On the Small Sample Properties of Weak Exogeneity Tests in Cointegrated VAR models

Ralf Brüggemann*

January 4, 2002

Abstract

We investigate the small sample properties of two types of weak exogeneity tests in cointegrated VAR models that are frequently used in applied work. The first one is the standard Likelihood Ratio (LR) test in the Johansen framework. The second test is based on mapping the cointegrated VAR model into VECM representation and then reducing the model using some model selection procedure before testing the significance of the α parameters. Results from Monte Carlo experiments indicate severe size distortions in both test types in small samples. We suggest a bootstrap version of the LR test, which can be used for size correction.

Keywords: Cointegration, weak exogeneity, bootstrap test, Subset VECM
JEL classification: C32, C51

*Humboldt Universität zu Berlin, Institut für Statistik und Ökonometrie, Spandauer Str. 1, 10178 Berlin, Germany, E-Mail: brueggem@wiwi.hu-berlin.de. I thank Helmut Lütkepohl and Andreas Stiehler for many helpful comments and suggestions. Financial support by the Deutsche Forschungsgemeinschaft, SFB 373, is gratefully acknowledged.

1 Introduction

Cointegrated vector autoregressive (VAR) models have become a standard modeling tool in applied econometric time series analysis during the last decade. Modeling multivariate cointegrated time series usually includes a number of model specification steps like, for instance, choosing the information set, selecting the lag length and the determination of the cointegration properties. Finally, modeling the short run adjustment structure, i.e. the feedbacks to deviations from the long run relations, is an important step, because it can reveal information on the underlying economic structure. Modeling the feedback mechanisms in cointegrated VAR models is typically done by testing the significance of the feedback or loading coefficients. These significance tests are often called weak exogeneity tests, because certain sets of zero restrictions imply long run weak exogeneity with respect to the cointegrating parameters. The concept of weak exogeneity was defined by Engle, Hendry & Richard (1983) and is closely related to testing the feedback coefficients. If all but one variable in a system are weakly exogenous than efficient inference about the cointegration parameters can be conducted in a single equation framework.

Some authors use tests on the feedback coefficients as ‘a form for data exploration rather than as specification testing in the strict sense’ (Johansen & Juselius (1990, p. 202)), because they have no strong *a priori* hypotheses on the feedback coefficients (see e.g. Juselius (2001), Marcellino & Mizon (2001)). They highlight the economic interpretation and impose zero restrictions on the feedback coefficients to learn more about the adjustment mechanisms in the underlying system. Choosing valid (weak exogeneity) restrictions is of major importance, because policy implications are sometimes based on the short run adjustment structure (see e.g. Juselius (2001)). In this paper we concentrate on two alternative strategies to test for long run weak exogeneity that have been frequently used in the literature. The first is a Likelihood Ratio (LR) test proposed by Johansen (see Johansen & Juselius (1990), Johansen (1995)) and is implemented in popular software packages such as PcFiml by Doornik & Hendry (1997). Numerous studies used the LR test, see *inter alia* Juselius (1995), Juselius (1996), Marcellino & Mizon (2001). The second strategy involves first mapping the cointegrated VAR into a vector error correction model (VECM) representation, reducing the parameter space by imposing additional zero restrictions on the short run dynamics and finally testing the significance of the feedback coefficients using a t- or F-test. The basic idea of this strategy is to increase the precision of the important tests on α by reducing the number of estimated parameters first. Because it involves imposing subset restrictions on the VECM in a first step, we call this procedure a Subset test. Similar modeling strategies have been used *inter alia* by Johansen & Juselius (1994), Hendry (1995, Chapter 16), Mizon (1995), Urbain (1995), Juselius (2001) and Lütkepohl & Wolters (1998, 2001).

In this paper we investigate the properties of the LR and the Subset test to see whether a particular strategy has a clear advantage over the other. To do so, we conduct a number of Monte Carlo experiments using both, data based and artificial DGPs, which we think mimic typical situations in applied macroeconometric time series analysis.

The paper is structured as follows: Section 2 describes the modeling framework as well as the tests considered and points out the main differences between them. Section 3 describes the Monte Carlo experiments and presents the main results, before concluding remarks are given in Section 4.

2 Weak Exogeneity Tests

The general modeling framework is a VAR(p) model of form

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \Xi D_t + u_t, \tag{2.1}$$

where y_t is a $K \times 1$ vector of time series, D_t a vector of deterministic terms, and $A_1, \ldots, A_p$ are $K \times K$ coefficient matrices. Ξ is the coefficient matrix associated with deterministic terms, such as a constant, trend and seasonal dummies. The disturbance u_t is a normally distributed $K \times 1$ unobservable zero mean white noise process with covariance matrix Σ_u. If the variables are cointegrated the VAR(p) model (2.1) has a vector error correction representation

$$\Delta y_t = \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + \Xi D_t + u_t, \tag{2.2}$$

denoted as VECM(p), which is obtained by subtracting y_{t-1} from both sides of (2.1) and rearranging terms (see e.g. Lütkepohl (2001) for details). In cointegrated models Π has reduced rank $r = rk(\Pi) < K$ and can be decomposed as $\Pi = \alpha\beta'$, where α and β are $K \times r$ matrices containing the loading (or feedback) coefficients and the cointegration vectors, respectively. Before the LR test for weak exogeneity is applied, researchers typically impose identifying assumptions on the cointegrating vectors β. Then testing for weak exogeneity means testing zero restrictions on the α matrix. We start by describing a LR test for general restrictions on α and β .

2.1 The Likelihood Ratio Test

Within the VECM (2.2) researchers are often interested in testing general restrictions on β and α. Boswijk (1995) suggested to express general restrictions on β as

$$\text{vec}\,\beta = H\phi + h, \tag{2.3}$$

where H is a known $Kr \times n$ matrix, ϕ is a $n \times 1$ vector containing the free parameters and h is a known $Kr \times 1$ vector corresponding to normalizing restrictions on β. Linear exclusion restrictions on α can be written as

$$\text{vec}\,\alpha' = G\gamma \tag{2.4}$$

where G is a known $Kr \times s$ matrix and γ is a $s \times 1$ vector containing the free adjustment parameters. Using this notation, fairly general restrictions can be captured, however, estimating the model under these general restrictions requires an iterative procedure. Boswijk (1995) shows that Maximum Likelihood (ML) estimators for ϕ, γ and Σ_u can be obtained by iterating on

$$\hat{\phi} = [H'(\alpha'\Sigma_u^{-1}\alpha \otimes S_{11})H]^{-1} \times (H'(\alpha'\Sigma_u^{-1} \otimes I_K)\text{vec } S_{10} - H'(\alpha'\Sigma_u^{-1}\alpha \otimes S_{11})h)$$

$$\hat{\gamma} = \left[G'(\Sigma_u^{-1} \otimes \beta' S_{11}\beta)G\right]^{-1} G'(\Sigma_u^{-1} \otimes \beta')\text{vec } S_{10}$$

$$\hat{\Sigma}_u = S_{00} - \alpha\beta' S_{10} - S_{01}\beta\alpha' + \alpha\beta' S_{11}\beta\alpha'$$

using suitable starting values. S_{ij} are the moment matrices from the reduced rank regression suggested by Johansen (1995). Once the restricted model has been estimated, a corresponding LR statistic can be calculated as

$$LR = T(\ln|\hat{\Sigma}_u^r| - \ln|\hat{\Sigma}_u|), \tag{2.5}$$

where $\hat{\Sigma}_u$ ($\hat{\Sigma}_u^r$) is the estimated covariance matrix without (with) the imposed restriction. LR is asymptotically $\chi^2(df)$ distributed with df being the degree of overidentification. Alternative algorithms to estimate a cointegrated VAR model under general restrictions on the cointegration space have been suggested by Doornik & Hendry (1997, Chapter 11).

