ANNALS OF ECONOMICS AND FINANCE **19-1**, 151–196 (2018)

A Critical Appraisal of Studies Analyzing Co-movement of International Stock Markets

Jan F. Kiviet and Zhenxi Chen*

Literature is reviewed on the analysis of co-movement between the price indices of stocks or their realized returns at various markets. Four major categories of frequently recurring methodological shortcomings are registered. These are: (i) omitted regressor problems, (ii) neglecting to verify agreement of estimation outcomes with adopted model assumptions, (iii) employing particular statistical tests in inappropriate situations and, occasionally, (iv) lack of identification. The devastating effects of the detected methodological defects are explained in mildly technical appendices and are also illustrated by simulations and empirical examples.

Key Words: Diagnostic tests; Interlinkage of stock markets; Maintained hypotheses; Model specification; Test methodology.

JEL Classification Numbers: C18, C52, C58, G15.

1. INTRODUCTION

The last decades witnessed impressive economic globalization and financial integration. Interactions among the financial markets, such as co-movement, became obvious phenomena. As a consequence, many researchers examined the developments in stock markets for various countries and their interlinkages with markets in their neighboring countries as well as with the major international financial centres. In these studies widely divergent research methods have been used to analyze and characterize

*Kiviet: Amsterdam School of Economics University of Amsterdam, The Netherlands. E-mail: j.f.kiviet@uva.nl; Chen: Corresponding author. School of Economics and Commerce; Research Center of Financial Engineering, South China University of Technology, Guangdong, China. E-mail: zchen2@e.ntu.edu.sg. This paper was written while Professor Kiviet enjoyed hospitality as Visiting Professor at the Division of Economics, School of Humanities and Social Sciences, Nanyang Technological University, 14 Nanyang Drive, Singapore 637332. The author Zhenxi Chen would like to acknowledge the funding support from The Youth Foundation of the Humanities and Social Sciences Research of the Ministry of Education of China [grant number 17YJC790016] and the Natural Science Foundation of Guangdong Province, China [grant number 2017A030310314].

1529-7373/2018

these relationships and, not-surprisingly, often have arrived at conflicting conclusions. In this paper we will not focus in the first place on the different empirical findings reported by all these earlier studies, but on the various methodologies that they have used. After all, only inferences obtained from sound research methods should deserve serious further attention.

In the vast literature on the analysis of relationships between financial markets many studies focus on bivariate relationships, just involving two markets; others include more than two in a multivariate approach. Irrespective of the number of markets taken into consideration, the literature can be divided roughly into three different approaches regarding the chosen market characteristic of major interest. This is either: (*i*) the price index p_{it} of stock market i at time period t or its natural logarithm $\log(p_{it})$; or (*ii*) the conditional expectation $E(r_{it} \mid \mathcal{I}_t)$, where $r_{it} = \log(p_{it}) - \log(p_{i,t-1})$ is return and $\mathcal{I}_t$ represents an information set usually containing many (mostly lagged stock market) variables; or (*iii*) an aspect of the distribution of return other than its conditional mean, such as for instance the conditional centered second moment $Var(r_{it} \mid \mathcal{I}_t)$, or possibly particular quantiles of the density of r_{it}. An additional issue is usually whether the stock market prices should be taken in terms of local currency or be expressed in, for instance, US dollars. Many studies explore both variants and often find little difference between these two options.

For a recent historical survey on approaches to model individual stock returns see Koundouri et al. (2016). Studies that aim to explain movements in the actual price variables of different markets often employ cointegration analysis after nonstationarity of the stock indices has been assessed. Studies that chose to focus on modeling return (sometimes because no cointegration between indices could be established) usually specify some type of model for its conditional expectation and variance in terms of a dynamic relationship involving the returns at various markets. Studies under (*ii*) often use a vector autoregressive (VAR) model and assume conditional homoskedasticity, whereas the studies under (*iii*) tend to use a very simple model for the conditional expectation and then employ many parameters for modeling the conditional heteroskedasticity. The latter approach may be induced by strong belief in the efficient market hypothesis, which suggests that the conditional expectation of return is constant. However, the approach under (*ii*) as a rule provides empirical evidence of its significant nonconstancy.

The majority of studies analyze daily closing values of the stock market indices. A common problem is that different markets are located in different time zones. Often the index of the preceding day is taken for the New York stock index when modeling its effects on Asian markets. However, as we shall clarify below, this can have a peculiar effect on Granger-causality tests. Weekend days are excluded from the data set and usually the index

of the preceding day is used for the missing data on bank holidays, which are not always the same in all countries. Fewer studies on stock market interlinkage analyze weekly or monthly data.

From this literature a seriously disturbing observation has to be made regarding the progression of science. A major characteristic of this literature is that its contributors usually generously cite empirical findings obtained in earlier studies, irrespective of whether these have been obtained by a conflicting model specification or methodology. However, any discussion is then usually just about differences in the outcomes, instead of focussing on evaluating the pro's and con's of the chosen model specifications and the employed statistical techniques. In the review below, we will primarily highlight — and criticize where we think this is justified — the chosen methodology of a great (but by far not exhaustive) number of studies published in peer reviewed journals over the last two decades. In our opinion, the resulting empirical findings merit serious substantive consideration only, if the underlying methodology has no obvious flaws. Few studies seem to belong to the latter category.

In Section 2 we review many earlier studies on linkages between emerging and more developed stock markets, grouped in a series of subsections according to the three approaches distinguished above, and comment on their chosen methodology. There we note four major categories of problems which undermine many published empirical findings. All of them as a rule lead to seriously affected inferences, such as biased estimates of reaction coefficients and glaring misrepresentation of the actual significance level of test outcomes and of their reported p-values. These four categories of methodological errors are: (a) omitted regressor problems or not controlling adequately for relevant covariates; (b) neglecting to verify harmony of obtained estimation outcomes and the adopted model assumptions on which inference has been built; (c) employing particular statistical test procedures in inappropriate situations because the maintained hypotheses (a notion to be clarified below) do not hold; and, occasionally, (d) identification problems which preclude any useful statistical analysis. Further explanations of these rather technical issues and some relevant derivations are presented in their most simple formal format in a series of appendices. In Section 3 we provide some illustrations, both empirical and synthetic (by simulation), to highlight the devastating consequences of committing any of the four mentioned categories of methodological sins. Finally, some conclusions are drawn in Section 4.

2. REVIEW OF EARLIER RESULTS

Here we discuss a great number of published studies with relevance for the relationships between stock market indices or returns. We will catego-

rize all these studies with respect to their bivariate or multivariate nature, and regarding the methodology used for analyzing the chosen characterization of the dependent variable. In particular we will verify how careful all these studies are regarding four rather fundamental aspects for any empirical statistical analysis. These four (not fully disjunct) aspects are: (a) has it formally been examined –especially in case of a bivariate analysis– whether the adopted approach may suffer from omitted variables bias, because an appropriate specification may require a more extended multivariate approach, see also Appendix A.1; (b) does the study use diagnostics in order to verify whether or not the statistical assumptions made are actually supported by the empirical findings, see also Appendix A.2; (c) are the statistical tests used suitable for the particular situation in which they are applied (see Appendix A.3 for an example of an often misused test), and (d) are the estimated parameters formally identified (see Appendix A.4).

We group the cited articles in separate subsections which collect studies that focus on the level of indices, on modeling the expectation of return, and on other aspects of the distribution of return respectively. In each subsection these reviews are presented in chronological order of the examined publications.

2.1. Studies with focus on the level of stock indices

Chung and Liu (1994) analyze weekly stock market data expressed in local currencies from early 1985 until May 1992 for the US, Japan, Taiwan, Hong Kong, Singapore and South Korea. Nowhere in their analysis the Black Monday stock market crash of 1987 plays a role. They do not allow for any (possibly temporary) structural changes in any of their models. They find all series to be nonstationary (after taking their natural logarithm). Next they apply the Johansen methodology to establish the number of stochastic trends and cointegration relationships. The numbers they find prove to be highly dependent on the chosen lag length in the VEC (vector error-correction) model. Their maximum likelihood based inferences require normality of the disturbances. However, normality is rejected for all individual countries. Though, by using alternative multivariate normality tests (which involve a much larger number of degrees of freedom) the meshes of the net are stretched so far that establishing misspecification escapes. Using high degree of freedom Ljung-Box tests the authors declare all their estimated equations "free of serial correlation problems", which disregards the poor power of such tests. Next a criterion based on the average absolute forecast errors over the last year (taken out of the estimation sample) is employed to select the model which uses 12 lags, instead of 24 or 36. Finally, this model (which imposes without testing constant coefficients and constant error variances over the whole sample)

is further used to characterize and to test various hypotheses on the long and short run dynamics of the six analyzed stock market indices.

Arshanapalli et al. (1995) use more classic (Engle-Granger) cointegration and error-correction analysis to examine the stock market linkages between East-Asian markets (excluding China) and the US. Based on daily data ranging from January 1986 to May 1992, and analyzing separately the data pre and post October 1987, they conclude that the cointegrating structure of these markets and their dependence on the US increased after October 1987. However, their multivariate error-correction regressions actually result from reformulated and restricted VAR models with maximum lag order 4 for the 7 considered stock market indices, in which one linear cointegration relationship has directly been imposed. This cointegration relationship has been obtained from a static constant parameter regression between all indices, in full trust that its putative super-consistency will repel any bias due to neglected simultaneity and omitted short-run dynamic adjustments. This imposition undermines the interpretability of the presented inflated t-statistics. Apparently not noticing that no alternatives were allowed in their analysis, the authors nevertheless infer that "the six major Asian stock exchanges are linked together with a long-run equilibrium relationship". First, however, by diagnostic testing (see Appendix A.2) the adequacy of the dynamic specification of the error-correction regressions should have been verified, especially by tests for omitted regressors (see Appendix A.1). Also LM (Lagrange multiplier) tests for (higher-order) serial correlation should have been used, and not just the DW statistic, which requires all regressors to be strictly exogenous and hence is inappropriate in error-correction models. Presentation of results for the static long-run regressions over shorter subperiods might have given genuine support to the claimed existence of different though constant parameter long-run relationships before and after October 1987.

Ghosh et al. (1999) present a classic (Engle-Granger) cointegration analysis too. They restrict themselves to separate bivariate models for nine South-East Asian developing stock market indices. In all these bivariate models the other variable is the developed stock market of either the US or Japan. They use 141 daily data for estimation and 60 for out of sample prediction. Estimation is just based on the end of March until end of October 1997 period. None of the model formulations used does in any way explicitly examine whether the parametrization should pay respect in this way or another to the fact that in July 1997 the Asia financial crisis broke out in Thailand. Hence, all models impose many untested restrictions. Moreover, in the methodology practiced in this paper rejection of stationarity of the errors of a static bivariate cointegration relationship is automatically interpreted as absence of cointegration, whereas it is quite possible that cointegration would be established after including other more or less devel-

oped markets as well. The obtained error correction model estimates are all interpreted as if they represent the true data generating process, although no diagnostic test outcomes that would support this have been presented. The authors even claim that their findings imply the existence of causality; apparently they wrongly take this to be equivalent to correlation. Finally, they show that out of sample predictions by their models are superior to naive (no change) forecasts, which does not surprise. They would better have compared their predictions with the actually observed realizations in such a way, that by a Chow-type predictive failure test a genuine diagnostic on the quality of their findings had been obtained.

Huang et al. (2000) use daily price data from 1992 to 1997 for the US and various Asian stock markets. Allowing for structural breaks, they reject (except for Shanghai and Shenzhen) bivariate cointegration between the data for log of price. Next they examine Granger causality by bivariate dynamic regressions in the return data (though, surprisingly, not allowing for any structural breaks here). From bivariate VAR(k) models, where the Schwartz criterion is used to select lag order $k = 1$, they claim to find unidirectional Granger causality from the US to Hong Kong and no Granger causality in either direction between Hong Kong and Shanghai, nor between the US and Shanghai. The limitations of bivariate Granger tests, where in case of omitted regressors the included regressors serve as proxies for the omitted regressors leading to estimation bias (see Appendix A.1), are not mentioned. Nor has any evidence been presented on the statistical adequacy of the dynamic regressions in the form of diagnostics on serial correlation, heteroskedasticity, structural breaks or omitted (lagged) regressors.

Cheng and Glascock (2005) analyze weekly stock market price indices denominated in US dollars for China, Hong Kong, Taiwan, Japan and the US, covering January 1993 to August 2004. First, they demonstrate for the three markets from the Greater China Economic Area (GCEA) that each is not weakly efficient, because predictions by simple random walk models are dominated by those obtained from univariate models using more information from the past. Next, various single equation Johansen-type cointegration tests follow, but excluding the case where both Japan and US could appear with all the GCEA states in one cointegration relationship. Cointegration is not found, neither when tests are conducted over two (not specified) consecutive subperiods. Of course, it would have made sense here to split the two subperiods such that they deliberately exclude the weeks strongly associated with the Asian financial crisis of 1997/1998 in order to avoid contagion effects. Next, models are estimated by linear least-squares which are simple transformations of first-order bivariate autoregressive distributed lag equations. Overlooking the dangers of omitted relevant other markets or of longer lags, and not realizing that validity

of the earlier inferences would imply nonstationarity of the error terms, no single diagnostic is presented. Nevertheless, from standard t-tests far fetched conclusions are being drawn regarding existing nonlinear pairwise relationships. Finally, an innovation accounting analysis follows, now based on estimating a multivariate VAR in the returns. It has not been reported how lag lengths were assessed, nor whether diagnostics produced any evidence supporting the chosen specification.

