

Response to the Discussants of ‘Deciding between Alternative Approaches in Macroeconomics’

David F. Hendry
Economics Department and Institute for New Economic Thinking
at the Oxford Martin School, University of Oxford, UK

July 2017

1 Introduction

I thank the discussants for their thoughtful analyses and numerous constructive suggestions. Their ideas and clarifications will help advance empirical modelling practice. Professor Proietti elegantly summarises my paper in his introduction, and Professor Perez-Quiros helpfully clarifies the alternative approaches in my paper by ‘walking the reader’ through the various stages of macroeconomic forecasting, noting the prevalence of both theory-driven small dynamic factor models, (see e.g., Stock and Watson, 2002), and data-driven specifications in large scale models as in Forni, Hallin, Lippi, and Reichlin (2000).

Building on an existing body of research reviewed in Hendry and Doornik (2014), the aim of my paper was to draw together at a general level how to decide between approaches, then interested readers could consult more extensive explanations for whatever aspects mattered most to them. There are a number of publications concerning the details, with Monte Carlo simulation studies and theory analyses, albeit that many more remain to be undertaken, as well as applications to macroeconomics and a diverse range of fields including dendrochronology, volcanology and climatology.

To respond to the issues the discussants raised, Section 2 considers the different roles of strategy and tactics in model selection; Section 3 illustrates the combined theory-evidence approach when retaining a theory that transpires not to be a good guide to the finally selected model; and Section 4 briefly turns to the role of forecasting in model selection.

2 Strategy and tactics in model selection

Strategy and tactics both matter in model selection, howsoever that task is undertaken. The former was the focus of my paper, and even the choice of model selection algorithms was treated as part of tactics, so was not discussed at any length. The strategy is one of seeking to nest the local data generation process in a general starting model that retains available theory insights, and includes in an orthogonal form alternatives likely to be relevant in a wide-sense non-stationary world, then selects over the latter by multi-path block searches to discover what additional features matter, thereby evaluating the theory specification. Better tactics will undoubtedly evolve over time, and have done so already: compare model selection approaches before and after the introduction of indicator saturation for finding breaks.

Indicator saturation allows the design of indicators to match ‘likely’ break forms, providing a useful flexibility, as in impulse-indicator saturation (IIS) for outliers, step-indicator saturation (SIS) for location shifts, a \mathcal{V} shape for the impacts on temperature of volcanic eruptions as in Pretis, Schneider, Smerdon, and Hendry (2016), etc. Based on the results in Hendry and Krolzig (2005), collinearity is less problematic for SIS than it looked at first sight, encouraging us to investigate the properties of trend-indicator

saturation. Moreover, the uncertainty around SIS-determined break dates can be estimated, as in Hendry and Pretis (2016), as can uncertainty bands for the trajectory of the mean: see Pretis (2015).

Although we have not yet tackled selecting shifts in second moments, that remains on the research agenda, and model selection of non-linearities goes some way towards that aim. Basis functions for approximating non-linearity are again an aspect of tactics, where future improvements are highly likely, and Professor Proietti's proposal of using B-splines for both shifts and non-linearity points a way ahead.

What Ericsson (2017) calls multiple-indicator saturation (MIS)—where every regressor is interacted with a saturating set of step indicators, so for k regressors and T observations, there are $k \times T$ candidate variables—now enables the detection of changes in the parameters of regressors. Kitov and Tabor (2015) demonstrated the success of this approach, despite the high dimensionality of the set of candidate variables. To understand intuitively how SIS or MIS are able to detect location shifts or parameter changes, consider knowing where a single (moderately large magnitude) shift occurred, and splitting your data into sub-samples before and after that point. Then you would rightly be surprised if fitting your correctly specified model of the data generation process (DGP) separately to the different sub-samples did not deliver appropriately different estimates of their DGP parameters. Choosing that split by SIS or MIS will add variability from particular error draws around the break point, which may offset the apparent shift date temporarily by making it appear slightly before or after the actual occurrence, but the correct indicator, or one close to it, will accomplish almost the same task as knowing the timing, so will be the likely selection.