Alternatively, if only hypotheses on α are of interest, a LR test statistic can be easily computed without using an iterative procedure. In fact, simple hypotheses on the feedback coefficients can also be expressed as

$$\alpha = G\gamma \tag{2.6}$$

where γ is a $s \times r$ matrix of free parameters and G is a known $K \times s$ matrix of ones and zeros. For example, if we have a four dimensional system ($K = 4$) with one cointegrating relation ($r = 1$) and wish to test whether the error correction term enters the second equation, i.e. we test weak exogeneity of the second variable, the null hypotheses can be written as

$$\alpha = \begin{bmatrix} \alpha_1 \\ 0 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}. \tag{2.7}$$

Since we only impose restrictions on α, hypotheses like (2.6) can be tested by solving a modified eigenvalue problem as described by Johansen (1995, Chapter 8). The corresponding LR statistic is computed as

$$LR = T\sum_{i=1}^{r} \ln[(1 - \tilde{\lambda}_i)/(1 - \hat{\lambda}_i)] \tag{2.8}$$

where $\hat{\lambda}_i$ ($\tilde{\lambda}_i$) are the eigenvalues calculated without (with) the restriction. LR is again asymptotically $\chi^2(df)$ distributed where df denotes the degree of overidentification. Note that testing hypotheses like (2.6) is only of interest if $r = 1$, because only then no additional identifying restrictions on β are needed. Therefore, in the Monte Carlo simulations presented in Section 3 we consider simple restrictions of type (2.6) in systems where $r = 1$, and more general restrictions (2.3) and (2.4) if $r > 1$.

2.2 The Subset Test

Suppose the researcher has imposed (over-)identifying restrictions on the cointegrating vectors β to give the long run relations an economic interpretation. In order to have a more parsimonious description of the data, he may want to impose additional restrictions. To achieve this, he typically maps the model into VECM representation by fixing β and computing the error correction term $ec_t = \beta' y_t$, so (2.2) becomes

$$\Delta y_t = \alpha ec_{t-1} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + \Xi D_t + u_t, \tag{2.9}$$

and uses some model selection algorithm to delete variables from the model. Typically, one selects zero restrictions on $\Gamma := (\Gamma_1 : \cdots : \Gamma_{p-1})$ first and finally decides on the significance of the α parameters. As already pointed out in the introduction, the intuition for this procedure is to reduce the number of parameters first and thereby increasing the precision of the important tests on α. Because it involves imposing subset restrictions on the VECM in a first step, we call this procedure a Subset test. By writing the model in VECM form, all variables are in $I(0)$ space and conventional t- or F-tests can be used to conduct the weak exogeneity test. The implementation of the this 'test' differs across studies, because of different model selection methods. In fact, a number of procedures to delete variables from a system have been proposed in the literature. For example, Hendry (1995) suggested the general-to-specific (Gets) methodology that is based on a sequence of t-tests and a number of misspecification tests to ensure an adequate final model. The method is frequently used in empirical modeling. More recently, Krolzig & Hendry (2001) have implemented the Gets method into an automated computer algorithm (PcGets) and Krolzig (2001) has shown the usefulness for reducing VAR models. Alternative methods typically include procedures that are based on information criteria (see e.g. Lütkepohl (1991, Chapter 5)). For the VAR framework Brüggemann & Lütkepohl (2001) compare different subset modeling methods and find that system based procedures are not superior to methods based on the single equations. Therefore, here we use a single equation strategy that amounts to sequentially deleting variables with lowest absolute t-ratios until all are greater than some threshold value τ. More formally, we can write the k-th equation of (2.9) as

$$\Delta y_{kt} = \alpha_k ec_{t-1} + x_{1t}\vartheta_1 + \cdots + x_{Mt}\vartheta_M + u_{kt}, \qquad t = 1, \ldots, T \tag{2.10}$$

where α_k is the k-th row of α and x_{mt} $(m = 1, \ldots, M)$ denotes all other regressors in that equation. Let $t_m^{(j)}$ denote the t-ratio from an OLS estimation associated with ϑ_m in the j-th step of the procedure. In reduction step j we delete x_{mt} if

$$|t_m^{(j)}| = \min_{i=1,\ldots,M-j+1} |t_i^{(j)}| \qquad \text{and} \qquad |t_m^{(j)}| \leq \tau.$$

The algorithm stops if all $|t_m^{(j)}| > \tau$. The critical value

$$\tau = \tau_j = \{[\exp(c_T/T) - 1](T - M + j - 1)\}^{1/2} \tag{2.11}$$

is a function of the sample size T, the number of initial regressors M and the reduction step j. τ also depends on c_T, which varies with the choice of a typical information criterion (AIC: $c_T = 2$, HQ: $c_T = 2 \log \log T$ and SC: $c_T = \log T$). The critical value in this testing procedure (TP) for each test is chosen according to (2.11) to insure that a related procedure based on a sequential elimination of regressors (SER) using information criteria leads to the same final model. Therefore, we refer to this strategy as SER/TP in the following. Krolzig (2001) compares the PcGets and SER/TP algorithm and finds that PcGets has slightly better size properties. However, the comparison is based on one specific DGP and clearly a more systematic comparison is necessary. Results available so far indicate no major advantage for the more sophisticated PcGets algorithm.