Yang et al. (2006) study the interdependence among the stock markets of US, Germany and four major Eastern European emerging countries. Based on daily data from early 1995 to mid 2002 they first apply constant parameter cointegration analysis to the price series measured in logs of the market indices. This is done separately to three pre-crisis and three post-crisis years, although later, by using a two-year rolling window, strong evidence is presented that even in noncrisis years both the long-run and the short-run reactions vary over time. Nevertheless, the authors claim that the established cointegration relationships, where the lag length in the VAR systems have been selected on the basis of the Akaike information criterion, do pass LM tests for first- and fourth-order autocorrelation. The empirical conclusions are mainly based on the so-called persistence profile technique and on generalized forecast error variance decompositions. These numerical measures have been obtained from models which impose doubtful parameter constancy, and are next interpreted just on the basis of their point estimates, without taking their (obviously hard to assess) standard errors into account.

Huyghebaert and Wang (2010) are well aware that different research methodologies may be at the base of the mixed results in many earlier studies. For seven East Asian stock markets and the US they use daily data from mid 1992 to mid 2003, distinguishing the mid-1998 until mid-1999 crisis year from the pre and post crisis periods, examining two variants for the latter. They choose for a multivariate VAR framework for the log of stock-exchange indices when performing cointegration tests, assuming constant parameters over the four distinguished periods. They expect — overenthusiatically— "to correctly specify" their various VAR models by using a likelihood-ratio test for lag length to "ensure that all dynamics in the data are being captured". Unlike Yang et al. (2006) for Eastern-Europe, they find for East-Asia cointegration only during the crisis period, but not before, nor some time after. Next, Granger-causality tests are performed based on a VEC specification in the log indices for the crisis period and on VAR specifications in the returns for the other periods. Apparently the authors do not realize that cointegration directly implies causality. In their causality analysis parameter constancy over the subperiods is taken for granted, no results for serial correlation tests are being mentioned. As it seems, only F-test outcomes and their p-values for joint significance of all

lag coefficients for the separate countries are being mentioned, neglecting that just one significant single coefficient would imply Granger-causality too. Moreover, in line with the (incorrect) interpretation of insignificant outcomes of Sims' likelihood ratio test as indicating truth of the null hypothesis, F-test outcomes with p-values just above 5% are now interpreted as establishing absence of Granger-causality. Whether or not the presented tests are robust for (untested) heteroskedasticity is not clear. This is all followed by a meticulous generalized impulse response analysis, which of course is built on –and is thus conditional on the not exhaustively tested validity of– the earlier established specifications. The short-term causal relationships inferred from them in terms of responses to one unit shocks are of course all random estimates, but presented without their standard errors these can barely be interpreted.

Burdekin and Siklos (2012) use daily data from early 1995 to mid 2010 to investigate the interaction between the Chinese, US and Asia-Pacific equity markets. Advocating an eclectic approach, they in fact produce a wide range of results which are mutually contradictory and all untenable from a statistical point of view. Seemingly meant to just present some initial simple stylized facts of the returns, such as estimates of pairwise unconditional correlations and of first-order autoregression coefficients, next significance tests of these are presented too. However, proper interpretation of the tests on correlations (see Appendix A.3) require assumptions which have to be rejected given the different coefficients obtained in the autoregressions. But also the latter are built on assumptions, such as no higher-order dependence and parameter constancy, which have not been verified. Next contagion tests are being produced in a set of fully symmetric equations [their (1.1)] which omit any dynamics but explicitly include endogenous regressors without imposing any restrictions which guarantee identification of the parameters of this simultaneous system. Although it is not mentioned which technique has been used to estimate the coefficients and their standard errors, due to the shortcomings just mentioned it should be obvious that no sensible interpretation of these results is possible either (see Appendix A.4). This is followed by estimating a dynamic conditional correlation (DCC) model. However, how the conditional expectation of the returns has been modeled this time, and whether this appeared satisfactory, is not mentioned, so probably again has not been verified. All attention goes here to the conditional covariance of the returns without any concern how this might be affected when the conditional expectation has not been modeled adequately. Finally static quantile cointegration regressions are estimated and their coefficients are interpreted on the basis of standard t-ratio's, which is inappropriate again.

Glick and Hutchison (2013) investigate China's financial linkages with Asia in both bond and equity markets using daily data from mid 2005 to

end of 2012. After some standard unit root analysis just differenced data are being considered, starting off with wrongly (see Appendix A.3) employing significance tests on simple correlation coefficients of returns. Next, a series of static bivariate, multivariate and pooled regressions are presented which expose extreme naivety regarding the properties of least-squares estimators and requirements to usefully interpret significance tests. Assuming for the moment that the simple multivariate regressions in their Table 2 are not misspecified, each significant coefficient rejects validity of a restricting bivariate regression, in which the presented significance test results are thus inappropriate, even if the regressors are hardly correlated because they are both serially dependent, given the causality tests presented in their Table A2. Moreover, the multivariate regressions also reject validity of the pooled regression, because the slope coefficients are clearly different for most of the individual Asian countries. However, the multivariate regressions are uninterpretable themselves too, because it seems unlikely that equity return in China is exogenous for all other Asian stock markets, so least-squares estimators are biased, not only due to possibly neglected reverse causality, but also due to omitted regressors because of neglected dynamics. Next, results are presented which undermine all findings presented before, because evidence is produced indicating that regression coefficients are non-constant through time. However, that evidence is just derived from pooled regressions, which involve restrictions on the slopes which had already been shown to be incredible. Hence, the many alternative regressions presented in this study simply illustrate that none of them fully respects the actual interdependencies between the variables examined. None of the empirical conclusions drawn in this paper is based on solid statistical evidence.

Wang (2014) examines the interdependence and causality among six East-Asian stock markets and the US. This is mainly a repetition of the Huyghebaert and Wang (2010) study, now using more recent daily data, namely from mid 2005 to mid 2013, thus focusing now on the effects of the 2008 global financial crisis instead of the 10 years earlier Asian financial crisis. Using the same methodology, the same criticism applies. In particular, without proper testing, parameter constancy is assumed within the three distinguished periods: pre-crisis, crisis and post-crisis. It seems that lag length tests have been performed rather mechanically, treating all East-Asian countries similarly, whereas possible misspecification of the VAR and VEC models has not been tested. The causality tests, for which it is not clear whether they are robust for heteroskedasticity, are joint on all lagged coefficients of each country, which obscures particular forms of causality.

2.2. Studies regarding stock return data

Some of the above mentioned studies, already turned to analyzing return after stock indices were found to be not cointegrated, at least over particular subperiods. Other studies turn to analyzing returns right away. Within this group of studies we can distinguish those that focus primarily on modelling the conditional expectation of the distribution of return and those that focus primarily on other aspects of this distribution.

2.2.1. Focus on conditional expectation

Groenewold et al. (2004) analyze the dynamic interrelationships between the share markets of Shanghai, Shenzhen, Hong Kong and Taiwan. Although correctly criticizing similar earlier research using just bivariate cointegration, they overlook that their own approach may suffer from the same criticism as they exclude the international global financial market from their analysis. They use daily data from end of 1992 until end of 2001 and also consider two sub-samples, namely before and after the crisis period 1997/98. No cointegration can be established, except for the two markets in mainland China (Shanghai and Shenzhen) over the first sub-sample. Next they combine the two markets in mainland China and estimate trivariate VAR(5) models which pass tests for 1st and 4th order serial correlation. Knowing from various other studies that at least the Hong Kong and Taiwan markets are affected by the New York stock exchange all VAR inference may be seriously hit by omitted variable bias, and so will then be the resulting impulse responses and forecast error-variance decompositions, which lack a measure of precision (standard errors) anyway.

Zhu et al. (2004) analyze causal linkages between returns at the stock markets of Shanghai, Shenzhen and Hong Kong, using daily data from early 1993 to the end of 2001. Not being able to find cointegration between the three indices, and not questioning whether this might be due to omissions such as the effect of global stock market indices and possible parameter nonconstancies, they move on to analyze causality between the three return series, with and without removing so-called calendar effects. They split the data period into two and estimate for each and the whole data period VAR models. Optimal lag lengths have been determined mechanically by the Schwartz criterion. No diagnostic checks have been used, so coefficient restriction tests may suffer seriously from adopting invalid maintained hypotheses. The causality tests are only based on joint F-tests, using the 5% significance level as a razor blade, whereas p-values of individual t-tests could have given useful supplementary evidence. A similar analysis is also

performed on the absolute values of the return series to analyze causality in volatilities.

Singh et al. (2010) analyze data for 15 stock markets, including US, Canada, UK, Germany, France and ten major Asian markets. They use daily return data over the 2000 through 2008 period obtained from indices both at opening and at closing of the market. First they estimate VAR(15) models using just one lag for both close-to-close returns and for open-to-open returns. No evidence on parameter constancy nor on lack of error serial correlation is provided. Next the two VAR models are adapted in the following way. For the two markets that open (close) earliest in the day (Japan and Korea) the VAR specification is unaltered (includes 15 lagged explanatories). The three markets that open (close) next are Singapore, Malaysia and Taiwan. For them in the VAR specification the lagged returns of Japan and Korea are replaced by current values. And for the remaining ten VAR equations all lagged returns at earlier opening (closing) markets are replaced by current returns, and in addition the second through fifth lag of the dependent variable has been added, but not for France and Germany. The authors do realize that this introduces reverse causality (simultaneity) in their specifications (except for Japan and Korea), at least when the returns of the earlier opening (closing) markets have been realized in part during overlapping opening hours, but they suggest (footnote 5) that this has only a minor effect. Instead of using an alternative and still consistent estimation methods they just use least-squares and wrongly interpret standard t-values in the standard way. For what reasons suddenly lag lengths larger than one have been included is not clarified. For none of these results the maintained hypotheses implicitly adopted given the estimation method used seem credible. Finally, chucking all earlier results for the conditional first moment of return, AR(1)-GARCH(1,1) models are presented, which do of course require a completely different set of maintained hypotheses, but these have not been tested either.

Jayasuriya (2011) examines interlinkages of stock returns for China and three of its emerging neighbors, namely Thailand, Indonesia and Philippines. Although citing various similar studies mentioning the significant role of the US, this study chooses to include only some control variables from the real economy, namely the growth rates in interest, inflation and exchange rates, which –without any further motivations– are all treated as exogenous. The analysis is based on the monthly stock returns from November 1993 to July 2008, to which Jarque-Bera tests have been employed, which is incorrect, because it is most unlikely that these returns are serially uncorrelated and have time-invariant first four moments, which

is a requirement for this test to apply. Because the returns are found to be stationary it is claimed that a VAR model is appropriate, overlooking that it should be examined as well whether a stationary error-correction term in the stock indices has to be included as well. All conclusions are based on VAR(1) models which have passed LM serial correlation tests at lags 12 and 24. Of course, order 1 and 2 tests should have been used as well, because if the problems are especially at those lags, the tests employed lack power. Parameter constancy is taken for granted, one simple dummy is supposed to account for the Asian financial crisis, and the omission of more developed and other neighboring stock markets is not supposed to lead to estimation bias. Hence, the presented impulse response functions and variance decompositions are clearly based on much too weak grounds.

Chow et al. (2011) analyze weekly returns and their absolute value (which they take as their measure of volatility) from 1992 until 2010 at the stock markets of Shanghai and New York. In their bivariate analyses they exclude, without testing, any possible effects from Hong Kong or any other East Asian markets, so no evidence has been produced to refute obvious criticism that further explanatory variables have wrongly been omitted undermining all their inferences. First some simple descriptions of the (absolute) returns at the two markets are presented in the form of means, variances and simple correlations, over the two decades separately, and also over the full 20 years period. These suggest nonconstancy, but later analyses (and earlier results in Chow and Lawler, 2003) indicate that these three data characteristics are in fact nonconstant within the two decades too, so apparently they just should be seen as rough indicators for which no precision measure can be given. Next (in their Table 3) two simultaneous and unidentified (see Appendix A.4) equations have been estimated (also over the two decades separately), and wrongly it has been inferred from two t statistics (which are algebraically equivalent, see Appendix A.6) that the New York market affects Shanghai and vice versa. Next they estimate time-varying coefficient models (not mentioning that if these are found to be appropriate all their earlier inferences have to be withdrawn) using Kalman filter techniques and making strong distributional assumptions on error terms and weekly random movements of the coefficient values. Plausibility of these assumptions has not been verified from the fitted models by the authors. Results demonstrating some of the serious shortcomings of these bivariate time-varying coefficient models can be found in Chen et al. (2015).

2.2.2. Focus on (partial) correlations

Aityan et al. (2010) is a relatively recent study in a long tradition (see further references in their paper) which is purely based on calculating simple bivariate correlation coefficients of the returns at two markets over subperiods. In this particular paper they use daily data and calculate correlations for separate years. The very serious shortcoming of this analysis is that interpretation of these correlations would only be possible if for both series the daily data are serially uncorrelated and have constant mean and variance. However, all studies discussed in the foregoing subsection demonstrate that these conditions are not satisfied. Hence, all statistical tests produced in this literature are invalid. The correlations presented do not establish estimates of a constant underlying population parameter. An indication of their accuracy by an appropriate standard error cannot be given. Proper analysis of these return series demonstrates that more sophisticated multivariate models with many more parameters are required in order to produce inferences that possibly make sense. Only for such more extended models their adopted maintained hypotheses are not easily refuted by empirical evidence.

Kenett et al. (2012) use partial correlations and thus extend the above criticized approach to a trivariate analysis. Also they use a much smaller window of a month and consider overlapping windows. However, our major criticism still applies. The approach does not start off from formulating a stochastic model for the interdependencies between the analyzed return series. Therefore, there is no maintained hypothesis for which tenability can be verified by testing it against alternatives. Also specific hypotheses regarding the obtained results cannot be tested statistically because the reliability of the estimates cannot be expressed in standard errors.

