EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 10, No. 1, 2017, 133-156
ISSN 1307-5543 – www.ejpam.com
Published by New York Business Global

Sir Clive W.J. Granger Memorial Special Issue on Econometrics

Sir Clive W.J. Granger Model Selection

Jennifer L. Castle

Magdalen College and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, UK.

Abstract. Clive Granger proposed *thick* modelling as an alternative to selecting a unique model based on a given criterion, or *thin* modelling. This stemmed from his research on forecast combination and portfolio selection in which using just the best asset or forecast can be suboptimal in many settings. This paper proposes to integrate thick modelling into the general-to-specific model selection literature, yielding the benefits of selecting a set of well-specified encompassing models while taking seriously Granger's critique of model selection. The paper argues that model uncertainty is addressed by applying selection to narrow down the class of models followed by pooling across the retained set of close specifications. An example using artificial data illustrates the approach.

2010 Mathematics Subject Classifications: 62F07, 62P20, 91B84
Key Words and Phrases: Clive Granger, model selection, thick modelling, model averaging.

1. Introduction

Economic modelling requires a synthesis of theory and empirics, typically with varying weights placed on each depending on the researcher's beliefs or preferences. Granger was clear that one of the main objectives of empirical modelling was to inform the decision making process. He was relatively agnostic about the preferred approach to model building as long as the models were carefully evaluated and validated using actual data, and the models were useful in addressing the research question asked, specifically in terms of the quality of decisions that are made based on the models. Model selection is inevitable in this framework. Imposing theory models with no testing would fail Granger's criteria, and any other form of model building must necessitate model selection. In Granger's Marshall Lectures [24] he explains how he sees model selection:

Email address: jennifer.castle@magd.ox.ac.uk

> A sculptor once said that the way he viewed his art was that he took a large block of stone and just chipped it away until he revealed the sculpture that was hidden inside it. Some empirical modelers view their task in a similar fashion starting with a mass of data and slowly discarding them to get at a correct representation. My perception is quite different. I think of a modeler as starting with some disparate pieces some wood, a few bricks, some nails, and so forth - and attempting to build an object for which he (or she) has only a very inadequate plan, or theory. The modeler can look at related constructs and can use institutional information and will eventually arrive at an approximation of the object that they are trying to represent, perhaps after several attempts. Model building will be a team effort with inputs from theorists, econometricians, local statisticians familiar with the data, and economists aware of local facts or relevant institutional constraints.

The above quote can be misinterpreted by those who are not familiar with the general-to-specific (*Gets*) model selection literature and perceive the approach to consist of *'starting with a mass of data and slowly discarding them to get at a correct representation'*. In fact, the approach is much more in line with Granger's view of model building, whereby theory, past evidence and institutional knowledge all inform the initial specification of the general unrestricted model (GUM) but rigorous evaluation and validation is conducted to reduce the model by eliminating those effects that are statistically irrelevant. The important distinction is in the specification of the GUM – without careful thought informed by the objective of the modelling exercise, theory and institutional knowledge, and an understanding of the data and its limitations, one is in danger of data mining. The edited volume by Granger (1990) (see [23]) provides a collection of critiques.

Granger's interest in model selection came later in his research career, see Hendry (2010) [37]. He showed remarkable foresight in Granger, King and White (1995) (see [31]) where he argues that model selection procedures were preferable to hypothesis testing when specifying empirical models, thereby anticipating developments in *Gets* modelling such as Hoover and Perez (1999) [51] and Hendry and Krolzig (1999) [42]. The general-to-specific literature, often termed the LSE approach due to its proponents at the London School of Economics in the 1960s and 1970s had been developing for some time, see, e.g. Pagan (1987), Phillips (1988), Mizon (1995) and Hendry (2003) ([60],[63],[58],[35]), but were only automated later. Granger, in an interview with David Hendry explored the *Gets* approach in detail in Granger and Hendry (2005) (see [29]). Although Granger's contributions to the field of model selection were not as prolific as his contributions to other fields of econometrics demonstrated by the papers in this special issue, he clearly had an interest in model selection as seen in Granger and Hendry (2005) [29]. One of the contributions of this paper is to consider how model selection performs under misspecification, which was a question that Granger highlighted in [29].