In the Monte Carlo simulations of the Subset test, we use SER/TP together with AIC to impose zero restrictions on Γ, because this strategy has performed relatively well in the comparison of Brüggemann & Lütkepohl (2001). The steps necessary in the Subset test can be summarized as follows:

- map cointegrated VAR to VECM representation (2.9)
- use SER/TP to impose zero restrictions on Γ (exclude α's from search)
- test (weak exogeneity) hypotheses on α using t- or F-tests

2.3 Small Sample Correction and Bootstrap Tests

In this paper we are mainly concerned with the test performance in finite sample situations faced by applied econometricians. The distribution of the LR test is only an asymptotic one and may be quite misleading. In fact, in some studies concerned with testing linear restrictions on cointegrating vectors the LR test has severe size distortions (see e.g. Gredenhoff & Jacobson (2001) and references therein). A similar problem may be present when testing hypotheses on α, as pointed out by Podivinsky (1992). Therefore, it may be useful to consider some kind of small sample adjustment. For the LR tests we use two small sample modifications. The first method is a simple degrees of freedom correction as suggested by Sims (1980, p. 17)

$$LR^M = (T - k)LR/T \rightarrow \chi^2(df), \tag{2.12}$$

where k is a correction for the degrees of freedom as discussed below. LR^M has the same asymptotic distribution as LR, but is less likely to reject in small samples. Moreover, we know from standard linear regression models that the F-version of the LR test is better behaved in small samples. Therefore, we also consider

$$F = LR/df \approx F(df, T-k), \tag{2.13}$$

where k again is some degrees of freedom correction. (2.13) is approximately F distributed (see Lütkepohl (1991, p. 123)). Choosing k is crucial because it affects the test statistic or the degrees of freedom and hence the test decision. A number of proposals have been made in the literature. Typically, k is the approximate number of estimated parameters in the system (2.2) or the number of estimated parameters in one equation of the system. Podivinsky (1992) suggested to choose k equal to the total number of parameters in Π and Γ of (2.2). This correction worked well in the specific example used in his simulation. However, if K and p get large relative to the sample size T, one ends up with negative degrees of freedom. We therefore choose k such that it equals the approximate number of estimated coefficients in one equation of (2.2), as suggested by Lütkepohl (1991, p. 123) in the VAR context. More precisely, we let

$$k = r + Kr + K(p-1) + 1 \tag{2.14}$$

where the first right hand side term gives the number of α parameters per equation, the second the number of β parameters, and the last two the number of parameters in Γ and Ξ in one equation of the system.[1] One might argue, that the value of k is too large because it actually includes the number of β parameters (Kr) of the whole system. On the other hand, choosing k too small reduces the effect of the correction. Using (2.14) we also sometimes get a negative value for $T-k$. For instance, for $T=30$, $K=4$, $r=3$ and $p=5$ we find $T-k=-2$. Therefore, we use

$$k = r + r + K(p-1) + 1 \tag{2.15}$$

as an alternative measure for the approximate number of estimated parameters in one equation. It is yet unclear which k is the preferred correction, but the discussion already highlights one major drawback of small sample corrections of test statistics: The choice of k is somewhat arbitrary and the optimal k might depend on the specific properties of the considered system. To avoid the problem of choosing k, we can alternatively estimate a bootstrap version of the LR statistic using the following procedure:

1. Estimate the cointegrated model under H_0, record the LR test statistic and save the parameters $\hat{\alpha}, \hat{\beta}, \hat{\Gamma}, \hat{\Xi}$.

2. Draw bootstrap residuals $u_1^*, \ldots, u_T^*$ with replacement from estimated centered residuals $\hat{u}_1 - \bar{u}, \ldots, \hat{u}_T - \bar{u}$, where $\bar{u} = (1/T)\sum_{t=1}^{T} \hat{u}_t$.

3. Generate bootstrap time series y_t^* recursively using $\hat{\alpha}, \hat{\beta}, \hat{\Gamma}, \hat{\Xi}$ and given presample values $y_{-p+1}, \ldots, y_0$.

[1] Here we consider the number of parameters in a model with intercept only.

4. Estimate the LR statistic using the bootstrap time series.

5. Repeat 2. - 4. many times.

6. Compare LR with the $(1-\alpha)$-quantile of the bootstrap distribution.

Since LR is asymptotically pivotal, this might be a classical situation where the bootstrap is useful (see Horowitz & Savin (2000)). To assess the small sample performance of all tests we use Monte Carlo simulation techniques.

3 Monte Carlo Evidence

The aim of the Monte Carlo study is to find out whether the LR test (possibly corrected in small samples) or the Subset test is the best way to model the loading coefficients α. Therefore, we first compare empirical sizes of the standard LR test (2.5) and (2.8) with those of the Subset test. Then, we present results for the modified statistics (2.12) and (2.13) and compare them to results of the bootstrap test.

We use two types of DGPs. DGPs of the first type are data based, i.e. in the Monte Carlo we use parameters that have been estimated by fitting a VAR model to real world data. We do so to get a realistic DGP in terms of dimension, dynamics and cointegration properties. The second set of DGPs are artificial ones, because this enables us to change the properties of the cointegration space (i.e. the cointegrating rank).

The parameters for the data based DGPs have been generated by fitting cointegrated VAR models to a data set for the U.K. monetary sector. The data were first analyzed by Hendry & Ericsson (1991) and later reconsidered by *inter alia* Johansen (1992), Hendry (1995) and Doornik, Hendry & Nielsen (1998). We use the data because we think they represent a typical system analyzed with cointegration techniques. The data include the log of M1 money (m), the log of TFE deflator (p), the log of real TFE in 1985 prices (inc) and a measure of opportunity costs for holding money (R), the typical ingredients for a money demand analysis. We estimate VAR models under the rank restriction $r = 1$ with lag order p ranging from 1 to 5 including an unrestricted constant. In addition, we impose the weak exogeneity restrictions $\alpha_2 = \alpha_3 = \alpha_4 = 0$ and save $\hat{\alpha}, \hat{\beta}, \hat{\Gamma}_1, \ldots, \hat{\Gamma}_{p-1}, \Xi$ and $\hat{\Sigma}_u$. To conserve space, we only list the cointegration parameters in Table 1. The remaining parameters are available from the author on request. Obviously the β parameters vary only very little when increasing the lag length. The data based DGPs therefore have very similar cointegration properties.

To assess the influence of the cointegrating rank, we have also considered artificial DGPs and give the cointegration parameters in the bottom panel of Table 1. Using δ_1 and δ_2 we can vary the cointegrating rank of the system between one and three. For instance, if $\delta_1 = \delta_2 = 1$ then the cointegrating rank of the DGP is $r = 3$. For each choice of r we also consider different dynamics by varying p from 1 to 5. For a given p, we use the same values for $\Gamma_1, \ldots, \Gamma_{p-1}, \Xi$ and Σ_u as in the corresponding empirical U.K. money demand system.