2.2.3. Focus on conditional variance and further aspects

Li (2007) presents a multivariate study based on daily return data from 2000 till mid 2005 for the S&P 500 and the stock markets of Shanghai, Shenzhen and Hong Kong. The mean of these four variables is simply modeled by a VAR(1) allowing its innovation errors to have a conditionally heteroskedastic covariance matrix. For the latter various forms have been tested, leading to the acceptance of an unrestricted four-variable asymmetric GARCH specification inspired by the BEKK parametrization. For $m = 4$ markets this has (next to four intercepts) $m^2 = 16$ coefficients regarding the one period lagged returns for modeling the mean, and $3m^2 = 48$ parameters to model the error variance. Tenability of the assumption that the errors are innovations is only supported by high-order (12 and 24)

Ljung-Box statistics. Their p-values (all above 0.15) inspire the author to conclude straightforwardly that the mean has been adequately specified. This conclusion seems rather rash, taking into account that usually no satisfactory parsimonious constant parameter specification can be found for the mean of returns. The Ljung-Box test is well known for being heavily undersized (much too low rejection probability) and having poor power. A more serious test for the chosen specification of the mean should examine the significance of higher-order lagged returns and possible structural changes in their values. Also the choice for daily data, given the diverging closing hours and disjunct bank holidays of the markets in the US and in China, may have curious effects on the obtained residuals. A poorly specified mean equation will lead to residuals which do not really represent true innovations and therefore may suggest findings regarding volatility linkage which will not be genuine but mere artefacts. Anyhow, on top of concluding to very mild unidirectional return linkages of mainland China with Hong Kong and of Hong Kong with the US, the study also claims to establish various rather subtile volatility linkages between the four markets. The latter are of course all conditional on the maintained hypothesis that the mean of the returns has been modeled adequately indeed by a constant parameter VAR(1). This supposition becomes more doubtful after the author has reduced the sample size in order to get rid of the 9/11 shocks. Re-estimation of the same model over the three years after 9/11 yields estimates for the mean of the returns which still support unidirectional dependence of Hong Kong on the US, but the weak dependence of mainland China on Hong Kong is no longer significant.

Zhang et al. (2009) analyze data for the Shanghai and Hong Kong composite stock indices over the period mid 1996 until late 2008. They start off with the bivariate equivalent of the specification used by Li (2007), oddly enough after they have already established the higher-order autoregressive nature of the return series for Hong Kong. Regarding return spillovers from Hong Kong to Shanghai and vice versa the same conclusions are drawn as in Li (2007), but not regarding volatility linkage, which they ascribe to the different data period. Although that may certainly be one of the reasons, it is strange that no remark has been made on the devastating effects of their choice for a bivariate analysis. This forces zero values for parameters that are significantly different from zero in Li's multivariate study. That Zhang et al. (2009) start off from a misspecified model for the mean of return is also communicated by the significant Ljung-Box statistics, which are presented in the tables, but receive no attention in the text at all. Nevertheless, Zhang et al. (2009) continue to use their limited specification of

the mean model in a more sophisticated copula-based analysis. Although evidence is put forward that for Shanghai the linear autoregressive model for the mean improves by allowing for a nonlinear dynamic extension, again clear signs are presented (though again neglected) indicating that the resulting residuals of this extended mean equation do not represent a series of innovations, which implies that the copula approach has been built on an inappropriate likelihood function.

Beirne et al. (2010) construct GDP weighted weekly returns for the global market (US, Japan and four major European countries) and for large regions (including Asia) and next perform trivariate analyses of global and regional spillovers in emerging stock markets by specifying VAR-GARCH(1,1)-in-mean processes with BEKK representation. They impose, without testing, that there is no spillover from local markets to regional and global markets, nor from regional markets to the global market. The only diagnostic used is the ten lags Ljung-Box test, which rejects for only a few countries. The study misses results for the much more critical test on significance of the second and higher order lag in the VAR specifications. The results for the Asian countries are based on data from mid 1993 until early 2008. No checks on parameter constancy have been reported.

Li (2012) examines the stock market linkages between China, Korea, Japan and US. Using daily data for return from mid 1992 to early 2010 he applies a 4x4 GARCH-BEKK model and conditional correlation to uncover the linkages. The results suggest that China is linked to the oversea markets. Here too the GARCH-BEKK model has a specification of just one lag in the mean equation, which is accepted by the author given Ljung-Box test results at 12 and 24 lags. Hence, the same criticism applies as to Beirne et al. (2010). Given the results in the papers focusing on the expectation of return, just using a VAR(1), especially if it concerns daily data, does not seem right. In that case, the residuals do not represent innovations, which would turn all GARCH-BEKK findings into mere artefacts.

Zhang and Li (2014) use daily data from 2000 until 2012 on the Shanghai stock exchange composite index and the Dow Jones industrial average index. Allowing for a structural break they cannot establish cointegration between the indices; their explanations for that do not include that any bivariate approach seems inappropriate. Next they focus on analyzing returns. First univariate models are used combining an ARMA(1,1) specification for the conditional expectation of return with GARCH(1,1) for the error terms, and next the two established standardized residual series are used in a trend analysis with structural breaks of the so-called conditional correlations, which they forget to define properly. Given the almost

complete lack of diagnostic testing of the many specification assumptions made we suppose that the results should be interpreted much more cautiously than the authors do. Finally, in a quantile regression model it is suddenly supposed that the relationship between the two returns requires in fact only 3 coefficients (between end of 2007 until early 2012), whereas the earlier analysis used at least 20 or more.

3. ILLUSTRATION OF CONSEQUENCES OF POOR ECONOMETRICS

In this section we demonstrate the consequences of various methodological flaws by some simple simulations and also by empirical illustrations.

3.1. Some simulation findings

To demonstrate pitfalls associated with using the Jarque-Bera (JB) test for normality on return data and of using Ljung-Box (LB) or Box-Pierce (BP) tests for serial correlation in VAR models we simulated data from a very simple bivariate VAR(1) model where the intercepts in both equations are zero. The first equation has a dummy explanatory variable d_t with coefficient β, and its disturbance $u_t^{(1)}$ may be ARMA(1,1) with parameters ρ and θ. The second equation is a simple AR(1) process. Further details on this system are that

$$y_t^{(1)} = 0.1y_{t-1}^{(1)} + 0.1y_{t-1}^{(2)} + \beta d_t + u_t^{(1)} \tag{1}$$

$$y_t^{(2)} = 0.2y_{t-1}^{(2)} + u_t^{(2)}, \tag{2}$$

for $t = 1, ..., n$, where

$$\begin{aligned} u_t^{(1)} &= \sigma_1[(1-\rho^2)/(1+2\rho\theta+\theta^2)]^{1/2}\xi_t \\ \xi_t &= \rho\xi_{t-1} + \varepsilon_t^{(1)} + \theta\varepsilon_{t-1}^{(1)} \\ u_t^{(2)} &= \sigma_2\varepsilon_t^{(2)}. \end{aligned}$$

We take $\sigma_1 = 0.01$ and $\sigma_2 = 0.05$, whereas the series $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$ are mutually independent serially uncorrelated drawings from the standard normal distribution. Hence, disturbance $u_t^{(2)}$ has variance σ_2^2 and the AR(1) series $y_t^{(2)}$ has standard deviation $0.051 = 0.05/\sqrt{1-0.2^2}$. Since series ξ_t is an ARMA(1,1) process with variance $(1+2\rho\theta+\theta^2)/(1-\rho^2)$, disturbance $u_t^{(1)}$ is an ARMA(1,1) process with variance σ_1^2. This yields series $y_t^{(1)}$ and $y_t^{(2)}$ which reasonably mimic some of the basic properties of actual return series.

We did choose $y_{-50}^{(1)} = y_{-50}^{(2)} = 0$ and discarded the obtained warming-up values for $t = -50, ..., -1$ in estimation. We generated this system 10000 times for just a few specific values of β, ρ, θ and sample size n. Each replication we applied particular tests at significance level $\alpha = 0.05$ and counted the frequency of rejection. The JB test for normality, which is χ_2^2 distributed under its null hypothesis, we applied to the two synthetic return series. Tests against p^{th} order serial correlation for $p \in \{1, 10\}$ we applied to the first equation when estimated by least-squares assuming (sometimes incorrectly) that the disturbances $u_t^{(1)}$ are serially uncorrelated. Next to the BP and LB tests, which are χ_p^2 distributed under their null, we examined the LM and the LMF tests, which under the null are χ_p^2 and $F_{p,n-k-p}$ distributed[1], where k is the number of regressors in the equation under the null hypothesis.

All these tests are asymptotic tests, so for n finite their actual type I error probability may deviate from 0.05. If it is much larger than 0.05 for a particular test then a rejection could both be due to validity and invalidity of the null hypothesis, which renders the test not very useful. If it is much smaller than 0.05 the probability to commit a type I error is much lower than intended, which results in much larger probability of type II errors and thus in low power of the test. So, in both cases (over-rejection and under-rejection under the null) the ability of the test to help choosing between the null and the alternative hypothesis is negatively affected.

We first focus on the popular Jarque-Bera normality test. This has been developed for series which are (asymptotically) IID (independent and identically distributed) like regression residuals when the model is adequately specified and the errors are IID themselves. Hence, when employed to a (demeaned) return series, the JB test can only fruitfully be interpreted if this return series has first and second unconditional moments that are time-invariant, and these returns are serially uncorrelated. However, these three characteristics are usually rejected empirically, implying that the maintained hypothesis of the JB test procedure does not hold. Hence, in such circumstances one should not employ the test. Nevertheless, we will use it to both variables $y^{(1)}$ and $y^{(2)}$ of the above system to examine its behavior. The dummy variable d_t has all its observations zero, except $d_{n/2} = 1$ (and we made sure n is even). Hence, the coefficient β models a one-off shock halfway the observed sample. We examined $\beta = 0$ and some negative val-

[1] We considered the versions which set pre-sample values of the least-squares residuals equal to zero.

ues, equal to 2, 4 and 8 times σ_1, which represent a one day stock market crisis.

Table 1 presents the rejection probability by JB when applied to series $y^{(1)}$ in case $\rho = \theta = 0$. We note that for $-0.02 \leq \beta \leq 0$ values remarkably close to the chosen significance level are found, but for $\beta \leq -0.04$ the null hypothesis is much more often rejected, apparently simply because the mean is nonconstant. From these results one could wrongly conclude that the serial dependence of the series seems not a crucial issue. However, this is due to the moderate values of the coefficients of the lagged regressors in the system. We also applied the JB test to the $y_t^{(2)}$ series and found a rejection probability slightly below 0.05. However, upon increasing the AR(1) coefficient from 0.2 to 0.9 the rejection probability increased to 0.25 for the smaller and to 0.45 for the larger sample size. Hence, this highlights that the JB test should not be applied to return series if evidence is available that these are not IID, but only to the residuals of a well specified model for returns.

TABLE 1.

Rejection probabilities of JB normality test for series $y^{(1)}$

	$n = 150$				$n = 500$			
	β				β			
	0.00	−0.02	−0.04	−0.08	0.00	−0.02	−0.04	−0.08
JB	0.041	0.057	0.371	0.999	0.043	0.054	0.269	0.997

When analyzing the serial correlation tests we removed the dummy variable when generating and estimating the first equation. The rejection probabilities that we calculated can be found in Table 2. Note that under the null hypothesis (when $\rho = 0$ and $\theta = 0$) both the BP and LB test show profound size problems, which are opposite in nature for $p = 1$ and $p = 10$. Only for very large samples and p substantial too the control over type I errors is appropriate. For small p we note that size control gets even significantly worse for these two tests when n gets larger. Apparently the BP and LB rejection probabilities can be non-monotonic in n, which is a very bad sign. On the other hand, at the sample sizes examined both LM and LMF show no serious size problems. For all tests the rejection probability increases when the null is not true, as it should. From the results for the Lagrange-Multiplier tests we note that when the serial correlation problem is of low order (as in the examined AR(1) and MA(1) cases) power of the test benefits when using a low order alternative. The often much higher

reject values of the BP and LB tests when the disturbances are AR(1) or MA(1) are meaningless, because the corresponding probabilities are also much higher than for LM and LMF when the null hypothesis is true.

TABLE 2.

Rejection probabilities of residual serial correlation tests in (1)

	$n = 150$			$n = 500$		
	$\rho = 0.0$ $\theta = 0.0$	$\rho = 0.2$ $\theta = 0.0$	$\rho = 0.0$ $\theta = 0.2$	$\rho = 0.0$ $\theta = 0.0$	$\rho = 0.2$ $\theta = 0.0$	$\rho = 0.0$ $\theta = 0.2$
BP1	0.250	0.678	0.730	0.351	0.945	0.972
BP10	0.001	0.039	0.050	0.042	0.647	0.746
LB1	0.253	0.682	0.733	0.352	0.945	0.972
LB10	0.001	0.044	0.055	0.042	0.650	0.749
LM1	0.051	0.190	0.232	0.054	0.497	0.622
LM10	0.042	0.074	0.087	0.050	0.210	0.269
LMF1	0.049	0.184	0.226	0.054	0.495	0.619
LMF10	0.042	0.075	0.088	0.050	0.210	0.270

From these results one cannot predict what the patterns will be for similar or different values of n in larger VAR systems with longer lags and different coefficient values, but they do make clear that in the cases investigated there is little to choose between the BP and LB tests and that both can have large and small rejection probability under both the null and the alternative hypothesis, which completely undermines their signpost function. The LM and LMF tests show very reasonable size control, but their results also expose that one should certainly not overestimate the actual power of serial correlation tests. Hence, for none of the tests examined, it will ever make much sense to boldly conclude — as many practitioners wrongly do – that there is no serial correlation problem when the test employed happens not to reject.

3.2. Some empirical illustrations

In this subsection we use empirical data to illustrate the occurrence and serious consequences of various possible methodological flaws when estimating VAR models and performing tests for unit roots, cointegration and for causality. We perform these analyses by using the EViews package, and calculate the following diagnostics: (i) the BP (now referring to Breusch-Pagan) heteroskedasticity test, which tests whether the disturbance variance seems functionally related to the regressor variables; (ii) the JB (Jarque-Bera) normality of the residuals test; and (iii) also three

variants of serial correlation tests, namely LB (Ljung-Box), LM (Lagrange multiplier χ^2 test), and LMF (Lagrange multiplier F-test). Each serial correlation test will be conducted at lag orders running from 1 up to 10.