A surprising finding from our research is that model selection, appropriately conducted in a setting where the general unrestricted model (GUM) is sufficiently well specified to nest the DGP, is almost as good as selecting from that DGP at the same significance level. Of course, nesting the DGP, or the local DGP, is unlikely and the entailed LDGP may be a poor approximation to the actual DGP. Moreover, GUMs may be underspecified in many possible ways. Nevertheless, Castle, Doornik, and Hendry (2011) show in simulations that the costs from estimating the DGP can exceed those from selecting from an underspecified GUM, as measured by the RMSEs of coefficient estimates relative to the DGP parameters. Castle and Hendry (2014) also investigate the consequences for automatic model selection facing shifts when using indicator saturation, reinforcing the advantages of seeking to include all likely substantively-relevant variables.

3 Is it problematic to retain a poor theory?

An underspecified set of variables is most likely to arise when only a theory model is specified and estimated, eschewing the advice in the paper to retain that theory while searching over a much larger set of candidates. To illustrate the combined approach in that setting, I return to the Davidson, Hendry, Srba, and Yeo (1978) (DHSY) study on quarterly UK data for constant price consumers' expenditure, c_t and real personal disposable income, i_t , over 1958(2)–1976(2). Starting from the simplest version of the permanent income hypothesis (PIH) current at the time of their study, namely $c_t = \beta_0 + \beta_1 i_t + \beta_2 c_{t-1} + e_t$, with added seasonal dummies, S_i , (equation (12) in DHSY), had *Autometrics* been available, could DHSY have found their model in an afternoon rather than several years, despite the initial theory being a poor guide to their finally selected model?

Re-estimating DHSY's PIH equation, but in a log-linear specification over the full sample yielded:

$$\begin{aligned}
 c_t &= \underset{(0.07)}{0.59} c_{t-1} + \underset{(0.054)}{0.31} i_t + \underset{(0.14)}{0.87} - \underset{(0.007)}{0.12} S_{1,t} - \underset{(0.005)}{0.01} S_{2,t} - \underset{(0.003)}{0.03} S_{3,t} \quad (1) \\
 \hat{\sigma} &= 1.0\% \quad R^2 = 0.995 \quad F_{ar}(5, 66) = 9.68^{**} \quad F_{arch}(4, 69) = 2.79^* \\
 \chi_{nd}^2(2) &= 5.14 \quad F_{het}(7, 69) = 3.97^{**} \quad F_{reset}(2, 69) = 0.57
 \end{aligned}$$

where estimated coefficient standard errors are shown in parentheses below estimated coefficients, $\hat{\sigma}$ is the residual standard deviation, R^2 is the coefficient of multiple correlation, F_{ar} is a test for residual autocorrelation (see Godfrey, 1978), F_{arch} tests for autoregressive conditional heteroskedasticity (see Engle, 1982), F_{het} is a test for residual heteroskedasticity (see White, 1980), $\chi_{nd}^2(2)$ is a test for non-Normality (see Doornik and Hansen, 2008), and F_{reset} is the RESET test (see Ramsey, 1969). Thus, tests for residual autocorrelation, ARCH and heteroskedasticity all reject: see Figure 2(a) for a graph of the resulting residuals.

The main result is the failure of that simple theory to characterize the evidence, so evaluation is accomplished, but there is no useful guidance on how to proceed towards a better formulation. Recipes for patching autocorrelation, heteroskedasticity, etc., do not take into account that rejection on such tests can arise from many failures of the assumptions needed for congruence, not necessarily the alternative hypothesis against which the test was designed to have power: see Mizon (1995) and Spanos (2017).