Table 1: DGPs for Monte Carlo Experiments

DGP	Cointegration Parameters	p	D_t
Data Based DGPs:			
(a)	$\alpha' = \begin{pmatrix} -0.096 & 0 & 0 & 0 \end{pmatrix}, \beta' = \begin{pmatrix} 1 & -0.842 & -1.541 & 5.580 \end{pmatrix}$	1	CON
(b)	$\alpha' = \begin{pmatrix} -0.103 & 0 & 0 & 0 \end{pmatrix}, \beta' = \begin{pmatrix} 1 & -0.911 & -1.365 & 6.390 \end{pmatrix}$	2	CON
(c)	$\alpha' = \begin{pmatrix} -0.097 & 0 & 0 & 0 \end{pmatrix}, \beta' = \begin{pmatrix} 1 & -0.930 & -1.356 & 6.813 \end{pmatrix}$	3	CON
(d)	$\alpha' = \begin{pmatrix} -0.128 & 0 & 0 & 0 \end{pmatrix}, \beta' = \begin{pmatrix} 1 & -0.950 & -1.272 & 6.905 \end{pmatrix}$	4	CON
(e)	$\alpha' = \begin{pmatrix} -0.153 & 0 & 0 & 0 \end{pmatrix}, \beta' = \begin{pmatrix} 1 & -0.957 & -1.234 & 6.891 \end{pmatrix}$	5	CON
Artificial DGPs:			
	$\alpha = \begin{pmatrix} -0.1 & 0 & 0 \\ 0 & -0.2\delta_1 & 0 \\ 0 & 0 & -0.3\delta_2 \\ 0 & 0.1\delta_1 & 0 \end{pmatrix}, \beta' = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1\delta_1 & -1\delta_1 & 0 \\ 0 & 0 & 1\delta_2 & -1\delta_2 \end{pmatrix}$		
(f)–(k)	$\delta_1 = 0, \delta_2 = 0$	1– 5	CON
(l)–(p)	$\delta_1 = 1, \delta_2 = 0$	1– 5	CON
(q)–(u)	$\delta_1 = 1, \delta_2 = 1$	1– 5	CON

Note: Parameter values for $\Gamma_1, \dots, \Gamma_{p-1}$ as well as Ξ correspond to estimated values from the U.K. money demand model with lag order p. CON is an unrestricted constant.

From both DGP types, we have generated 1000 replications of length T ranging from 30 to 1000 and fit a cointegrated VAR model. In this paper, we are primarily interested in the test for weak exogeneity and therefore estimate the models using the correct lag length p and the correct cointegration rank r. Moreover, for DGPs with $r > 1$, we also estimate the model using 'correct' identifying assumptions that exactly identify all cointegration vectors. For example, when using DGPs (l)–(p), we impose the following restrictions on β:

$$\beta' = \begin{pmatrix} 1 & -1 & * & * \\ * & 1 & -1 & * \end{pmatrix}, \tag{3.1}$$

where $*$ indicates free elements of β. In principle, imposing correct p, r and identifying assumptions should improve the performance of the weak exogeneity tests relative to the situation faced in practice. Nevertheless, we conduct the experiments as if p, r and the identifying assumptions were known, because we want to assess the test itself and not other factors affecting its performance.

Using the test statistics (2.5),(2.8), (2.12) and (2.13) we then test the following hypotheses

- $H_0^1 : \alpha_{21} = 0$
- $H_0^2 : \alpha_{21} = \alpha_{31} = 0$
- $H_0^3 : \alpha_{21} = \alpha_{31} = \alpha_{41} = 0$

Note that in this case, df, the degree of overidentification is simply the number of α coefficients to be tested, because β is exactly identified. If we use the Subset test, we first reduce the short run dynamics, i.e. we impose zero restrictions on Γ, before conducting a t-test for H_0^1 and F-tests for H_0^2 and H_0^3.

For all tests, we record the relative rejection frequencies of the hypotheses, i.e. the empirical size. Given the nominal size of $P = 0.05$ and 1000 Monte Carlo replications the standard error of the Monte Carlo is $\sqrt{P(1-P)/1000}$ and hence the 2 standard error confidence interval around the nominal size is $(0.036; 0.064)$. All simulations were performed using GAUSS v3.2 for Windows. For convenience we present the results graphically and discuss them in the following sections.

3.1 LR- vs. Subset Test

Figure 1 shows the empirical size (i.e. the relative rejection frequency of the true H_0) of the standard LR and the Subset test for sample sizes T ranging from 30 to 100 and for the data based DGPs (a) to (e). The confidence band around the nominal size is indicated by the dashed horizontal lines. For $T = 30$ and $T = 60$, the empirical sizes of both tests are clearly above the desired level of 5% in all cases. We find that the empirical size increases with the lag length of the system. Although applied researchers most likely avoid to fit large lag dynamics to short time series, we have included results for $T = 30$ to get a sense on how bad things can get. In these very small samples, we find severe size distortions for both test types with sizes ranging anywhere between 13 and 99 %. We also find that the performance of both tests deteriorates with increasing degrees of overidentification, i.e. the empirical size increases when moving from H_0^1 to H_0^2 and H_0^3. In almost all cases the LR test has sizes considerably closer to the nominal level than the Subset test. For $p = 1$, however, we find that the Subset test and the LR test perform very similar. This can be expected, because there is no search for zero restrictions if $p = 1$ in the Subset test and hence, both tests are very similar. In contrast, for $p > 1$, we search for zero restrictions in Γ when using the Subset test and we consistently get more severe size distortions than for LR. We also observe that the Subset test performs increasingly worse relative to LR when increasing the lag length p. This might be an indication that the performance of the Subset test is adversely affected by the number of model selection steps involved. Interestingly, this phenomenon is not simply a small sample problem. To see this, we have repeated the Monte Carlo for $T = 200, 500, 1000$ and give results in Figure 2. For the LR test we find empirical sizes fairly close to the 5% level and almost always within the confidence band. In other words, the LR test works very well in sufficiently large samples. In contrast, the Subset test is still severely oversized even if $T = 1000$, a case virtually never encountered in real world macroeconometric time series modeling. We have also repeated the Monte Carlo comparison for the artificial DGPs and basically find the same picture. To conserve space we do not show these results here.

To sum up, the results suggest to use the LR test for hypotheses on β and α first, before imposing exclusion restrictions on Γ. However, we also find size distortions of the LR test in small samples and hence, it may be worthwhile to consider the small sample modifications and the bootstrap version discussed in Section 2.3.

3.2 Small Sample Correction and the Bootstrap LR Test

Figure 3 again shows the size of LR now with additional results for LR^M, F and LR^*. Results for the LR^M have been obtained by using the correction (2.14). We do not show results for using (2.15), because these corrections are less effective and perform worse than the ones shown. Results for the bootstrap version are based on 200 bootstrap draws in each Monte Carlo replication.

The results for the data based DGPs (a)-(e), Figure 3, show that the F approximation reduces the actual size only very little and there is only a minor improvement compared to LR. In contrast, the LR^M test can reduce the empirical size quite substantially, although the resulting size is still larger than the nominal level. Similarly, LR^* brings down the empirical size very close to the desired level with the exception of the case of $T = 30$ and large p. As pointed out before, these cases are rarely encountered in practice. In the majority of cases, the bootstrap version LR^* has the best size properties of all considered test statistics, although for the DGPs (a)-(e) the difference between LR^* and LR^M is small.

To get a more informative and general picture of the relative performance of all four tests and to investigate the influence of the cointegration rank, we also show results for the artificial DGPs.

