Literature* 34,

647–668.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics* 54, 375–421.

Sweeting, T. J. (1992). Asymptotic ancillarity and conditional inference for stochastic processes. *Annals of Statistics* 20(1), 580–589.

Taylor, A. M. and M. P. Taylor (2004). The purchasing power parity debate. Working paper 10607, NBER, Cambridge, USA.

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

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

van Garderen, K. J. (1997b). Exact geometry of explosive autoregressive models. Discussion Paper 9768, CORE, Belgium.

van Garderen, K. J. (1999). Exact geometry of autoregressive models. *Journal of Time Series Analysis* 20(1), 1–21.

van Garderen, K. J. (2005). Bias correction for adjustment parameters in a Cointegrating VAR. Mimeo, University of Amsterdam.