Granger coined the phrase 'thick modelling' to argue that model selection need not focus on selecting just one model specification from the infinite range of possibilities, which would be defined as the 'best' model based on some pre-specified criterion. He proposed keeping all alternative close specifications that are validated on empirical criteria and pool the resulting outcomes, be they parameter estimates, impulse responses, policy simulations, hypothesis tests or forecasts, see Granger and Jeon (2004) and Granger (2005) [30, 26]. Given Granger's interest in portfolio theory (Granger, 2005, [27]) this is a natural thing to do, and corresponds to his pragmatic approach to empirical modelling discussed in Granger (2009) (see [28]). It also ties with his research on forecasting and forecast combinations stemming from Bates and Granger (1969) (see [5]).

The close correspondence between the views on model selection of Granger's thick modelling and Hendry's *Gets* modelling, and their many personal and professional interactions on the subject, suggest that the two approaches can be integrated. This paper investigates the approach to model selection of undertaking general-to-specific selection and then applying thick modelling techniques of averaging across the selected models, where one or more models can be retained. We find that thick modelling does not resolve problems of misspecification, but can capture model uncertainty reflected by retaining different model specifications commencing from the same starting point due to searching many reduction paths. Thick modelling across well-specified models captures the relevant model uncertainty implied by model selection. Section 2 outlines the selection approach implemented by *Autometrics* and discusses how thick modelling can be implemented, section 3 notes the literature in which thick modelling is employed, section 4 demonstrates the approach with an artificial data series and section 5 concludes.

2. Model selection, model averaging and thick modelling

Commencing from an information set, model selection can be interpreted as pooling the available information set and selecting possible model specifications therefrom. Thick modelling will retain a set of specifications and thin modelling will retain a single specification. Model averaging pools the model specifications resulting from subsets of information. The distinction between pooling information and pooling models has implications for congruency, which will now be discussed.

2.1. Model selection

Granger emphasized the importance of establishing the purpose of the econometric modelling exercise. Let us assume that the objective is to identify a model(s) that matches the unknown Data Generation Process (DGP) in all measured aspects. This introduces the notion of congruence which relates to the model being statistically well-specified, see e.g., Bontemps and Mizon (2003) [6]. Congruence requires that the empirical model is statistically 'close' to the evidence. The Theory of Reduction establishes the properties of the unknown local data generating process (LDGP) which delivers a mean-zero, homoskedastic innovation process that the model aims to capture. Hendry (1995, ch.9) (see [34]) provides extensive discussion.[*] A further step in the *Gets* literature is that the model

[*]The Theory of Reduction commences from an unmanageably large DGP denoted by $\mathsf{D}_{\mathbf{u}}\left(\mathbf{U}_T^1|\mathbf{U}_0,\psi_T^1\right)$, with $\psi_T^1 \in \mathbf{\Psi} \subseteq \mathbb{R}^{kT}$, where $\mathbf{U}_T^1 = (\mathbf{u}_1,\ldots,\mathbf{u}_T)$ is the full sample vector of random variables defined on probability space $(\mathbf{\Omega},\mathcal{F},\mathsf{P})$. A series of reduction steps are applied to the DGP, including sequential

should encompass every other model that is a valid restriction of the general unrestricted model. Encompassing refers to the ability of a model to explain the results of rival models and hence make them redundant, enabling selection between two or more congruent models. Mutually encompassing models are able to explain the results of each other's model, so the models cannot be ranked. The LDGP will be congruent and will encompass all other models, so models that are non-congruent or non-encompassing fail to model the LDGP. Encompassing has been extensively discussed in, *inter alia* Mizon and Richard (1986) [59], Hendry and Richard (1989) [45], Mizon (1994) [57] and Hendry (1995, ch.14) [34].

The most recent generation of automated *Gets* software, *Autometrics* (see Doornik, 2009, [19] and Hendry and Doornik, 2014, [39]), uses a tree search to eliminate irrelevant variables, with various methods for speeding up the full multipath search. Diagnostic checks are used to ensure the resulting terminal models are congruent. Encompassing tests are implemented to ensure the terminal models encompass the GUM so there is no significant loss of information by undertaking the reduction. They are also performed against the union of the terminal models. This may result in a single unique terminal model being retained, but in general there may be multiple terminal models, all of which are valid reductions of the GUM and mutually encompass each other. In practice, the automated algorithm selects one model based on the Schwarz (1978) (see [64]) information criterion, but a range of other methods to select between congruent encompassing models could be introduced at this stage, or the user may have a preference for a particular terminal model on theory or aesthetic grounds.