The empirical data used are the daily stock indices of SH (Shanghai's SSE composite index), HK (Hong Kong's Hang Seng), US (S&P 500), SG (Singapore's STI) and KR (Korea's KOSPI index) over the period January 2003 till the end of 2006, all extracted from Yahoo finance. All these stock indices are expressed in local currencies. Deliberately the sample period (just over 1000 observations) is chosen such that financial crises periods are avoided, in order to keep away from extra hard model specification problems. To take into account the different time zones, for the US market the index of the preceding day has been taken. For all markets we use the preceding available observation for a bank holiday. Figure 1 plots the time series of the five indices. By visual check, we do not observe pronounced breaks except for Shanghai, for which there seems a structural break in the course of 2005.

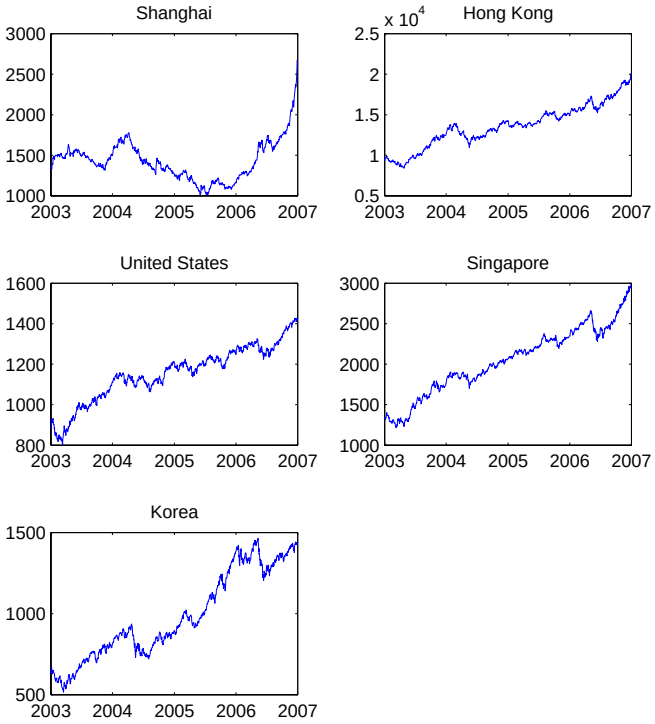
3.2.1. Univariate unit root analysis

First, we will illustrate inference problems which may occur in the context of standard unit root analysis. We have estimated univariate autoregressive models for the natural logarithm of all the five stock indices at lag length $l = 0, 1, 2...$ with an intercept plus (or without) linear trend, and performed (augmented) Dickey-Fuller tests over the full sample and over the first and last part of the sample.

In Table 3 we report for illustrative purposes the results for Singapore and in Table 4 for the US. We present the test statistic (DF) and its corresponding 5% critical value (DFC), and also the p-value for each of the five diagnostics, where serial correlation is tested both for order 2 and order 10. In addition, the minimum p-value of the LMF test is reported and also the lag (≤ 10) at which it was obtained. Note that the unit root test presumes that the autoregressive model is well specified, meaning that the error terms should be serially uncorrelated, have time-invariant variance and are normally distributed. The top panel of the tables include the linear trend and the lower panel excludes it. The full sample results are for 2003 through 2006, whereas sub-sample "first" is for 2003-2004 and "last" for 2005-2006.

For Singapore, both the information criteria of Akaike and of Schwartz (AIC and SIC) prefer lag length zero, irrespective of the inclusion of a

FIG. 1. Time-series of the five stock indices: Shanghai (SH), Hong Kong (HK), United States (US), Singapore (SG), Korea (KR)



trend. At the same time, however, we note that the disturbances seem heteroskedastic and nonnormal (this is also found without first taking logs of the index). When a less parsimonious dynamic model is estimated the BP and JB tests do not improve, but some evidence of higher order serial correlation is detected as well. Neglecting the diagnostics, none of the DF values of Table 3 forces to reject the unit root hypothesis. However, the results also illustrate that the positive statement "the Singapore index is a unit root process" does not find firm support from these findings, also because of the substantial differences between DF values over the subsamples.

From the results on the US presented in Table 4 we find again many rejections by the BP and JB tests. Irrespective of the inclusion of a linear

TABLE 3.

Dicky-Fuller unit root tests and p-values of some diagnostic tests in AR(l) models for the natural logarithm of the Singapore stock index

		intercept and linear deterministic trend included											
sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	−2.40	−3.41	0.00	0.00	0.61	0.60	0.60	0.28	0.28	0.28	0.13	6
	1	−2.44	−3.41	0.00	0.00	0.81	0.57	0.57	0.32	0.25	0.26	0.13	6
	2	−2.41	−3.41	0.00	0.00	1.00	0.71	0.71	0.34	0.16	0.17	0.07	6
	3	−2.49	−3.41	0.00	0.00	1.00	0.86	0.86	0.29	0.17	0.18	0.06	4
	4	−2.50	−3.41	0.00	0.00	1.00	0.35	0.36	0.33	0.01	0.01	0.01	8
first	2	−2.08	−3.42	0.00	0.00	1.00	0.53	0.53	0.76	0.59	0.60	0.27	1
	3	−2.13	−3.42	0.00	0.00	1.00	0.47	0.48	0.80	0.60	0.61	0.23	4
last	2	−1.20	−3.42	0.00	0.00	0.99	0.08	0.08	0.07	0.06	0.06	0.03	5
	3	−1.18	−3.42	0.00	0.00	0.99	0.06	0.06	0.07	0.06	0.06	0.01	3
		intercept included, no trend											
sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	−0.64	−2.86	0.00	0.00	0.60	0.60	0.60	0.28	0.28	0.28	0.15	6
	1	−0.63	−2.86	0.00	0.00	0.72	0.50	0.50	0.30	0.25	0.25	0.13	7
	2	−0.67	−2.86	0.00	0.00	1.00	0.63	0.63	0.34	0.17	0.17	0.08	7
	3	−0.75	−2.86	0.00	0.00	1.00	0.76	0.76	0.30	0.17	0.17	0.08	7
	4	−0.69	−2.86	0.00	0.00	1.00	0.45	0.46	0.32	0.01	0.01	0.00	7
first	2	−0.93	−2.87	0.00	0.00	1.00	0.57	0.58	0.74	0.58	0.59	0.31	1
	3	−1.02	−2.87	0.00	0.00	1.00	0.46	0.46	0.76	0.60	0.61	0.24	1
last	2	0.49	−2.87	0.00	0.00	0.99	0.11	0.11	0.09	0.08	0.08	0.04	8
	3	0.51	−2.87	0.00	0.00	0.99	0.05	0.05	0.10	0.08	0.08	0.02	3

trend the SIC criterion favors $l = 0$ whereas AIC prefers $l = 7$, whereas also the LMF tests reject low order autoregressive specifications. Again, for none of the rows in the table the unit root hypothesis is rejected, but at the same time no specification passes all the diagnostic tests. This table illustrates as well that as a rule the Ljung-Box test is much too mild, and that the two information criteria are of little use for finding an adequate model specification.

3.2.2. Cointegration analysis

Under the (somewhat doubtful) supposition that the natural logarithm of all five stock index series are genuine unit root processes, we next will illustrate inference problems which may easily occur in the context of cointegration analysis. Therefore we estimate VAR(l) models, with overall lag

TABLE 4.

Dicky-Fuller unit root tests and p-values of some diagnostic tests in AR(l) models for the natural logarithm of the US stock index

		intercept and linear deterministic trend included											
sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	−2.82	−3.41	0.00	0.00	0.02	0.01	0.01	0.04	0.03	0.03	0.01	7
	1	−2.61	−3.41	0.00	0.00	0.34	0.17	0.17	0.23	0.02	0.02	0.00	7
	2	−2.42	−3.41	0.00	0.00	0.95	0.06	0.06	0.44	0.13	0.13	0.04	7
	3	−2.50	−3.41	0.00	0.00	0.95	0.02	0.02	0.46	0.05	0.05	0.01	7
	7	−2.27	−3.41	0.00	0.00	1.00	0.93	0.93	1.00	0.49	0.50	0.41	9
first	2	−1.99	−3.42	0.00	0.00	0.94	0.07	0.07	0.67	0.33	0.34	0.07	2
	3	−2.08	−3.42	0.00	0.00	0.95	0.05	0.06	0.71	0.22	0.23	0.05	3
last	2	−2.80	−3.42	0.00	0.08	1.00	0.75	0.75	0.95	0.58	0.59	0.41	8
	3	−2.75	−3.42	0.00	0.08	1.00	0.75	0.75	0.95	0.52	0.53	0.24	5
		intercept included, no trend											
sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	−1.12	−2.86	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.00	7
	1	−1.06	−2.86	0.00	0.00	0.25	0.14	0.15	0.16	0.01	0.01	0.00	7
	2	−0.85	−2.86	0.00	0.00	0.96	0.10	0.10	0.35	0.10	0.10	0.03	7
	3	−0.91	−2.86	0.00	0.00	0.96	0.01	0.01	0.36	0.04	0.04	0.01	7
	7	−0.81	−2.86	0.00	0.00	1.00	0.86	0.87	1.00	0.68	0.69	0.62	9
first	2	−0.59	−2.87	0.00	0.00	0.94	0.08	0.08	0.64	0.28	0.29	0.08	1
	3	−0.67	−2.87	0.00	0.00	0.94	0.04	0.04	0.66	0.19	0.20	0.04	1
last	2	−0.21	−2.87	0.00	0.04	1.00	0.80	0.80	0.85	0.33	0.33	0.19	8
	3	−0.16	−2.87	0.00	0.00	1.00	0.79	0.79	0.86	0.27	0.28	0.15	8

length $l = 0, 1, 2, ...$, first for an initial number of market indices ($m = 2$) and next on an extended number ($m > 2$). For all tables we will again report which lag length is preferred by either the AIC or SIC criterion. The tables present the p-values of both Johansen's Maximum Eigenvalue test ($\lambda_{\max}$) and the Likelihood Ratio trace test (LR_{tr}) to assess the number of cointegration relationships (CRs). As in Tables 3 and 4 we also report the p-value for each of the five diagnostics, where serial correlation tests are again given for order 2 and order 10, but also the lag order for which the minimal p-value of LMF has been obtained. Now we use the multivariate versions of the diagnostics as provided by EViews. This means that the approximate critical values for the system LB test are taken to be χ^2 with $m^2(h - l)$ degrees of freedom, where m is the dimension of the VAR(l) system and h the order of serial correlation tested (here $h = 1, ..., 10$), so

the test is only feasible for $h > l$. The LM type tests test m^2h restrictions. As Eviews does not provide the LMF test, we follow the procedure of Edgerton and Shukur (1999) to implement the LMF test. Note that the Johansen Maximum-Likelihood approach presumes that the VAR model is well specified, meaning that the error terms are serially uncorrelated, have time-invariant variance and are normally distributed. Like all studies reviewed in Section 2.1 we will allow for an intercept in the VAR(l) models but not for a linear deterministic trend. We will again report results for the full and for the two sub-samples.

TABLE 5.

P-values of Johansen cointegration tests and diagnostics in VAR(l) models with intercept (without trend) for the natural logarithm of the stock indices of Hong Kong and Singapore.

							— order 2 —			— order 10 —			– Min –
l	CRs	$\lambda_{\max}$	LR_{tr}	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
	full sample												
0	0	0.17	0.04	0.00	0.00	0.14	0.20	0.03	0.49	0.36	1.00	0.02	1
	≤ 1	0.07	0.07										
1	0	0.15	0.06	0.00	0.00	0.53	0.18	1.00	0.81	0.41	1.00	0.34	1
	≤ 1	0.13	0.13										
2	0	0.13	0.04	0.00	0.00	NA	0.92	1.00	0.87	0.36	1.00	1.00	1
	≤ 1	0.09	0.09										
8	0	0.36	0.12	0.00	0.00	NA	0.18	0.63	0.39	0.24	0.83	0.02	6
	≤ 1	0.09	0.09										
	first sub-sample												
8	0	0.84	0.73	0.00	0.00	NA	0.30	0.47	0.17	0.03	0.85	0.03	7
	≤ 1	0.48	0.48										
	last sub-sample												
8	0	0.16	0.09	0.00	0.00	NA	0.59	1.00	0.92	0.92	1.00	1.00	1
	≤ 1	0.20	0.20										

Table 5 presents results for bivariate VAR(l) models ($m = 2$) for the natural logarithm of the indices of Hong Kong and Singapore. BP and JB reject all examined VAR(l) models, suggesting heteroskedasticity and non-normality. AIC and SIC suggest the appropriate lag order to be 2 and 0 respectively for the full sample. The LR_{tr} test perceives one cointegration relationship over the whole sample at $l = 0$. However, this result should not be trusted due to the serial correlation. Evaluating at $l = 2$ as suggested by AIC, there is no longer serious serial correlation, and now the LR_{tr} test also indicates one cointegration relationship. However, at $l = 8$ no cointegration

relationship is found in both full sample and sub-samples, and both full sample and first sub-sample suffer from serious serial correlation according to the LMF test.

Table 6 presents cointegration test results for trivariate VAR(l) models ($m = 3$) by adding Korea to the two markets system consisting already of Hong Kong and Singapore. BP and JB are again significant at all lag orders. Here both AIC and BIC favor the lag order to be 0, at which no cointegration relationship is detected. Concern for serial correlation is alleviated for the full sample as all p-values are above the 10% level, which is not the case for the first sub-sample at $l = 8$. Note that no cointegration is found, which of course undermines the results at lag order 2 in Table 5.

TABLE 6.