To be more precise, for DGPs (f)-(k), i.e. $r = 1$, we plot the results in Figure 4. Compared to the oversized LR statistic, we again find that the F approximation only leads to very small improvements. On the other hand, LR^M works relatively nicely when p is large. This can be explained by the fact that the correction is a function of p. In many cases, we find LR^* to have the smallest empirical size, often fairly close to the desired level (especially for $T = 60$ and $T = 100$).

If we increase the cointegrating rank of the underlying DGP to $r = 2$ and $r = 3$ (Figures 5 and 6), we once more find evidence for massive size distortion for the LR test. In contrast to what has been said, we now find that neither F nor LR^M work satisfactorily. For $T = 30$, LR^M has a tendency to overcorrect, i.e. the modified tests have empirical sizes that is significantly smaller than the nominal size. In some cases, it even drops down to zero (see Figure 5 and 6, column 1). This behavior is normally associated with a loss in power against alternatives and hence, not desirable. In most cases, LR^* does again the best job and its empirical size is very close to the desired level. It is interesting to note, that the bootstrap works better when $r = 2$ or $r = 3$. This fact may be due to the exact identifying assumptions. Obviously, it pays to introduce the right structure on the long run coefficients. The small sample corrected versions, however, only work sometimes and often result in empirical sizes smaller than the nominal size.

We compare the power of the bootstrap test LR^* to the size-adjusted power of the standard LR test to check whether using the bootstrap is associated with loss in power against alternatives. For the power simulations we use DGPs from Table 1 and let α_{21} vary according to

$$\alpha_{21} = \frac{b}{\sqrt{T}}, \qquad b \in \{0, 0.2, 0.4, \ldots, 3\}, \tag{3.2}$$

and record the rejection frequencies for hypotheses H_0^1, H_0^2, H_0^3. To make results comparable, we compute the size-adjusted power of the standard LR test, i.e. we adjust the critical values such that the empirical size of LR is exactly 0.05. Moreover, we only compare cases, where the empirical size of the bootstrap test LR^* is very close to 0.05. As a typical result from the power simulations, Figure 7 shows the comparison for DGP (s), i.e. a DGP with cointegrating rank $r = 3$ and $p = 3$. For H_0^1 there is virtually no difference between the power of LR^* and LR, whereas for H_0^2 and H_0^3 the bootstrap test is more powerful than LR. For $T = 100$, however, the differences are again very small. The comparison indicates that using the bootstrap is not associated with a loss in power. In fact, the bootstrap test has sometimes even more power than the standard test.

Overall, the results from the Monte Carlo experiments indicate that the bootstrap test LR^* has the most favorable size properties and its power is comparable to the standard LR test. Since the performance of the LR test depends on the sample size, the dynamics p, the degree of overidentification and the cointegrating properties of the underlying system, LR^* does a better job than the suggested small sample modifications. Consequently, it is advisable to use LR^* for testing restrictions on α.

4 Conclusions

We have considered different methods for testing the significance of loading (or feedback) coefficients in cointegrated VAR models. Testing hypotheses on the loading coefficients is also closely related to the concept of weak exogeneity with respect to the long run parameters. We argue that these tests are important to identify and interpret the short run structure in a cointegrated VAR model in a meaningful way. Therefore, we are interested in the size properties of these tests in small samples. Both test types have frequently been used in applied work. The first one is the standard LR test in the Johansen framework. The second test is based on mapping the cointegrated VAR model into VECM representation and then reducing the model using some model selection procedure before testing the significance of the α parameters.

We have conducted a number of Monte Carlo experiments and find considerable size distortions in small sample situations. More precisely, both tests reject the true null too often. Only when no model selection of the short run dynamics is conducted within the Subset test, it performs similar to the LR test. In all other cases, we find that the LR test has more favorable size properties in small samples. The Monte Carlo study reveals that the size distortions for the Subset test are not simply a small sample phenomenon, but a problem that does not vanish in large samples. Obviously, these size distortions are related to the model selection of lagged differences. Overall, the results from the comparison suggest to use the LR- rather than the Subset test.

Since the LR test also has size distortions in small samples, we also investigated the performance of two small sample modifications and a bootstrap version of this tests. We find that the F approximation cannot successfully reduce the size distortions, while a crude small sample modification works in some of the considered cases. In some other cases, however, LR^M has a tendency to overcorrect the size distortion. Since the performance of tests on α typically depends on a number of factors, such as sample size, cointegrating properties, lag length, we suggest to use the bootstrap test LR^* in applied work, because it provides the most reliable size correction and hence, the most favorable small sample behavior.

The results have important implications for empirical model building: Testing the significance of loading parameters should be done within the Johansen framework, possibly using a bootstrap corrected test. In other words, the long run parameters β and the short run adjustment structure α should be modeled carefully in a first step, before mapping the model to $I(0)$ and imposing additional restrictions on coefficients Γ for the lagged differences. Given the properties of the Subset test procedure, it can most likely not successfully be used as 'a form of data exploration' to identify a meaningful adjustment structure. Researchers using a strategy similar to the Subset test most certainly use additional modeling tools, such as a battery of misspecification tests to derive their final model. Therefore, in practice, researchers may revise the model specified by the Subset test based on additional evidence and expert knowledge. Clearly, we cannot mimic this behavior in our Monte Carlo comparison. Therefore, the Subset test may work better in practice than our simulation results suggest. In practice, it may thus be advisable to start with the LR^* test for exclusion restrictions on α. Then, one may set up the VECM, impose restrictions on Γ and finally check again the significance of α parameters to derive the final model.

References

Boswijk, H. P. (1995). Identifiability of cointegrated systems, *Tinbergen Working Paper* .

Brüggemann, R. & Lütkepohl, H. (2001). Lag selection in Subset VAR models with an application to a U.S. monetary system, *in* R. Friedmann, L. Knüppel & H. Lütkepohl (eds), *Econometric Studies - A Festschrift in Honour of Joachim Frohn*, LIT: Münster, pp. 107–128.

Doornik, J. A. & Hendry, D. F. (1997). *Modelling Dynamic Systems Using PcFiml 9.0 for Windows*, Thomson Publishing, Institute of Economics and Statistics, University of Oxford.

Doornik, J. A., Hendry, D. F. & Nielsen, B. (1998). Inference in cointegrating models: UK M1 revisited, *Journal of Economic Surveys* **12**(5): 533–571.

Engle, R., Hendry, D. F. & Richard, J.-F. (1983). Exogeneity, *Econometrica* **51**: 277–304.

Gredenhoff, M. & Jacobson, T. (2001). Bootstrap testing linear restrictions on cointegration vectors, *Journal of Business and Economic Statistics* **19**(1): 63–71.

Hendry, D. F. (1995). *Dynamic Econometrics*, Oxford University Press, Oxford.

Hendry, D. F. & Ericsson, N. (1991). Modeling the demand for narrow money in the United Kingdom and the united states, *European Economic Review* **35**: 833–881.