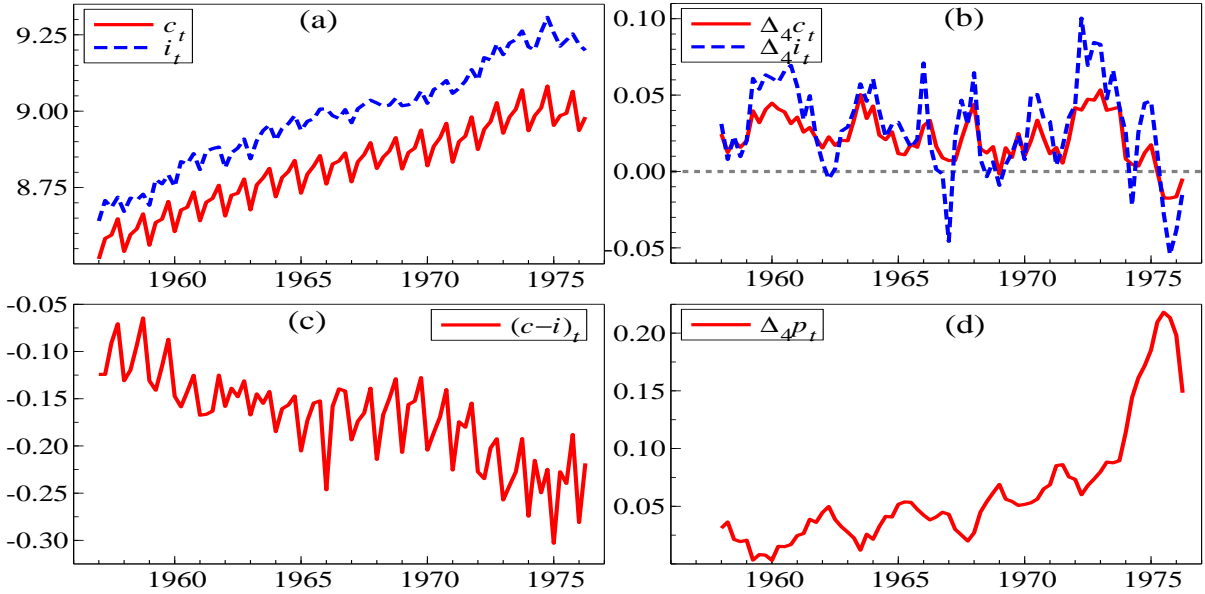


Figure 1: DHSY time series: (a) c_t and i_t ; (b) $\Delta_4 c_t$ and $\Delta_4 i_t$; (c) $(c - i)_t$; (d) $\Delta_4 p_t$.

The DHSY data on c_t, i_t shown in Figure 1(a) suggests that seasonal lags might matter (here 5 lags), and also including inflation, $\Delta_4 p_t$ and its lag, and their tax dummy, $\Delta_4 D_t$, creates the GUM. The continuous variables were orthogonalized against c_{t-1} and i_t , denoted $\tilde{\cdot}$. Next, this GUM was estimated to check that the coefficient estimates of the retained variables in (1) were unaffected, as (2) confirms, where bold denotes coefficients of added variables that are individually significant at 1%.

$$\begin{aligned}
c_t = & \begin{matrix} 0.59 \\ (0.04) \end{matrix} c_{t-1} + \begin{matrix} 0.31 \\ (0.03) \end{matrix} i_t + \begin{matrix} 0.88 \\ (0.09) \end{matrix} - \begin{matrix} 0.12 \\ (0.004) \end{matrix} S_{1,t} - \begin{matrix} 0.01 \\ (0.003) \end{matrix} S_{2,t} - \begin{matrix} 0.03 \\ (0.002) \end{matrix} S_{3,t} \\
& + \mathbf{0.007} \Delta_4 D_t - \begin{matrix} 0.04 \\ (0.09) \end{matrix} \tilde{c}_{t-2} + \begin{matrix} 0.05 \\ (0.09) \end{matrix} \tilde{c}_{t-3} + \mathbf{0.73} \tilde{c}_{t-4} - \begin{matrix} 0.02 \\ (0.12) \end{matrix} \tilde{c}_{t-5} + \mathbf{0.16} \tilde{i}_{t-1} \\
& - \begin{matrix} 0.031 \\ (0.05) \end{matrix} \tilde{i}_{t-2} + \begin{matrix} 0.04 \\ (0.05) \end{matrix} \tilde{i}_{t-3} - \begin{matrix} 0.10 \\ (0.05) \end{matrix} \tilde{i}_{t-4} - \mathbf{0.20} \tilde{i}_{t-5} - \mathbf{0.33} \widetilde{\Delta_4 p_t} + \begin{matrix} 0.19 \\ (0.09) \end{matrix} \widetilde{\Delta_4 p_{t-1}} \\
\hat{\sigma} = & 0.58\% \quad R^2 = 0.998 \quad F_{ar}(5, 46) = 1.56 \quad F_{arch}(4, 61) = 2.98^* \\
\chi_{nd}^2(2) = & 0.08 \quad F_{het}(31, 37) = 1.47 \quad F_{reset}(2, 49) = 4.92^* \quad F_{add}(12, 51) = 13.1^{**}
\end{aligned} \tag{2}$$

The fit is improved, and the F_{add} -test on the joint significance of the added orthogonal variables strongly rejects the completeness and correctness of the initial theory: Figure 2(b) plots the resulting residuals.