P-values of Johansen cointegration tests and diagnostics in VAR(l) models with intercept (without trend) for the natural logarithm of the stock indices of Hong Kong, Singapore and Korea.

						— order 2 —			— order 10 —			– Min –	
l	CRs	$\lambda_{\max}$	LR_{tr}	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
	full sample												
0	0	0.54	0.27	0.00	0.00	0.48	0.45	0.73	0.70	0.72	1.00	0.12	1
	≤ 1	0.33	0.28										
	≤ 2	0.39	0.39										
1	0	0.46	0.32	0.00	0.00	0.91	0.42	1.00	0.86	0.68	1.00	1.00	1
	≤ 1	0.53	0.42										
	≤ 2	0.38	0.38										
2	0	0.39	0.22	0.00	0.00	NA	0.78	1.00	0.87	0.64	1.00	1.00	5
	≤ 1	0.41	0.31										
	≤ 2	0.36	0.36										
8	0	0.84	0.58	0.00	0.00	NA	0.28	0.46	0.54	0.57	1.00	0.42	1
	≤ 1	0.45	0.42										
	≤ 2	0.50	0.50										
	first sub-sample												
8	0	0.91	0.87	0.00	0.00	NA	0.12	0.42	0.09	0.19	1.00	0.00	1
	≤ 1	0.88	0.78										
	≤ 2	0.50	0.50										
	last sub-sample												
8	0	0.48	0.36	0.00	0.00	NA	0.42	1.00	0.92	0.97	1.00	0.95	1
	≤ 1	0.67	0.46										
	≤ 2	0.29	0.29										

In Table 7 we add Shanghai and the US to the system. As before BP and JB are significant at all lag orders. Now AIC and SIC suggest orders

of 2 and 0 again. For these lag orders the LR_{tr} test indicates no and one cointegration relationship respectively. Note that at these two lag orders there are serious serial correlation problems according to the LM and LMF tests, thus these cointegration results should not be trusted. The serial correlation results also indicate clearly that the quality of inference regarding the appropriate lag order by the information criteria is most doubtful. Increasing the lag order, still no cointegration relationship is found, whereas at lag order 8 serial correlation problems seem no longer present. Investigation of the two sub-samples at lag order 8 indicates zero and two cointegration relationships, although the first sub-sample again suffers from serious serial correlation. Hence, we still did not establish genuine long-run cointegration.

At this stage we want to draw the following conclusions from the above. Since information criteria take it for granted that a VAR(l) model specification is adequate for the markets considered, they clearly should not be trusted to determine the lag order for performing cointegration tests when at the same time the specification does not pass diagnostic tests. However, no battery of diagnostics can ever guarantee that a particular specification sufficiently respects the actual data generating process. From the cointegration result of Table 7 using $l = 8$ and the last sub-sample one may conclude that if any cointegration relationship between these four Asian markets exists, this relationship most probably must also involve the US. So note that this implies that all inferences in Tables 5 and 6, even those with satisfactory diagnostics, are misleading. Apparently, none of its test results can be associated with any degree of credible significance. This is evidenced in Table 8 which lists the two established normalized cointegration relationships in the last sub-sample for the five-markets model at $l = 8$. The US is significant for the two cointegration relationships. Hence, note that the results on the five markets destroy the trustworthiness of the results on fewer markets. Probably the same would happen with the five markets model if we added further markets.

An issue of further concern for VAR modeling, not mentioned in any of the studies referred to above, is an uncomfortable consequence of the time zone difference between US and Asia. We can illustrate this for a simple binary VAR(1) system $y_t = Ay_{t-1} + \varepsilon_t$, where, for instance, $y_t = (s_{us,t-1}, s_{sg,t})'$ contains the log stock indices of the US and Singapore. This entails the equations

$$\begin{cases} s_{us,t-1} = a_{11}s_{us,t-2} + a_{12}s_{sg,t-1} + \varepsilon_{1t}, \\ s_{sg,t} = a_{21}s_{us,t-2} + a_{22}s_{sg,t-1} + \varepsilon_{2t}. \end{cases}$$

TABLE 7.

P-values of Johansen cointegration tests and diagnostics in VAR(l) models with intercept (without trend) for the natural logarithm of the stock indices of Hong Kong, Singapore, Korea, Shanghai and the US

						— order 2 —			— order 10 —			– Min –
l	CRs	$\lambda_{\max}$	LR_{tr}	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF lag
	full sample											
0	0	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.47	0.00	1
	≤ 1	0.24	0.12										
	≤ 2	0.60	0.31										
	≤ 3	0.39	0.29										
	≤ 4	0.34	0.34										
2	0	0.21	0.07	0.00	0.00	NA	0.02	1.00	0.24	0.70	1.00	0.59	7
	≤ 1	0.23	0.23										
	≤ 2	0.75	0.53										
	≤ 3	0.57	0.46										
	≤ 4	0.40	0.40										
8	0	0.29	0.13	0.00	0.00	NA	0.27	1.00	0.73	0.75	1.00	0.59	1
	≤ 1	0.32	0.30										
	≤ 2	0.80	0.57										
	≤ 3	0.65	0.45										
	≤ 4	0.31	0.31										
	first sub-sample												
8	0	0.79	0.64	0.00	0.00	NA	0.00	0.59	0.05	0.46	1.00	0.00	1
	≤ 1	0.83	0.69										
	≤ 2	0.62	0.67										
	≤ 3	0.85	0.78										
	≤ 4	0.57	0.57										
	last sub-sample												
8	0	0.00	0.00	0.00	0.00	NA	0.78	1.00	0.72	0.54	1.00	1.00	1
	≤ 1	0.00	0.02										
	≤ 2	0.70	0.65										
	≤ 3	0.67	0.67										
	≤ 4	0.64	0.64										

Due to the habitual one period lag applied to the series for the US, the first equation is as intended, because the realization of $s_{sg,t-1}$ precedes the generation of $s_{us,t-1}$. However, due to taking this lag, the equation for Singapore becomes unsatisfactory, because for explaining $s_{sg,t}$ the latest available information on the US, being $s_{us,t-1}$, is not utilized. Consequences of this for VAR models in returns we will examine in the next subsection.

TABLE 8.
Normalized cointegration relationships as perceived from the five markets models (std. errors between parentheses)*

sample	Hong Kong	Singapore	Korea	Shanghai	US	Constant
last sub-sample	1	0	0.24	0.03	-4.60^{***}	21.14^{***}
			(0.20)	(0.14)	(0.90)	(4.61)
		1	0.28^{**}	0.16^{*}	-3.90^{***}	16.92^{***}
			(0.13)	(0.09)	(0.58)	(2.98)

* Significance at level 10, 5 and 1% is indicated by: *, ** and ***.

3.2.3. Causality analysis

We continue to illustrate the devastating effects on inference when using underspecified models and now turn to Granger-causality tests. For that purpose we estimate VAR(l) models for $l = 0, 1, 2, ...$ on returns (the first difference of the natural logarithm of the stock index) and again examine the effects of increasing the number of included markets $m = 2, ..., 5$. All equations have an intercept but not a linear trend (because its presence would be unacceptable according to finance theory). We just give results for the full sample size. For cases where existence of cointegration relationships is very likely the VAR model in returns omits the "error correction term", so ideally it should not pass diagnostic checks and it should not be used for Granger-causality testing (also because joint dependence is already implied by cointegration). The tables report results for each of the dependent variables (Dep) of the system in turn. For each group of lagged regressors the p-value of the F-test for their joint-significance is given, followed by the smallest p-value for the significance test of an individual lagged variable, followed by the lag order at which it occurs. Again results on the five diagnostics are given in the usual fashion, although, unlike what we did in the preceding subsection, we now present their single equation versions (so they do not involve residuals from the other equations). Due to the concern regarding the treatment of different time zones, we will (when the US is included in the VAR system) also present results for separate single equation regressions which always have as regressors the most recently already realized returns at the other markets and their lags.

In Table 9 bivariate VAR results are given for Hong Kong and Singapore, for which Table 5 provided some dubious evidence of possible cointegration. SIC and AIC suggest the lag order to be 0 and 1 respectively. Since some t-values are significant lag-order zero is clearly not acceptable. JB is always significant. However, for the F and t causality tests nonnormality of the disturbances should not be a serious problem, given the size of the

sample. At $l = 1$, we detect no heteroskedasticity, and evidence of one causality relation is found, namely that Singapore Granger-causes Hong Kong. However, this result is doubtful, because of serious serial correlation at lag 1 in this relation. Increasing the lag order to 2 resolves the serial correlation problem and yields the same causality findings, although heteroskedasticity emerges now in the relation for Singapore. Oddly enough, at $l = 8$ also the relationship for Hong Kong becomes heteroskedastic, whereas serial correlation re-emerges for the Singapore equation.

In Table 10 trivariate VAR results are given for Hong Kong, Singapore and Korea, for which in Table 6 no evidence of cointegration could be established. Both AIC and SIC suggest the lag order to be 0 again, which is clearly misleading, given the significance of t and F tests. At $l = 1$, two causality relationships seem detected, namely that Singapore Granger-causes both Hong Kong and Korea. However, both equations show significant first-order serial correlation, so before drawing any conclusions first the lag order of the VAR model should be increased. For lag order 2 the causality relationships are still the same, but BP becomes significant, so at least heteroskedasticity robust t and F statistics should be used now. At lag order 8 again some serial correlation diagnostics become significant, whereas now Singapore seems to be Granger-caused by both Hong Kong and Korea too. Note that it is hard to explain the rather erratic findings in Tables 9 and 10, unless one supposes that in fact none of the VAR models in these tables are nested in the actual data generating process for the returns of these three markets, simply because unintentionally particular profound determining factors have been omitted.

TABLE 9.

P-values of Granger-causality tests and single equation diagnostics in VAR(l) models for the returns of the stock indices of Hong Kong and Singapore.

		— HK —			— SG —			— order 2 —					— order 10 —			— Min —	
l	Dep	F	t	lag	F	t	lag	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.90	0.90	1	0.03	0.03	2	0.35	0.00	0.61	0.06	0.06	0.97	0.61	0.61	0.02	1
	SG	0.67	0.67	1	0.85	0.85	1	0.29	0.00	0.69	0.63	0.63	0.27	0.26	0.26	0.14	7
2	HK	0.85	0.55	2	0.01	0.03	2	0.20	0.00	1.00	0.97	0.97	0.98	0.98	0.98	0.82	1
	SG	0.85	0.61	1	0.69	0.40	2	0.00	0.00	1.00	0.91	0.91	0.34	0.30	0.31	0.17	7
8	HK	0.92	0.27	8	0.11	0.04	1	0.00	0.00	1.00	0.56	0.56	1.00	0.56	0.58	0.14	5
	SG	0.86	0.13	7	0.20	0.05	6	0.00	0.00	1.00	0.58	0.58	1.00	0.04	0.04	0.02	7

Table 11 reports results of pentavariate VAR models for all five markets. BP and JB are always significant at the various lag orders; we removed their p-values from the table in order to save space. SIC and AIC favor

TABLE 10.

P-values of Granger-causality tests and single equation diagnostics in VAR(l) models for the returns of the stock indices of Hong Kong,Singapore and Korea.

		— HK —			— SG —			— KR —			— order 2 —				— order 10 —			— Min —	
l	Dep	F	t	lag	F	t	lag	F	t	lag	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.78	0.78	1	0.02	0.02	1	0.30	0.30	1	0.24	0.00	0.65	0.13	0.13	0.96	0.73	0.74	0.05	1
	SG	0.39	0.39	1	0.61	0.61	1	0.21	0.21	1	0.29	0.00	0.72	0.70	0.70	0.26	0.24	0.24	0.14	7
	KR	0.52	0.52	1	0.02	0.02	1	0.32	0.32	1	0.31	0.00	0.53	0.08	0.08	0.47	0.16	0.17	0.03	1
2	HK	0.75	0.51	2	0.01	0.02	1	0.65	0.37	1	0.00	0.00	1.00	0.96	0.96	0.98	0.98	0.98	0.79	1
	SG	0.63	0.36	1	0.68	0.45	2	0.49	0.23	1	0.00	0.00	1.00	0.89	0.89	0.32	0.26	0.26	0.15	7
	KR	0.71	0.44	1	0.02	0.03	1	0.68	0.40	1	0.00	0.00	0.97	0.17	0.17	0.55	0.33	0.34	0.17	2
8	HK	0.95	0.30	7	0.11	0.03	1	0.89	0.34	4	0.00	0.00	1.00	0.14	0.15	1.00	0.33	0.35	0.09	1
	SG	0.53	0.02	7	0.30	0.03	6	0.36	0.01	7	0.00	0.00	0.97	0.10	0.10	1.00	0.02	0.03	0.01	6
	KR	0.85	0.28	5	0.04	0.01	5	0.55	0.12	7	0.00	0.00	1.00	0.68	0.69	1.00	0.05	0.05	0.03	8

the lag orders 0 and 2 respectively. At $l = 1$ three causality relationships are found, namely that Singapore Granger-causes Hong Kong, Korea and the US. However, the equations for Singapore, Korea and the US suffer from serial correlation at the 5% level. This suggests to increase the lag order. For $l = 2$ four additional causality relationships emerge, namely that the US Granger-causes Hong Kong, Singapore and Korea, and that Korea, like Singapore, Granger-causes the US. However, the relation for Singapore shows significant serial correlation. Even taking $l = 8$ does not resolve this problem, but in fact aggravates it. Moreover, the VAR(8) model suggests another seven causality relationships. Table 12 reports for the same five markets the results of separate single equation regressions in which those for the Asian markets now include the most recent realized return for the US. The most striking differences with Table 11 are that for $l = 1$, now the US seems to Granger-cause Hong Kong, Singapore and Korea, while the Granger-causality relationships from Singapore to Hong Kong and Korea become insignificant. This highlights that just mechanically coping with the time zone difference in VAR models is inappropriate and can have devastating effects on inference.