One of the criticisms of such an approach is that there is no accounting for 'model' or 'specification' uncertainty. It is argued that the standard errors of the selected terminal model capture the estimation uncertainty due to sampling, but not the uncertainty of the choice of model selected. In response to this criticism, searching many paths increases the probability of locating the LDGP. However, thick modelling provides a route to capturing model uncertainty due to selection, while retaining the benefits of selection. Model selection is designed to handle more complex problems than just locating relevant variables, by jointly including variables, dynamics, outliers and structural breaks, non-stationarities and non-linearities. Castle, Doornik and Hendry (2011) (see [9]) provide simulation evidence establishing that *Autometrics* is able to recover the LDGP almost as often as when commencing from the LDGP itself, and hence the costs of searching over a more general initial specification are small relative to the costs of inference directly conducted on the

factorization to obtain an innovation process, marginalization and conditional factorization (to discard the marginal distribution if weak exogeneity is satisfied), and further transformations including lag truncation, functional form, constancy, etc, to result in an LDGP given by $\mathbf{A}(\mathbf{L})\mathbf{g}(\mathbf{y}_t) = \mathbf{B}(\mathbf{L})\mathbf{h}(\mathbf{z}_t) + \epsilon_t$ where $\epsilon_t \underset{app}{\sim} \mathsf{N}_n(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$ where ϵ_t is a mean-zero homoskedastic innovation process with variance $\mathbf{\Sigma}_\epsilon$ and $\mathbf{A}(\mathbf{L})$ and $\mathbf{B}(\mathbf{L})$ are lag polynomials.

. The properties of the LDGP are determined by the reductions applied, so by ensuring there is no loss of information in the reduction process we face the null hypothesis that the LDGP has homoskedastic near normal innovation errors, with weakly exogenous conditioning variables and constant parameters and encompasses the DGP, see Hendry (2009) [36] and Hendry and Doornik (2014, ch.6) [39]. So despite not knowing the DGP or LDGP specification, we can establish the properties that the LDGP should possess.

LDGP.

Model selection does incur costs—relevant variables can be omitted and irrelevant variables can be incorrectly retained—but these costs are quantifiable using measures of potency and gauge. Potency, which calculates how likely relevant variables are retained (the average non-null rejection frequency for the null hypothesis that a coefficient is equal to zero), should be close to the power of the corresponding test in the LDGP. The nominal null rejection frequency chosen for selection, i.e. the significance level for hypothesis tests of the kind $\mathsf{H}_0 : \beta_j = 0$, should match the retention of irrelevant variables (the empirical null rejection frequency), which is called the gauge. These measures refute the criticism by Leamer (1983) (see [55]) that the model selection process is often hidden when the final model is reported. If there are many irrelevant variables, then uncertainty appears to be large as different irrelevant variables are likely to be selected on different draws by chance sampling, commencing from the same information set, but those irrelevant variables will have little impact given that their retention rate is controlled by the nominal significance level. The important aspect of selection is to retain the same set of relevant variables in different draws commencing from the same information set, which will be a function of the non-centralities (population $\mathsf{t}-$statistics) of the relevant variables and the nominal significance level. Forcing the 'theory-relevant' variables and implementing selection over only the additional candidate variables will control this cost, see Hendry and Johansen (2015) [40].

2.2. Model Averaging

Model averaging is an alternative to model selection that claims to address model uncertainty. By averaging over 2^k possible models within the given model class, where k is the number of potential variables, it is thought that the uncertainty over which model is correct is captured. See Hoeting, Madigan, Raftery and Volinsky (1999) [50] for a Bayesian interpretation and Hansen (2008) [33] for a classical motivation within forecasting. There is a huge literature discussing the appropriate weights for model averaging depending on the purpose of the modelling exercise, for example Akaike (1973), Schwarz (1978) or Hannan and Quinn (1979) (see [2][64][32]) information criteria for in-sample averaging or out-of-sample mean square forecast error criterion if the purpose of the exercise is to forecast.

Model averaging is most commonly used in forecasting, where combining forecasts can outperform the individual forecasts if there are offsetting biases or breaks, see Hendry and Clements (2004) [38]. Model or forecast averaging can also help if diversification provides a reduction in variance, see Granger (2002) [25]. However, if proponents of model averaging do wish to account for model uncertainty, then all possible models must be considered rather than a reduced set based on some model selection criterion, unless this selection uncertainty is also accounted for in measuring model uncertainty, which is not typically done in the literature. Many procedures and algorithms have been proposed to narrow down the search space, including Occams razor, branch and bound techniques (see Hocking and Leslie, 1967, [47], LaMotte and Hocking, 1970, [54], and Gatu and Kontoghiorghes, 2006, [21]), ridge regression (see Hoerl and Kennard, 1970, [49, 48]), the non-negative garrote (see Breiman, 1995, [7]) and the least absolute shrinkage and selection operator (LASSO: see Tibshirani, 1996, [65]). The cost of averaging over all possible models is that a large subset will necessarily be mis-specified and could contaminate the averaged model. Moreover, the average calculated from different data draws could still differ greatly when regressors are highly collinear and the information set over which models are averaged is not a good representation of the DGP.