Next, selection from the additional variables in (2) was undertaken at $\alpha = 1\%$, retaining the variables in (1) without selection. The sample size is $T = 73$ and the total number of additional variables is $N = 12$, so a significance level of 1% entails that on average only **one** variable will be selected by chance roughly every **8** times under the null that the additional variables are irrelevant. Nevertheless, $\Delta_4 D_t$ and the orthogonal components \tilde{c}_{t-4} , \tilde{i}_{t-1} , \tilde{i}_{t-5} , and $\widetilde{\Delta_4 p_t}$ were selected, with a joint significance of $F_{\text{add}}(4, 58) = 31.3^{**}$, which strongly rejects their irrelevance at any sensible significance level.

Given the theory rejection, it is natural to replace the orthogonalized variables by their original measures to formulate an untransformed GUM, and select over all the candidates, including the theory variables (estimation of the levels formulation is justified by Sims, Stock, and Watson, 1990). This yielded:

$$\begin{aligned}
 c_t = & \underset{(0.03)}{0.92} c_{t-4} + \underset{(0.04)}{0.27} i_t + \underset{(0.04)}{0.19} i_{t-1} - \underset{(0.05)}{0.14} i_{t-4} - \underset{(0.04)}{0.24} i_{t-5} \\
 & - \underset{(0.08)}{0.34} \Delta_4 p_t + \underset{(0.08)}{0.21} \Delta_4 p_{t-1} + \underset{(0.002)}{0.007} \Delta_4 D_t \\
 \hat{\sigma} = & 0.58\% \quad F_{\text{ar}}(5, 56) = 0.58 \quad F_{\text{arch}}(4, 61) = 1.49 \\
 \chi_{\text{nd}}^2(2) = & 0.18 \quad F_{\text{het}}(16, 52) = 0.84 \quad F_{\text{reset}}(2, 59) = 3.8^*
 \end{aligned} \tag{3}$$

where only the RESET mis-specification test marginally rejects: see Figure 2(c).¹