TABLE 11: P-values of Granger-causality tests and single equation diagnostics in VAR(l) models for the returns of the stock indices of Hong Kong, Singapore, Korea, Shanghai and US.

		— HK —			— SG —			— KR —			— SH —			— US —			— order 2 —			— order 10 —			— Min —	
l	Dep	F	t	lag	F	t	lag	F	t	lag	F	t	lag	F	t	lag	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.80	0.80	1	0.02	0.02	1	0.30	0.30	1	0.24	0.24	1	0.82	0.82	1	0.62	0.24	0.24	0.96	0.85	0.85	0.09	1
	SG	0.30	0.30	1	0.54	0.54	1	0.25	0.25	1	0.82	0.82	1	0.31	0.31	1	0.80	0.18	0.18	0.19	0.09	0.09	0.04	6
	KR	0.55	0.55	1	0.02	0.02	1	0.32	0.32	1	0.85	0.85	1	0.82	0.82	1	0.52	0.11	0.11	0.46	0.20	0.20	0.04	1
	SH	0.20	0.20	1	0.67	0.67	1	0.51	0.51	1	0.51	0.51	1	0.30	0.30	1	0.94	0.76	0.76	0.30	0.23	0.23	0.23	10
	US	0.24	0.24	1	0.00	0.00	1	0.14	0.14	1	0.98	0.98	1	0.00	0.00	1	0.27	0.00	0.00	0.27	0.02	0.02	0.00	1
2	HK	0.99	0.86	1	0.00	0.02	1	0.49	0.31	1	0.24	0.23	2	0.01	0.00	2	0.98	0.28	0.29	0.98	0.85	0.86	0.29	2
	SG	0.60	0.36	1	0.55	0.31	2	0.44	0.22	1	0.86	0.61	2	0.00	0.00	2	0.96	0.01	0.01	0.25	0.02	0.03	0.01	2
	KR	0.65	0.49	2	0.01	0.03	1	0.62	0.32	1	0.94	0.75	2	0.05	0.02	2	0.96	0.16	0.16	0.45	0.24	0.25	0.16	2
	SH	0.12	0.10	2	0.47	0.25	2	0.72	0.50	1	0.78	0.48	1	0.61	0.33	1	1.00	0.43	0.43	0.31	0.18	0.19	0.19	10
	US	0.41	0.20	1	0.00	0.00	1	0.00	0.00	2	0.96	0.78	2	0.00	0.00	1	1.00	1.00	1.00	0.46	0.45	0.46	0.21	7
8	HK	0.95	0.22	7	0.04	0.01	2	0.72	0.11	4	0.01	0.00	6	0.19	0.00	2	0.99	0.13	0.14	1.00	0.13	0.15	0.03	5
	SG	0.39	0.01	7	0.15	0.01	6	0.29	0.01	7	0.27	0.08	8	0.00	0.00	2	0.99	0.62	0.64	0.99	0.03	0.04	0.03	8
	KR	0.74	0.29	6	0.01	0.01	5	0.45	0.14	4	0.02	0.01	4	0.08	0.02	2	0.97	0.06	0.06	1.00	0.03	0.04	0.04	8
	SH	0.19	0.09	8	0.19	0.00	5	0.85	0.29	6	0.20	0.06	3	0.79	0.12	8	0.99	0.46	0.47	0.89	0.34	0.38	0.38	10
	US	0.40	0.04	7	0.01	0.00	1	0.01	0.00	2	0.14	0.00	6	0.00	0.00	1	1.00	0.62	0.63	1.00	0.27	0.30	0.29	8

TABLE 12: P-values of Granger-causality tests and diagnostics of single equation for the returns of the stock indices of Hong Kong, Singapore, Korea, Shanghai and US.

		— HK —			— SG —			— KR —			— SH —			— US —			— order 2 —			— order 10 —			— Min —	
l	Dep	F	t	lag	F	t	lag	F	t	lag	F	t	lag	F	t	lag	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.88	0.88	1	0.19	0.19	1	0.12	0.12	1	0.23	0.23	1	0.00	0.00	1	0.03	0.00	0.00	0.55	0.15	0.15	0.00	2
	SG	0.41	0.41	1	0.58	0.58	1	0.09	0.09	1	0.83	0.83	1	0.00	0.00	1	0.06	0.03	0.03	0.02	0.01	0.01	0.00	6
	KR	0.53	0.53	1	0.16	0.16	1	0.17	0.17	1	0.78	0.78	1	0.00	0.00	1	0.02	0.00	0.00	0.08	0.02	0.02	0.00	2
	SH	0.13	0.13	1	0.59	0.59	1	0.57	0.57	1	0.54	0.54	1	0.91	0.91	1	0.96	0.85	0.85	0.20	0.15	0.15	0.15	10
	US	0.24	0.24	1	0.00	0.00	1	0.14	0.14	1	0.98	0.98	1	0.00	0.00	1	0.27	0.00	0.00	0.27	0.02	0.02	0.00	1
2	HK	0.62	0.34	2	0.07	0.04	2	0.07	0.09	2	0.21	0.14	1	0.00	0.00	1	0.16	0.00	0.00	0.86	0.01	0.01	0.00	2
	SG	0.66	0.49	1	0.59	0.41	1	0.13	0.10	1	0.96	0.83	2	0.00	0.00	1	0.27	0.00	0.00	0.09	0.00	0.00	0.00	2
	KR	0.96	0.77	1	0.13	0.09	2	0.24	0.13	1	0.84	0.56	2	0.00	0.00	1	0.53	0.00	0.00	0.53	0.05	0.05	0.00	2
	SH	0.12	0.10	2	0.44	0.22	2	0.73	0.50	1	0.76	0.48	1	0.61	0.33	2	1.00	0.90	0.90	0.21	0.14	0.15	0.15	10
	US	0.42	0.20	1	0.00	0.00	1	0.00	0.00	2	0.96	0.77	2	0.00	0.00	1	1.00	1.00	1.00	0.46	0.45	0.46	0.21	7
8	HK	0.93	0.22	8	0.03	0.00	2	0.17	0.05	2	0.02	0.03	4	0.00	0.00	1	1.00	0.99	0.99	1.00	0.94	0.95	0.90	7
	SG	0.76	0.09	7	0.01	0.01	5	0.10	0.01	7	0.24	0.05	3	0.00	0.00	1	0.98	0.26	0.27	1.00	0.05	0.06	0.04	6
	KR	0.86	0.24	6	0.00	0.01	2	0.26	0.16	7	0.01	0.01	4	0.00	0.00	1	0.97	0.16	0.18	1.00	0.82	0.84	0.09	1
	SH	0.14	0.05	8	0.14	0.01	5	0.74	0.22	6	0.17	0.05	8	0.99	0.45	6	1.00	0.81	0.82	0.88	0.07	0.08	0.08	10
	US	0.40	0.04	7	0.01	0.00	1	0.01	0.00	2	0.14	0.00	6	0.00	0.00	1	1.00	0.62	0.63	1.00	0.27	0.30	0.29	8

Focussing only on those relationships in Table 12 which are not rejected by any of the serial correlation tests, and giving much more weight to the p-value of an F test than to an individual t tests (especially when l is large and a significant t test is just found at a high lag order), one could with substantial reservation infer that: (i) Hong Kong seems Granger-caused by all four other markets; (ii) Korea seems Granger-caused by Singapore, Shanghai and the US; Shanghai seems Granger-caused by none of the four markets, because the significant individual t-values at lag 5 for Singapore and at lag 8 for Hong Kong could just be incidental; the US seems Granger-caused by Singapore and Korea (and possibly very mildly by Shanghai); whereas for Singapore we have not found a regression model from this limited data set which passes the serial correlation diagnostics.

However, in this study our purpose was not to make strong claims regarding the major determining factors of the indices and returns at these five markets. Our primary aim was to illustrate that a few isolated estimates and tests may seem very suggestive regarding the significance of particular relationships, but only when such results pass a relevant battery of diagnostic tests and also withstand extensions of the model, by including more markets and deeper lags they deserve to be taken seriously. Note that Table 10 suggests that Korea does not Granger-cause Hong Kong, whereas Table 12, after allowing the US to play its role as well, indicates the contrary.

These illustrations of causality testing have not only indicated again that information criteria cannot be trusted in determining the optimal lag order structure of a VAR model, but also that they certainly cannot reveal whether a VAR model suffers from omitted relevant determining variables such as lagged variables from other markets. In both respects diagnostic tests seem much more useful, but they are certainly no panacea. The only trustworthy tests against omitted variables seem to be the constructive tests which include and test the significance of the actual earlier omitted variables concerned. The common practice of VAR causality testing of return series without much concern regarding possibly omitted markets can be much improved, as we showed, especially for the situation that different time zones are involved. In that case, instead of a system, separate adequately specified single equations should be subjected to causality tests.

4. CONCLUSION

Much too often practitioners produce (or uncritically cite published) empirical results for which too few (or improper) diagnostics have been eval-

TABLE 13.

Overview of major econometric shortcomings in the reviewed studies

	A	A1	B	B1	C1	C2	C3	C4	C5	D	E
Aityan et al. (2010)	✓		✓								
Arshanapalli et al. (1995)	✓	✓	✓	✓		✓			✓		
Beirne et al. (2010)	✓	✓							✓		
Burdekin & Siklos (2012)	✓					✓	✓				
Cheng & Glascock (2005)	✓	✓	✓	✓		✓					✓
Chow et al. (2011)	✓		✓			✓				✓	✓
Chung & Liu (1994)	✓	✓			✓				✓		
Glick & Hutchison (2013)	✓	✓	✓	✓			✓				✓
Gosh et al. (1999)	✓	✓	✓	✓							
Groenewold et al. (2004)	✓										
Huang et al. (2000)	✓	✓	✓	✓							
Huygebaert & Wang (2010)	✓	✓	✓	✓				✓			
Jayasuriya (2011)	✓	✓			✓				✓		
Kenett et al. (2012)	✓		✓								
Li (2007)	✓	✓	✓	✓					✓		
Li (2012)	✓								✓		
Singh et al. (2010)	✓	✓	✓	✓		✓					✓
Wang (2014)	✓	✓	✓	✓				✓			
Yang et al. (2006)	✓	✓									
Zhang & Li (2014)	✓		✓								✓
Zhang et al. (2009)	✓		✓	✓					✓		
Zhu et al. (2004)	✓		✓	✓				✓			

A = insufficient checks for omitted variables
A1 = and in particular insufficient checks for coefficient nonconstancy
B = insufficient diagnostic checking
B1 = and in particular insufficient appropriate testing for error serial correlation
C1 = inappropriate use of tests for normality
C2 = inappropriate use of t-tests
C3 = inappropriate use of tests of correlation
C4 = using joint restrictions test, where single restrictions are required too
C5 = using only high-order serial correlation tests
D = neglecting identification problems
E = presenting alternative results with mutually conflicting maintained hypotheses

uated that do support their findings. As a rule all statistical evidence is based on an extensive set of adopted model assumptions. Evidence supporting the likely validity of many of these assumptions can usually be provided by so-called diagnostic tests or by demonstrating that the main-

tained model is not rejected when tested against a less restrictive alternative model formulation. By scrutinizing a great number of peer reviewed published papers on analyzing the interlinkages between various stock markets we demonstrate that in this literature there exists a tendency to take any empirical findings serious irrespective by what methodology and statistical techniques they have been obtained, whereas employing unsound statistical methodology is ubiquitous. We plea for bringing methodological issues more to the forefront, and stopping citing empirical results which have clearly been obtained by methods that as a rule will lead to strongly biased results for which neither statistical accuracy nor significance can be assessed in a proper way. This widespread submissive uncritical citation attitude makes one wonder too what the intrinsic value is of citation based indices on the quality of academic journals, because in the niche of studies examined here it is obvious that citations are obtained simply because of the particular topic being studied and not primarily because the study has been a useful stepping stone for further and deeper study. In Table 13 we give an overview of the various observations made regarding methodological shortcomings, dividing them in five major categories and some subcategories.

Evaluations using empirical data illustrate the serious effects of wrongly omitted explanatory variables, such as variables of other influential markets and significant higher order lagged variables. Such kind of omitted variables affect the results of cointegration tests and Granger-causality tests, which are often the basis for further analyses and inference. In addition, particular popular information criteria are shown to provide seriously misleading results for assessing the appropriate lag length of VAR models. It is demonstrated that it is of great importance to conduct thorough checks and diagnostic tests to make sure that the underlying assumptions of the regressions are met, whereas by simulation it is shown that particular often used tests for serial correlation are deficient in this respect, whereas more adequate alternative versions have been developed.

APPENDIX

A.1. ON OMITTED REGRESSORS

Suppose a relationship is examined for which a simple multiple regression model would be adequate and ordinary least-squares techniques would provide valid inferences (at least in samples of substantial size). What

would be the consequences if some of the regressors were omitted from the regression? To answer this question it suffices to consider the simple two regressor model

$$y_t = \beta x_t + \gamma w_t + \varepsilon_t,$$

where x_t and w_t are scalar variables that have been observed together with the dependent variable y_t for observations $t = 1, ..., n$, whereas all three have been taken in deviation from their sample average, which permits to remove the intercept term from the regression. Extension to the case where both x_t and w_t are in fact vectors of variables, possibly containing various sequences of lagged variables as in VAR models as used in Granger-causality tests, are straight-forward, though mathematically more involved, and therefore avoided here.