Horowitz, J. L. & Savin, N. (2000). Empirically relevant critical values for hypotheses tests: A bootstrap approach, *Journal of Econometrics* **98**: 375–389.

Johansen, S. (1992). Weak exogeneity and cointegration in U.K. money, *Journal of Policy Modeling* **14**(3): 313–334.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autogressive Models*, Oxford University Press.

Johansen, S. & Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration – with applications to the demand for money, *Oxford Bulletin of Economics and Statistics* **52**(2): 169–210.

Johansen, S. & Juselius, K. (1994). Identification of the long-run and the short-run structure - An application to the islm model, *Journal of Econometrics* **63**: 7–36.

Juselius, K. (1995). Do purchasing power parity and uncovered interest rate parity hold in the long run? An example of likelihood inference in a multivariate time-series model, *Journal of Econometrics* **69**: 211–240.

Juselius, K. (1996). An empirical analysis of the changing role of the German Bundesbank after 1983, *Oxford Bulletin of Economics and Statistics* **58**(4): 791–?

Juselius, K. (2001). European integration and monetary transmission mechanisms: The case of Italy, *Journal of Applied Econometrics* **16**: 314–358.

Krolzig, H.-M. (2001). General-to-specific reductions of vector autoregressive processes, *in* R. Friedmann, L. Knüppel & H. Lütkepohl (eds), *Econometric Studies - A Festschrift in Honour of Joachim Frohn*, LIT: Münster, pp. 129–157.

Krolzig, H.-M. & Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures, *Journal of Economic Dynamics and Control* **25**: 831–866.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*, Berlin: Springer-Verlag.

Lütkepohl, H. (2001). Vector autoregressions, *in* B. Baltagi (ed.), *Companion to Theoretical Econometrics*, Blackwell, Oxford, pp. 678–699.

Lütkepohl, H. & Wolters, J. (1998). A money demand system for German M3, *Empirical Economics* **23**: 371–386.

Lütkepohl, H. & Wolters, J. (2001). The transmission of german monetary policy in the pre-Euro period, *Discussion Paper 87*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Marcellino, M. & Mizon, G. E. (2001). Small-system modelling of real wages, inflation, unemployment and output per capita in italy, *Journal of Applied Econometrics* **16**: 359–370.

Mizon, G. E. (1995). Progressive modeling of macroeconomic time series - the LSE methodology, *in* K. D. Hoover (ed.), *Macroeconometrics:*, Kluwer Academic Publisher, pp. 107–170.

Podivinsky, J. M. (1992). Small sample properties of tests of linear restrictions on cointegrating vectors and their weights, *Economics Letters* **39**: 13–18.

Sims, C. A. (1980). Macroeconomics and reality, *Econometrica* **48**: 1–48.

Urbain, J.-P. (1995). Partial versus full system modelling of cointegrated systems: An empirical illustration, *Journal of Econometrics* **69**: 177–210.

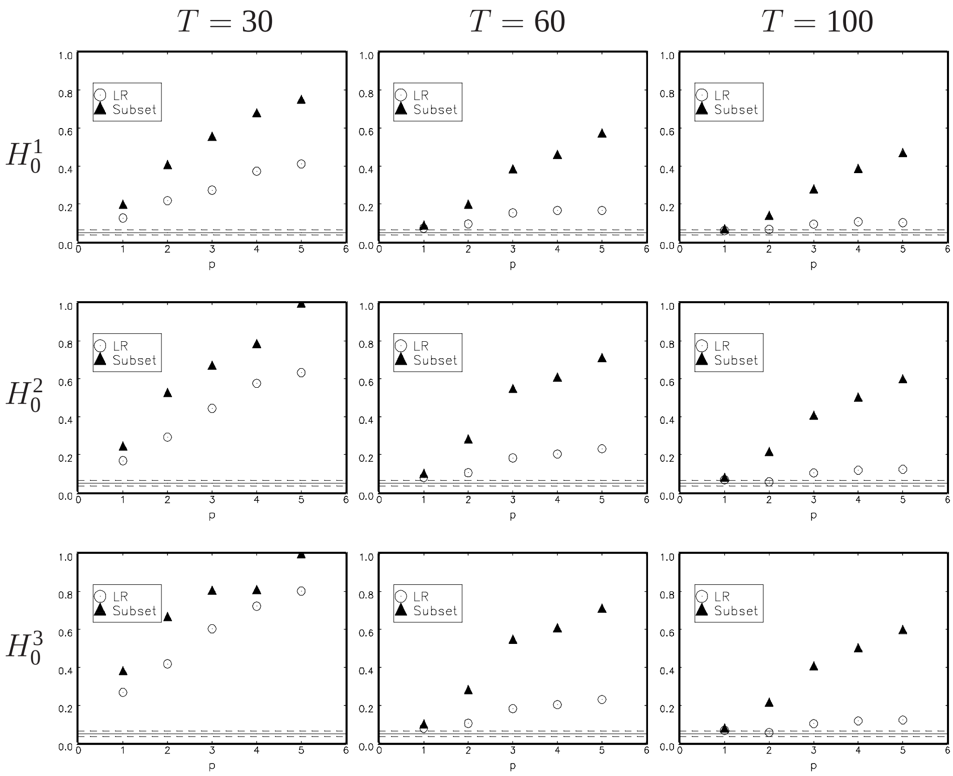

Figure 1: Size of LR- and Subset test DGPs (a)-(e), $T = 30, 60, 100$ (columns)