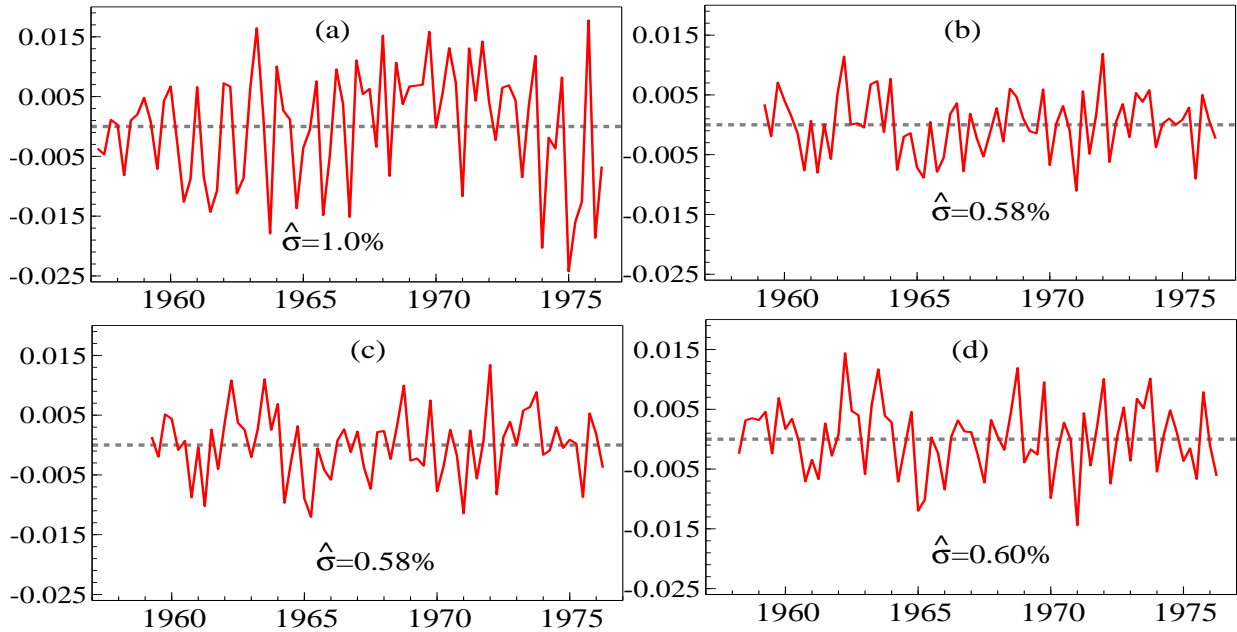


Figure 2: Residuals from the four models of UK consumers' expenditure: (a) the PIH formulation (1); (b) the orthogonalized (and original) variables GUM (2); (c) the selected level's representation (3); (d) the selected approximation to DHSY (4).

To effect a reduction to a non-integrated representation, I transformed the GUM in the original measures to seasonal differences and the differential $(c - i)_{t-4}$, selecting at 1% not retaining any variables.

¹This result differs from <http://voxeu.org/article/improved-approach-empirical-modelling-0>, which commenced with 5 lags on inflation.

This yielded the same result as transforming (3):

$$\begin{aligned}
\Delta_4 c_t &= \underset{(0.03)}{0.48} \Delta_4 i_t - \underset{(0.04)}{0.22} \Delta \Delta_4 i_t - \underset{(0.02)}{0.13} \Delta_4 p_t - \underset{(0.08)}{0.22} \Delta \Delta_4 p_t \\
&\quad - \underset{(0.011)}{0.095} (c - i)_{t-4} + \underset{(0.002)}{0.007} \Delta_4 D_t \\
\hat{\sigma} &= 0.60\% \quad F_{\text{ar}}(5, 62) = 0.36 \quad F_{\text{arch}}(4, 65) = 2.12 \\
\chi_{\text{nd}}^2(2) &= 0.20 \quad F_{\text{het}}(12, 60) = 0.75 \quad F_{\text{reset}}(2, 65) = 0.75
\end{aligned} \tag{4}$$

Now no mis-specification tests reject, and Figure 2(d) reports the final set of residuals. This equation essentially recovers the DHSY model in a few hours, showing the advances in econometrics and technology since the 1970s. While the general idea of combining theory retention with variable selection in Hendry and Johansen (2015) could have been proposed 50 years ago, automatic software to implement it is much more recent. Despite the near irrelevance of the initially retained theory, the combined approach has led to an improved congruent and interpretable model, and reveals the dangers of relying on the ‘significance’ of coefficients and the fit of an imposed model.

4 The role of forecasting in model selection

My paper questioned assigning a major role to forecast accuracy when choosing between empirical models. The estimated in-sample DGP need not be consistently ‘best’ on most forecasting criteria, especially when there are unanticipated shifts. While there is merit in a shifting world to emphasising the most recent data, a pseudo-forecast sample is often small relative to the whole available data, so idiosyncratic influences (such as measurement errors) can play a detrimental role. Nevertheless, if forecasting is the primary objective of an empirical exercise, then forecast accuracy (on some measure) such as the accuracy of interval forecasts, or forecast distributions, must be the main criteria for model selection. That said, there are difficulties in measuring how ‘accurate’ forecasts are, addressed in Clements and Hendry (1993), with a recent ‘direction free’ proposal in Hendry and Martinez (2017) for multi-step system forecasts despite relatively few forecast-error observations.

Allen and Fildes (2005) question the role of mis-specification tests in choosing forecasting models, but how selection is implemented, and what decision criteria are used, can differ greatly between just forecasting from those used for forecasting in combination with economic policy by a governmental agency: both forecast accuracy and valid policy advice matter. Evaluating a model by its predictions—by making statements outside its constructed remit, as in Eddington’s confirming that gravity bent light waves by a magnitude derived from Einstein’s theory of general relativity—is distinct from seeking to test a model by its forecasts of out-of-sample events. As Spanos (2007) showed, the goodness of fit (and e.g., forecasts of lunar eclipses) by the Ptolemaic and Copernican models of planetary orbits are similar, despite their clashing formulations, primarily because the solar system has a near-stationary distribution. Nonetheless, Spanos shows the former can be rejected as there are *systematic* residuals.