We focus on the case where the data are time-series. Let $x^t = (x_1 \cdots x_t)'$ and $w^t = (w_1 \cdots w_t)'$. Appropriate interpretability of standard least-squares analysis requires validity of the three assumptions: (1) predeterminedness of all regressors, that is $E(\varepsilon_t \mid x^t, w^t) = 0$; (2) serial uncorrelatedness of the disturbances, that is $E(\varepsilon_t \varepsilon_s) = 0$ for $t \neq s$; and (3) homoskedasticity of the disturbances, or $E(\varepsilon_t^2) = \sigma^2$. Violation of (1) leads to biased estimates, irrespective of the size of the sample, and so does violation of (2) in VAR models. A major consequence of violation of (2) or (3) is that the standard estimator for the variance of the coefficient estimates will be biased, which renders inferences based on standard test statistics and confidence intervals misleading. If the sample is not too small normality of the distribution of the disturbances is not essential. Of course, in practice validity of the three assumptions should always be verified as far as is possible. When interpreting standard least-squares inference on β and γ these three assumptions establish the so-called *maintained hypotheses*, which are required to be plainly valid.

Now suppose that a researcher deliberately or unknowingly omits regressor w_t. Then least-squares yields for β the estimator

$$\hat{\beta} = \Sigma_t x_t y_t / \Sigma_t x_t^2 = \beta + \gamma \Sigma_t x_t w_t / \Sigma_t x_t^2 + \Sigma_t x_t \varepsilon_t / \Sigma_t x_t^2.$$

Here the third term is the unavoidable random estimation error which varies around zero and decreases in magnitude with sample size. The second term represents a systematic deviation of $\hat{\beta}$ from β. This has a magnitude which does not decrease with the size of the sample; it does depend on four

autonomous factors, since

$$\gamma \Sigma_t x_t w_t / \Sigma_t x_t^2 = \gamma r_{x,w} s_w \frac{1}{s_x},$$

where $r_{x,w}$ is the sample correlation between variables x_t and w_t, s_w is the sample standard deviation of the observations w_t, and similar for s_x. Hence, this systematic error (the inconsistency of $\hat{\beta}$) is small provided the product of these four factors is small. So, in case some of them are large, the remaining ones should be sufficiently small. If this is the case, estimator $\hat{\beta}$ will as a rule not differ much from its estimate obtained without omission of regressor w_t. However, even if that occurs, omission will nevertheless be harmful for least-squares inference in case the disturbances of the model omitting w_t (these are now represented by $\gamma w_t + \varepsilon_t$) do not obey the three fundamental conditions. It should be obvious that in case $\gamma \neq 0$ and w_t represents a series which is correlated with x_t then the predeterminedness assumption does no longer hold (but the consequences may be mild if γ and/or s_w/s_x are small). If $\gamma \neq 0$ and the w_t series exposes heteroskedasticity and/or serial correlation then the other two conditions may not hold in the model that through omission of w_t incorrectly imposes $\gamma = 0$. In that case inference on β will be unreliable because the required maintained hypotheses are not all satisfied. If it is just condition (3) that does not hold, it is possible to repair the inaccuracies of least-squares inference (make them robust to heteroskedasticity). In VAR-type models this is generally not possible when condition (2) is not fulfilled.

Hence, the model omitting w_t is only really superior to the model including w_t if $\gamma = 0$ and omission is fully justified. Otherwise, when $\gamma \neq 0$ and the omitted variable w_t is correlated with x_t (for instance, because it is just the one-period lag of smooth series x_t) or if omission does not lead to substantial bias, but the omitted variable is itself not time-independent, then proper interpretation of least-squares inference from the model wrongly imposing $\gamma = 0$ is impossible. This situation should be prevented by always carefully verifying whether the three maintained hypotheses on which least-squares inference is built do actually hold (see Appendix A.2). If they do not, the effect of x_t on y_t can only be assessed either after a proper re-specification of the model, or by using an alternative for standard least-squares estimation.

A.2. ON DIAGNOSTIC TESTING

Diagnostic testing refers to all attempts made to verify validity of the *maintained hypotheses*. By these we indicate here all statistical aspects of the model that have to hold in order to be able to accurately interpret all inference techniques that one intends to employ regarding its parameter values. These statistical aspects should establish at least: (a) consistency of the estimators (estimators that are right on target in infinitely large samples), (b) confidence intervals for the unknown coefficient values with an actual coverage probability which converges for increasing sample size to the chosen nominal level (often 95%), and (c) actual significance levels (type I error probabilities) of tests on coefficient values which converge to their chosen nominal level of $100\times\alpha\%$. When one chooses to analyze a linear model by standard least-squares techniques the maintained hypotheses are the three stated in Appendix A.1.

From Appendix A.1 it should be clear that when in a linear model that has been estimated by ordinary least-squares one finds significant test outcomes for heteroskedasticity or for serial correlation the proper diagnosis might be that in fact improper restrictions have been imposed on regression coefficients. Hence, this may not call for some form of weighted least-squares or autoregressive transformation of the model, but for finding the omitted regressors. Candidates are not only really additional new variables, but also transformations of already included variables, such as lags, logs, squares, cross-products (interaction terms), dummy variables and cross-products with dummy variables (to represent variation in reaction coefficients over sub-samples).

Note that the predeterminedness assumption of a regressor may be violated due to an omitted regressor with which it happens to be correlated. This is resolved by including the omitted regressor. On the other hand, a regressor may be endogenous instead of predetermined because of reverse causality. This occurs when y_t is determined by x_t, while y_t itself is also one of the explanatory variables of x_t. Then x_t, since it depends on y_t, must also be correlated with ε_t. In that case variables y_t and x_t are called jointly dependent and valid inference on β (the effect of x_t on y_t) cannot be obtained by ordinary least-squares. However, provided the equation for y_t has sufficient unique characteristics (usually in the form of omitted variables whose valid omission is beyond doubt) which guarantee identification (see Appendix A.4), valid inference can be obtained by using the correctly omitted regressors as so-called external instrumental variables in two-stage least-squares estimation. For the latter the predeterminedness assumption

on x_t in the maintained hypotheses has to be replaced by so-called exclusion restrictions (correctly omitted regressors). So, the moral is here: wrongly omitted regressors are as a rule harmful for inference, whereas rightly omitted regressors are obligate for valid inference when reverse causality may be at play. In the latter case the precision of inference will be poor when x_t depends only weakly on the instrumental variables.

So, useful diagnostic tests are tests for heteroskedasticity (standard software provides tests against various types of heteroskedasticity), tests for serial correlation of suitable order (the popular Durbin-Watson test just checks first-order serial correlation and presupposes that all regressors are strictly exogenous, so there should be no lagged-dependent variables among the regressors as in VAR models), tests for the significance of omitted relevant variables (many implementations are possible, also covering tests for structural changes and for nonlinear dependencies). Also appropriate tests for serial correlation and tests for reverse causality take the form of omitted regressor tests because they boil down to testing the significance of particular (transformations of) residuals when added to the original model. Only when a model specification has been found which passes a battery of diagnostic tests (all yielding p-values preferably well above 0.10, or at least above 0.05), then building on the maintained hypotheses implied by the chosen estimation technique seems warranted. This finally opens the door to the phase in which inference on the coefficients of the model can be produced and interpreted with reasonable degree of trust.

A.3. ON SIGNIFICANCE TESTS OF SIMPLE CORRELATIONS

As is well-known a test on the significance of β in the simple linear regression model $y_t = \alpha + \beta x_t + \varepsilon_t$ $(t = 1, ..., n)$ is obtained by the least-squares based test statistic $t = \hat{\beta}/se(\hat{\beta})$, where $\hat{\beta} = \Sigma_t \tilde{x}_t \tilde{y}_t / \Sigma_t \tilde{x}_t^2$ and $se(\hat{\beta}) = [(n-2)^{-1} \Sigma_t e_t^2 / \Sigma_t \tilde{x}_t^2]^{1/2}$. Here $\tilde{x}_t = x_t - \bar{x}$ with $\bar{x} = n^{-1} \Sigma_t x_t$ (and likewise for $\tilde{y}_t$) and $e_t = y_t - \hat{\alpha} - \hat{\beta} x_t = \tilde{y}_t - \hat{\beta} \tilde{x}_t$, because $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$. Some simple algebraic manipulations show that $t = \hat{\rho}/[(1-\hat{\rho}^2)/(n-2)]^{1/2}$, where $\hat{\rho} = \Sigma_t \tilde{x}_t \tilde{y}_t / [\Sigma_t \tilde{x}_t^2 \Sigma_t \tilde{y}_t^2]^{1/2}$ is the simple sample correlation coefficient of the series $\{y_t, x_t; t = 1, ..., n\}$.

Therefore, testing whether or not the correlation $\rho = Cov(y_t, x_t)/[Var(x_t) \cdot Var(y_t)]^{1/2}$ is zero by reference to the statistic t and a critical value of the Student distribution with $n-2$ degrees of freedom requires validity of similar maintained hypotheses as testing the significance of β in the simple linear regression model. Sufficient for that is $\varepsilon_t \mid x^t \sim IIN(0, \sigma^2)$, which

represents the three maintained hypotheses mentioned in Appendix A.1 plus normality of the disturbances. In samples of reasonable size the normality is of minor importance. However, for asymptotic validity of the test it is essential that subtracting from y_t its conditional expectation expressed linearly in just the single variable x_t should remove all its time-dependence and heteroskedasticity and thus yield random white-noise.

Hence, a test on the significance of a simple correlation coefficient can only be credibly and usefully interpreted when supplemented at least by insignificant test statistics on both (higher-order) serial correlation and heteroskedasticity of the disturbances of the corresponding simple linear regression, and preferably also some further evidence on constancy of the coefficients α and β and absence of the devastating effects of possibly omitted explanatory variables. Thus, if a VAR specification for y_t seems credible, then performing the above test on correlation (which assumes a static regression) does not make sense.

A.4. ON LACK OF IDENTIFICATION

Consider a system, for which scalar variables $y_t^{(1)}$, $y_t^{(2)}$ and x_t $(t = 1, ..., n)$ have been observed for a sample of size n, that is given by the two equations

$$y_t^{(1)} = \beta_1 y_t^{(2)} + \gamma_1 x_t + \varepsilon_t^{(1)},$$
$$y_t^{(2)} = \beta_2 y_t^{(1)} + \gamma_2 x_t + \varepsilon_t^{(2)}.$$

Here the unobserved variables $\varepsilon_t^{(j)}$ $(j = 1, 2)$ are disturbance terms, and x_t is in both equations an explanatory variable with unknown coefficient values γ_1 and γ_2. The variables $y_t^{(j)}$ $(j = 1, 2)$ are characterized by reverse causality, provided both β_1 and β_2 are nonzero. To keep things simple, we assume that both disturbances are white-noise series, which are not necessarily mutually uncorrelated, and that x_t (which could easily be extended to a vector of control variables) is predetermined and thus uncorrelated with both current disturbance terms.

Without any further assumptions, for instance on the actual value of at least one of the four coefficients β_j and γ_j $(j = 1, 2)$, all these coefficients are unidentified. This expresses that, whatever method one uses to obtain estimates $\hat{\beta}_j$ and $\hat{\gamma}_j$ $(j = 1, 2)$, it is impossible to find out whether these refer to the above system parameters or to some hidden linear transformations of them.

This can be clarified as follows. Let c_1 and c_2 be arbitrary real scalars. Now consider the rescaled system

$$c_1 y_t^{(1)} = c_1\beta_1 y_t^{(2)} + c_1\gamma_1 x_t + c_1\varepsilon_t^{(1)},$$
$$c_2 y_t^{(2)} = c_2\beta_2 y_t^{(1)} + c_2\gamma_2 x_t + c_2\varepsilon_t^{(2)}.$$

By addition of these two equations we obtain

$$(c_1 - c_2\beta_2)y_t^{(1)} = (c_1\beta_1 - c_2)y_t^{(2)} + (c_1\gamma_1 + c_2\gamma_2)x_t + (c_1\varepsilon_t^{(1)} + c_2\varepsilon_t^{(2)}).$$

From this we can again establish a two-equation system by normalizing the coefficients of $y_t^{(1)}$ and $y_t^{(2)}$ respectively, giving

$$y_t^{(1)} = \frac{c_1\beta_1 - c_2}{c_1 - c_2\beta_2}y_t^{(2)} + \frac{c_1\gamma_1 + c_2\gamma_2}{c_1 - c_2\beta_2}x_t + \frac{c_1\varepsilon_t^{(1)} + c_2\varepsilon_t^{(2)}}{c_1 - c_2\beta_2},$$
$$y_t^{(2)} = \frac{c_2\beta_2 - c_1}{c_2 - c_1\beta_1}y_t^{(1)} + \frac{c_1\gamma_1 + c_2\gamma_2}{c_2 - c_1\beta_1}x_t + \frac{c_1\varepsilon_t^{(1)} + c_2\varepsilon_t^{(2)}}{c_2 - c_1\beta_1}.$$

Of course we assumed here that c_1 and c_2 are such that $c_1/c_2 \neq \beta_2$ and $c_2/c_1 \neq \beta_1$. Using obvious definitions for the new symbols introduced below, the latter system can also be expressed as

$$y_t^{(1)} = \beta_1^* y_t^{(2)} + \gamma_1^* x_t + u_t^{(1)},$$
$$y_t^{(2)} = \beta_2^* y_t^{(1)} + \gamma_2^* x_t + u_t^{(2)},$$

which demonstrates that it is indistinguishable from the initial one. This highlights that each of its equations cannot be identified from an arbitrary linear combination of the two equations after this has been scaled with respect to the same left-hand variable.

In fact, if one would use least-squares to estimate the equation with $y_t^{(1)}$ as the dependent variable, then one will obtain coefficient estimates which minimize the residual sum of squares. This corresponds to implicitly choosing c_1 and c_2 such that, if $Var(\varepsilon_t^{(1)}) = \sigma_1^2$, $Var(\varepsilon_t^{(2)}) = \sigma_2^2$ and $Cov(\varepsilon_t^{(1)}, \varepsilon_t^{(2)}) = \sigma_{12}$, the resulting

$$Var(u_t^{(1)}) = \frac{c_1^2\sigma_1^2 + 2c_1c_2\sigma_{12} + c_2^2\sigma_2^2}{(c_1 - c_2\beta_2)^2}$$

will be minimal.