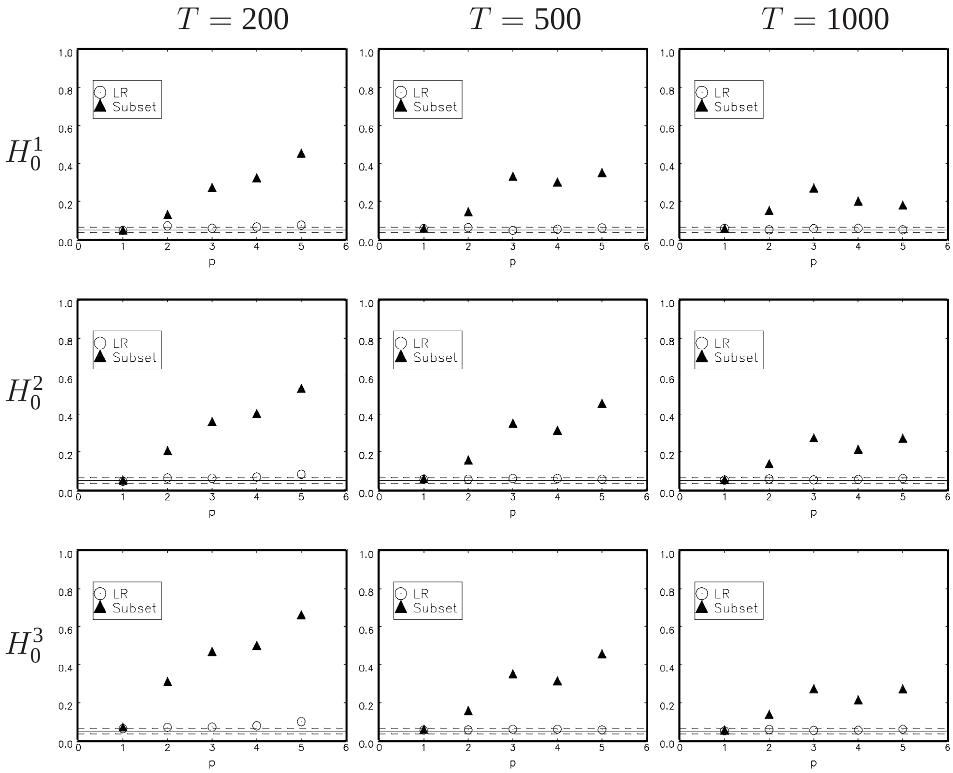

Figure 2: Size of LR- and Subset test DGPs (a)-(e), $T = 200, 500, 1000$ (columns)

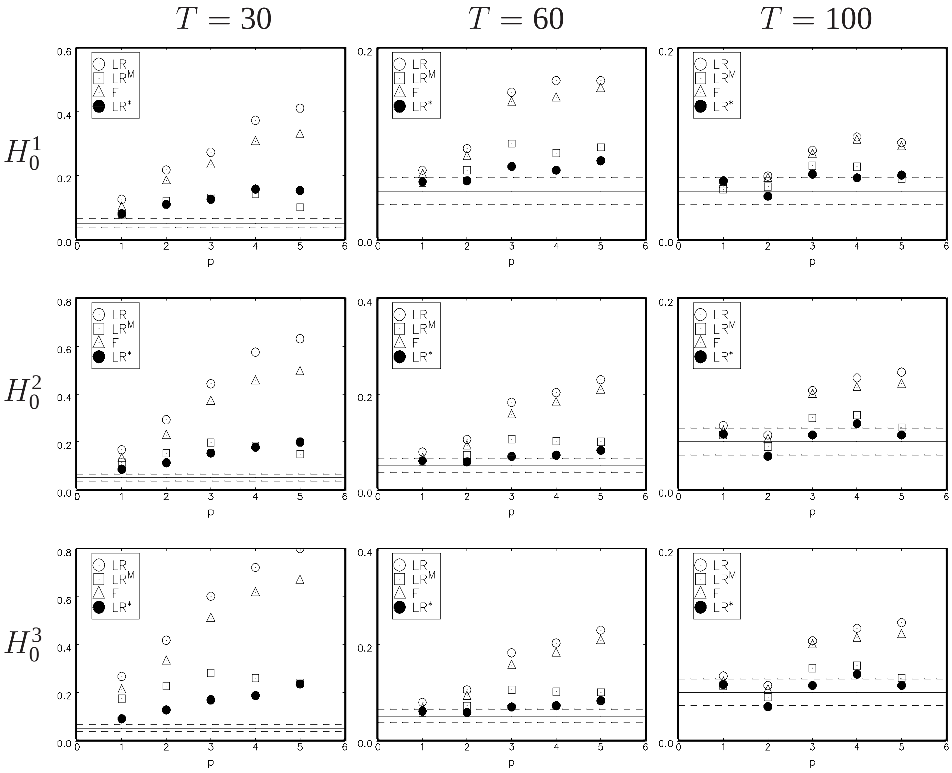

Figure 3: Size of LR, LR^M, F and LR^*, DGPs (a)-(e), $T = 30, 60, 100$ (columns)

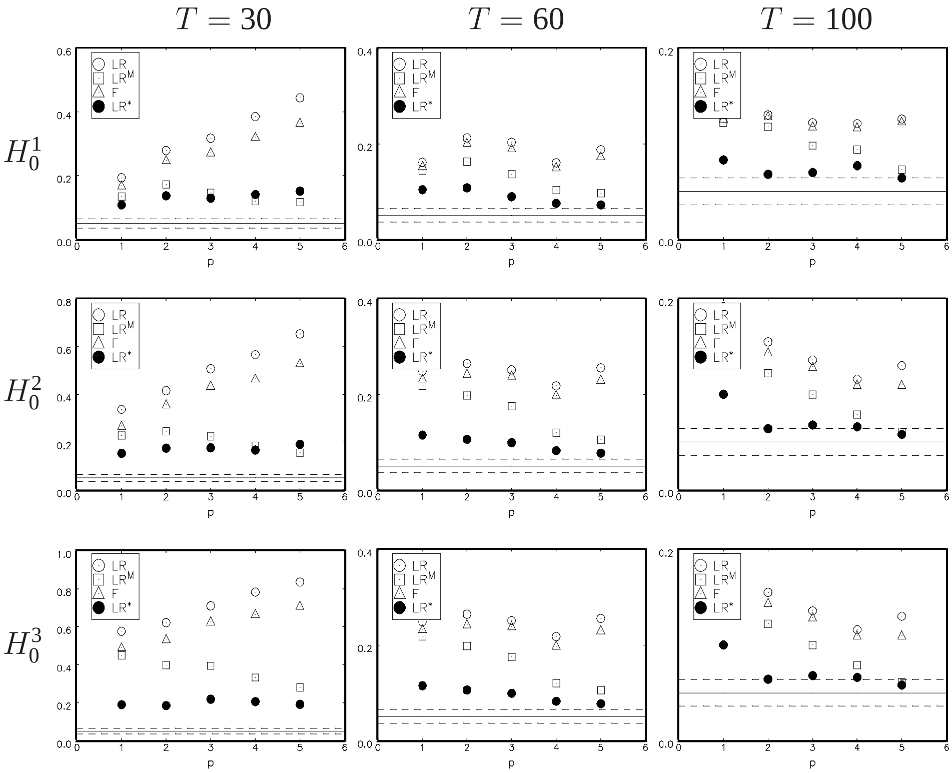

Figure 4: Size of LR, LR^M, F and LR^*, DGPs (f)-(k), $T = 30, 60, 100$ (columns)