As with selection, the tactics of handling forecast failure after location shifts can surely be improved, offsetting the slow progress in forecasting shifts. I agree with Professor Proietti that variants of robust methods could well do better, including averaging over ‘non-poisonous’ devices to reflect the extent of failure as it evolves, and using adaptive models to nest both shifts and their absence. Research on appropriate learning mechanisms combined with rapid updating merit consideration, as robustification is at most helpful for short horizons when the causes of unanticipated shifts are not understood.

In the context of the data available, such as big data in factor approaches, and principal components or basis functions to approximate non-linearities, summaries of what information might be relevant are

often needed. When 9000 variables are ‘potentially relevant’, prior reduction is essential: the tight selection criterion of 0.1% we use for SIS would still retain 9 adventitiously significant variables. Thus, even within the strategy of commencing from an extremely general initial specification, the selection tactics can depend greatly on the context.

References

- Allen, P. G. and R. A. Fildes (2005). Levels, differences and ECMs—principles for improved econometric forecasting. *Oxford Bulletin of Economics and Statistics* 67, 881–904.
- Castle, J. L., J. A. Doornik, and D. F. Hendry (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics* 3 (1), DOI: 10.2202/1941–1928.1097.
- Castle, J. L. and D. F. Hendry (2014). Model selection in under-specified equations with breaks. *Journal of Econometrics* 178, 286–293.
- Clements, M. P. and D. F. Hendry (1993). On the limitations of comparing mean squared forecast errors (with discussion). *Journal of Forecasting* 12, 617–637.
- Davidson, J. E. H., D. F. Hendry, F. Srba, and J. S. Yeo (1978). Econometric modelling of the aggregate time-series relationship between consumers’ expenditure and income in the United Kingdom. *Economic Journal* 88, 661–692.
- Doornik, J. A. and H. Hansen (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* 70, 927–939.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.
- Ericsson, N. R. (2017). How biased are U.S. Government forecasts of the Federal Debt? *International Journal of Forecasting* 33, 543–559.
- Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized factor model: Identification and estimation. *Review of Economics and Statistics* 82, 540–554.
- Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica* 46, 1303–1313.
- Hendry, D. F. and J. A. Doornik (2014). *Empirical Model Discovery and Theory Evaluation*. Cambridge, Mass.: MIT Press.
- Hendry, D. F. and S. Johansen (2015). Model discovery and Trygve Haavelmo’s legacy. *Econometric Theory* 31, 93–114.
- Hendry, D. F. and H.-M. Krolzig (2005). The properties of automatic Gets modelling. *Economic Journal* 115, C32–C61.
- Hendry, D. F. and A. B. Martinez (2017). Evaluating multi-step system forecasts with relatively few forecast-error observations. *International Journal of Forecasting* 33, 359–372.
- Hendry, D. F. and F. Pretis (2016). Quantifying the uncertainty around break dates in models using indicator saturation. Working paper, Economics Department, Oxford University.
- Kitov, O. I. and M. N. Tabor (2015). Detecting structural changes in linear models: A variable selection approach using multiplicative indicator saturation. Unpublished paper, University of Oxford.
- Mizon, G. E. (1995). A simple message for autocorrelation correctors: Don’t. *Journal of Econometrics* 69, 267–288.

- Pretis, F. (2015). Testing for time-varying predictive accuracy using bias-corrected indicator saturation. Working paper, Economics Department, Oxford University.
- Pretis, F., L. Schneider, J. E. Smerdon, and D. F. Hendry (2016). Detecting volcanic eruptions in temperature reconstructions by designed break-indicator saturation. *Journal of Economic Surveys* 30, 403–429.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B* 31, 350–371.
- Sims, C. A., J. H. Stock, and M. W. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica* 58, 113–144.
- Spanos, A. (2007). Curve-fitting, the reliability of inductive inference and the error-statistical approach. *Philosophy of Science* 74, 1046–1066.
- Spanos, A. (2017). Mis-specification testing in retrospect. *Journal of Economic Surveys*, DOI: 10.1111/joes.12200.
- Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indices. *Journal of Business and Economic Statistics* 20, 147–162.
- White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–838.