A.5. CHICKEN AND EGG IN TESTING

In Appendix A.2 we warned against interpreting inference on coefficients of a model produced by estimators and test statistics before sufficient evidence has been produced by diagnostic tests demonstrating the credibility of all the maintained hypotheses. Now one may wonder whether these diagnostic tests themselves can safely be interpreted. They may require their own maintained hypotheses whose validity has not been verified yet. How to proceed?

Let the current specification of a relationship be represented by the simple model $y_t = \beta x_t + \varepsilon_t$ $(t = 1, ..., n)$, where for the sake of simplicity x_t and β are again scalars. The candidate set of maintained hypotheses for inference based on standard least-squares estimates is: $E(\varepsilon_t \mid x^t) = 0$, $E(\varepsilon_t \varepsilon_s) = 0$ for $t \neq s$, and $E(\varepsilon_t^2) = \sigma^2$. Now consider the extended model $y_t = \beta x_t + \gamma w_t + \varepsilon_t^*$, where $\varepsilon_t^* = \varepsilon_t - \gamma w_t$. Also w_t is taken as a scalar here, but it could easily be generalized to be a vector containing possibly omitted regressors (for instance dummy variables and interactions of the elements of x_t with those dummies, or squares of elements of x_t or really additional alternative variables) or w_t may contain constructs like lagged residuals or reduced form residuals as in particular diagnostic tests on serial correlation of disturbances and endogeneity of some of the regressors. Now, in case γ is scalar, the diagnostic test is based on the usual t-ratio for testing the null-hypothesis $\gamma = 0$ in the extended regression. Rejection or not of this hypothesis, which will be interpreted as rejection or not of the maintained hypotheses of the model with just the regressors x_t, is based on comparing this test statistic with a critical value of the t distribution. Why is this a sound procedure?

Note that (at least in large samples) the test statistic will be distributed according to the appropriate t distribution if the maintained hypotheses in the regression with just regressor x_t do hold indeed and additionally $E(\varepsilon_t \mid w_t) = 0$ which implies $\gamma = 0$. When finding a significant test statistic this may be either due to validity of the maintained hypotheses, though having obtained (with a small probability converging towards $100\alpha\%$) an unlikely extreme draw from the t distribution (this then leads to a type I error), or due to $E(\varepsilon_t \mid w_t) \neq 0$ and thus $\gamma \neq 0$. The latter in principle implies invalidity of the maintained hypotheses under test, because it seems very likely that $E(\varepsilon_t \mid x_t) = E(\gamma w_t + \varepsilon_t^* \mid x_t) \neq 0$ when $E(w_t \mid x_t) \neq 0$. In addition, possibly $Var(w_t \mid x_t)$ is nonconstant and $Cov(w_t w_s \mid x_t) \neq 0$ for some $t \neq s$, which would also imply invalidity of the maintained hypotheses. Only in case $E(\varepsilon_t \mid w_t) \neq 0$ and thus $\gamma \neq 0$, whereas at the same time w_t is

such that $E(w_t \mid x_t) = 0$, and $Var(w_t \mid x_t) = \sigma_w^2$ and also $Cov(w_t w_s \mid x_t) = 0$ for $t \neq s$, the test statistic will not follow the t distribution and thus may reject with probability exceeding α, although the maintained hypotheses could be valid. In these latter extra-ordinary circumstances (note that the three required characteristics of the moments of w_t can be verified from the observed data) γw_t is such that omitting it is relatively harmless, because it can fully be absorbed in the disturbances without affecting the maintained hypotheses. However, using the significant diagnostic to re-specify the model may in fact be preferable. Its maintained hypotheses will be valid too if those of the original model were valid.

Hence, we conclude that a significant diagnostic should always lead to re-specification of the model and a corresponding reformulation of the maintained hypotheses. On the other hand, very little can be concluded from an insignificant diagnostic. Of course, it could be a consequence of validity of the maintained hypotheses, but it could as well be due to lack of power of the particular diagnostic test. This will occur when (some elements of) the maintained hypotheses are invalid, but variable w_t embodies insufficiently the aspects that have been omitted from the current specification of the model for the analyzed actual data generating process. This, for instance, happens when a regressor v_t has wrongly been omitted, whereas the variable w_t used in the diagnostic test is hardly correlated with v_t.

Note that the type I error probability of diagnostic testing can be controlled, irrespective of the quality of the specifications of the model with and the one without the extra regressors w_t. This is because its null hypothesis (full validity of the maintained hypotheses in the model without w_t) involves all the assumptions that lead to the test statistic having its known null-distribution. When the test statistic is not significant this does not imply that these maintained hypotheses are valid, but if the diagnostic test statistic is significant this (apart from possible type I errors) clearly demonstrates invalidity of these maintained hypotheses, although it does not convey that the model extended with variables w_t implies valid reformulated maintained hypotheses.

What is the difference with a standard t test in the initial model of testing a null-hypothesis like $\beta = c$ (with c some real number)? The difference is that its usual interpretation when the test statistic has a significant value is simply $\beta \neq c$, whereas as long as the maintained hypotheses have not yet been verified, the interpretation should be that either $\beta \neq c$ and/or elements of the maintained hypotheses do not hold. The latter less comfortable addition may only be swept under the carpet if, prior to performing

tests on β, extensive testing of the maintained hypotheses did not lead to its rejection.

A.6. ON TESTING IN AN UNIDENTIFIED EQUATION

Consider the simple system of two unidentified regression models

$$y_1 = X_1\beta_1 + \varepsilon_1 = y_2\beta_{11} + X_0\beta_{10} + \varepsilon_1,$$
$$y_2 = X_2\beta_2 + \varepsilon_2 = y_1\beta_{21} + X_0\beta_{20} + \varepsilon_2.$$

For $j \in \{1,2\}$ the $n \times 1$ vectors y_j and ε_j contain the observations of the two endogenous variables and the disturbance terms respectively, the $n \times K$ matrices X_j represent the regressor matrices, which have $K-1$ columns in common, namely X_0, and the $K \times 1$ vectors β_j denote the regression coefficients. Using the notation $P_j = X_j(X_j'X_j)^{-1}X_j'$ and $M_j = I - P_j$ for $j \in \{0,1,2\}$ in standard results for partitioned regression, the least-squares estimates for the scalar coefficients β_{11} and β_{21} can be expressed as

$$\hat{\beta}_{11} = y_2'M_0y_1/y_2'M_0y_2 \text{ and } \hat{\beta}_{21} = y_1'M_0y_2/y_1'M_0y_1.$$

For $j \in \{1,2\}$ the residual vectors of the two regressions are given by $e_j = M_jy_j$, their estimated disturbance variances by $s_j^2 = y_j'M_jy_j/(n-K)$ and the estimated variance of their first coefficients by $\widehat{Var}(\hat{\beta}_{11}) = s_1^2/y_2'M_0y_2$ and $\widehat{Var}(\hat{\beta}_{21}) = s_2^2/y_1'M_0y_1$. Now we find for the t-ratio for coefficient β_{11} the expression

$$\frac{\hat{\beta}_{11}}{se(\hat{\beta}_{11})} = \frac{y_2'M_0y_1/y_2'M_0y_2}{s_1/(y_2'M_0y_2)^{1/2}} = \frac{(n-K)^{1/2}y_2'M_0y_1}{(y_1'M_1y_1)^{1/2}(y_2'M_0y_2)^{1/2}}$$

and for that of β_{21}

$$\frac{\hat{\beta}_{21}}{se(\hat{\beta}_{21})} = \frac{(n-K)^{1/2}y_1'M_0y_2}{(y_2'M_2y_2)^{1/2}(y_1'M_0y_1)^{1/2}}.$$

Note that these two expressions have the same numerator. Next we will demonstrate that they also have the same denominator by using for the partitioned regressor matrices a result regarding the projection matrices P_j for $j \in \{1,2\}$, namely $P_1 = P_0 + P_{M_0y_2}$, which yields $M_1 = M_0 - P_{M_0y_2}$, where self-evidently $P_{M_0y_2} = M_0y_2(y_2'M_0y_2)^{-1}y_2'M_0$. Likewise $M_2 = M_0 - P_{M_0y_1}$. Substitution in the squared denominator for the t-ratio for β_{11} yields

$$y_1'M_1y_1y_2'M_0y_2 = y_1'[M_0 - P_{M_0y_2}]y_1y_2'M_0y_2 = y_1'M_0y_1y_2'M_0y_2 - (y_1'M_0y_2)^2$$

and the same is found for the other. Both test statistics are found to be equal to

$$\frac{\hat{\beta}_{j1}}{se(\hat{\beta}_{j1})} = \frac{\hat{\rho}_{12,0}}{[(1-\hat{\rho}^2_{12,0})/(n-K)]^{1/2}}, \text{ where } \hat{\rho}_{12,0} = \frac{y_2' M_0 y_1}{(y_1' M_0 y_1)^{1/2}(y_2' M_0 y_2)^{1/2}}.$$

Both test exactly the same hypothesis, namely whether (under a set of maintained hypotheses) the marginal correlation between the variables y_1 and y_2, after both variables have been cleared from their linear dependence on the variables in X_0, is zero or not. No separate conclusions can be drawn from the two equivalent t-ratio's.

REFERENCES

Aityan, S. G., A. K. Ivanov-Schitz and S.S. Izotov, 2010. Time-shift asymmetric correlation analysis of global stock markets. *Journal of International Financial Markets, Institutions & Money* **20**, 590-605.

Arshanapalli, B., J. Doukas, and L. H. P. Lang, 1995. Pre and post-October 1987 stock market linkages between U.S. and Asian markets. *Pacific-Basin Finance Journal* **3**, 57-73.

Beirne, J., G. M. Caporale, M. Schulze-Ghattas, and N. Spagnolo, 2010. Global and regional spillovers in emerging stock markets: A multivariate garch-in-mean analysis. *Emerging Markets Review* **11**, 250-260.

Burdekin, R. C. K. and P. L. Siklos, 2012. Enter the dragon: Interactions between Chinese, US and Asia-Pacific equity markets, 1995-2010. *Pacific-Basin Finance Journal* **20**, 521-541.

Chen, Z., J. F. Kiviet, and W. Huang, 2015. On the integration of China's main stock exchange with the international financial market. http://egc.sss.ntu.edu.sg/Research/workingpp/Pages/2015.aspx.

Chow, G. C. and C. C. Lawler, 2003. A time series analysis of Shanghai and New York stock price indices. *Annals of Economics and Finance* **4**, 17-35.

Chow, G. C., C. Liu, and L. Niu, 2011. Co-movements of Shanghai and New York stock prices by time-varying regressions. *Journal of Comparative Economics* **39**, 577-583.

Chung, P. J. and D. J. Liu, 1994. Common stochastic trends in pacific rim stock markets. *The Quarterly Review of Economics and Finance* **34**, 241-259.

Edgerton, D. and G. Shukur, 1999. Testing autocorrelation in a system perspective. *Econometric Reviews* **18**, 343-386.

Glick, R. and M. Hutchison, 2013. China's financial linkages with asia and the global financial crisis. *Journal of International Money and Finance***39**, 186-206.

Gosh, A., R. Saidi, and K. H. Johnson, 1999. Who moves the Asia-Pacific stock markets–US or Japan? Empirical evidence based on the theory of cointegration. *The Financial Review* **34**, 159-170.

Groenewold, N., S. H. K. Tang, and Y. Wu, 2004. The dynamic interrelationships between the greater China share markets. *China Economic Review* **15**, 45-62.

Huang, B.-N., C.-W. Yang, and J. W.-S. Hu, 2000. Causality and cointegration of stock markets among the United States, Japan and the South China growth triangle. *International Review of Financial Analysis* **9**, 281-297.

Huyghebaert, N. and L. Wang, 2010. The co-movement of stock markets in East Asia; Did the 1997-1998 Asian financial crisis really strengthen stock market integration? *China Economic Review* **21**, 98-112.

Jayasuriya, S. A., 2011. Stock market correlations between china and its emerging market neighbors. *Emerging Markets Review* **12**, 418-431.

Kenett, D.Y., M. Raddant, T. Lux, and E. Ben-Jacob, 2012. Evolvement of uniformity and volatility in the stressed global financial village. *PLoS ONE* **7(2)**, e31144.

Koundouri, P., N. Kourogenis, and N. Pittis, 2016. Statistical modeling of stock returns: explanatory or descriptive? A historical survey with some methodological reflections. *Journal of Economic Surveys* **30**, 149-164.

Li, H., 2007. International linkages of the Chinese stock exchanges: A multivariate GARCH analysis. *Applied Financial Economics* **17**, 285-297.

Li, H., 2012. The impact of China's stock market reforms on its international stock market linkages. *The Quarterly Review of Economics and Finance* **52**, 358-368.

Singh, P., B. Kumar, and A. Pandey, 2010. Price and volatility spillovers across North American, European and Asian stock markets. *International Review of Financial Analysis* **19**, 55-64.

Wang, L., 2014. Who moves East Asian stock markets? The role of the 2007-2009 global financial crisis. *Journal of International Financial Markets, Institutions & Money* **28**, 182-203.

Yang, J., C. Hsiao, Q. Li, and Z. Wang, 2006. The emerging market crisis and stock market linkages: further evidence. *Journal of Applied Econometrics* **21**, 727-744.

Zhang, B. and X.-M. Li, 2014. Has there been any change in the comovement between the Chinese and US stock markets? *International Review of Economics and Finance* **29**, 525-536.

Zhang, S., I. Paya, and D. Peel, 2009. Linkages between Shanghai and Hong Kong stock indices. *Applied Financial Economics* **19**, 1847-1857.

Zhu, H., Z. Lu, S. Wang, and A. S. Soofi, 2004. Causal linkages among Shanghai, Shenzhen, and Hong Kong stock markets. *International Journal of Theoretical and Applied Finance* **7**, 135-149.