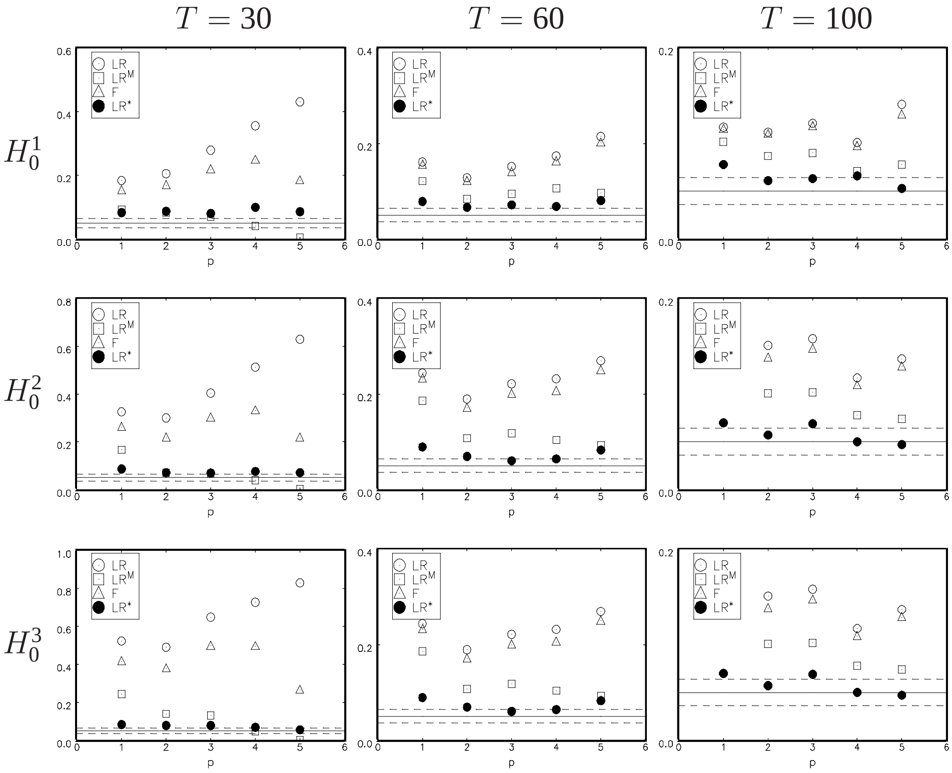

Figure 5: Size of LR, LR^M, F and LR^*, DGPs (l)-(p), $T = 30, 60, 100$ (columns)

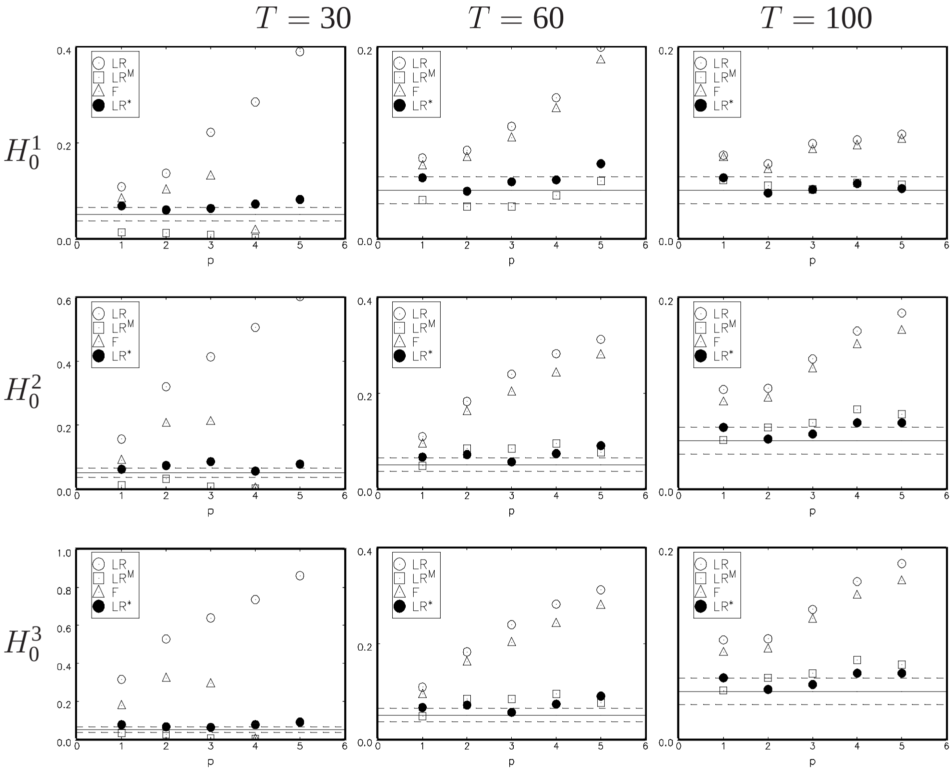

Figure 6: Size of LR, LR^M, F and LR^*, DGPs (q)-(u), $T = 30, 60, 100$ (columns)

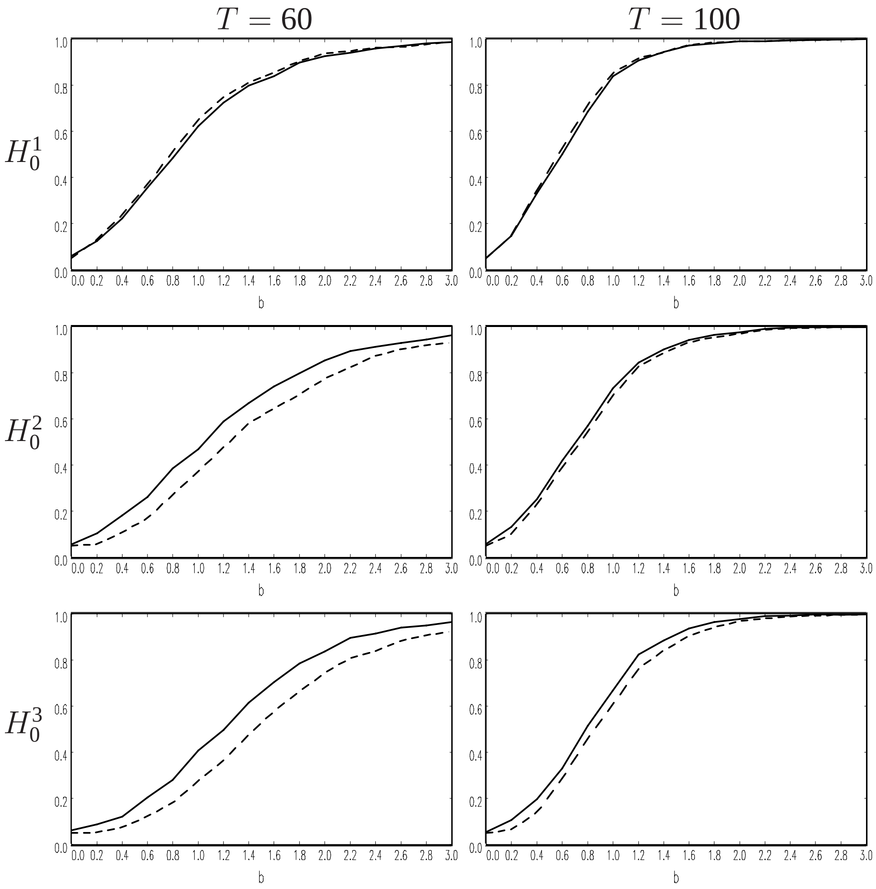

Figure 7: Power estimates of LR^* (———) and LR (size-adjusted) (- - - - -). Results are based on DGP (s) with $\alpha_{21} = b/\sqrt{T}$