2.3. Thick Modelling

Granger's thick modelling can be viewed as an attempt to reconcile the distinct methodologies of model selection and model averaging. The argument that justifies model averaging, namely accounting for model uncertainty, is fallacious—averaging over all possible models of which many must necessarily be non-congruent and therefore lead to biased coefficient estimates cannot be a sensible way to proceed, even if weights are rebased and intercepts are forced, see Hendry and Reade (2008) [44]. Furthermore, model averaging does not help if there are structural breaks in the data, whereas *Gets* selection seeks well-specified encompassing models of the LDGP and in doing so handles possible structural breaks, which is seen as more important than determining the uncertainty in a badly specified representation. However, discarding potentially useful information in ignored specifications because only one model is selected can also lead to poor outcomes, be they parameter estimates or forecasts. Thick modelling is an approach that keeps all close alternative specifications and pools across these. In the *Gets* terminology, the retained terminal models are all congruent encompassing models of the LDGP, so pooling across these models will give a thick representation of the LDGP.

We can summarise the approaches as the choice of weights on the model combination. If there are k potential variables, there are 2^k combinations of subsets of variables, including the null and full set, resulting in 2^k possible models. Let $\mathcal{M}$ denote the full set of models and $\mathcal{M}_s$ the selected set of models, where $\mathcal{M}_s \subseteq \mathcal{M}$. Given a criterion c_m, the weights are given by:

$$w_m = \begin{cases} \frac{c_m}{\sum_{j \in \mathcal{M}_s} c_j} & m \in \mathcal{M}_s \\ 0 & m \notin \mathcal{M}_s \end{cases} \tag{1}$$

allowing us to characterize:

1. Model selection (thin modelling in Granger's terminology): $\mathcal{M}_s$ has a single member so $w_m = 1$ for the selected model and 0 otherwise.

2. Unweighted model averaging: $\mathcal{M}_s = \mathcal{M}$ and $c_j = 1, \forall j$, resulting in $w_m = 2^{-k}$.

3. Thick modelling: assuming L models within $\mathcal{M}_s$, equal weighted averaging over $\mathcal{M}_s$ with $w_m = L^{-1}$.

The criterion for *Autometrics* is the nominal significance level α which is set by the user. This implies the set of selected models is a function of α, $\mathcal{M}_s(\alpha)$, and varying α will

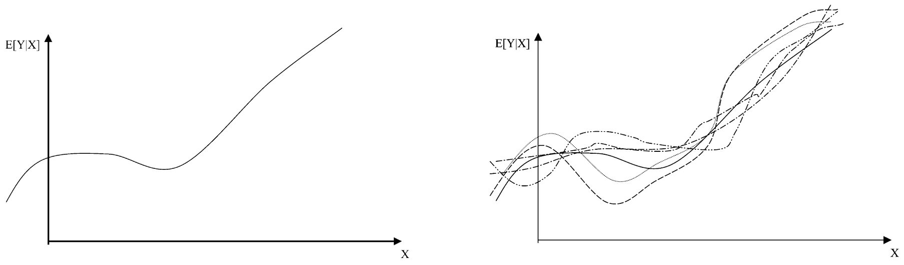

Fig. 1. Thin modeling.

Fig. 2. Thick modeling.

Figure 1: Figures from Granger and Jeon (2004) describing thin and thick modelling

vary the number of selected models. When $\alpha = 1$ the GUM will be retained, so $\mathcal{M}_s$ will have a single member although no selection has been applied. A tight α will likely lead to a smaller subset of congruent mutually encompassing models retained, so fewer models to average over for the thick modelling.

Figure 1, taken from Granger and Jeon (2004) (see [30]), captures the principle of thick modelling. Although Granger did not frame his proposal of thick modelling within *Gets* selection, the benefits of doing so are twofold. First, *Gets* aims to select congruent encompassing specifications. Models that are discarded fail this criteria, either because there are irrelevant variables in the model, the model fails to encompass the GUM so information is lost, or the model fails diagnostic tests so cannot be a close approximation to the LDGP whose properties are known even though its specification is unknown. Therefore, averaging over the retained set ensures a well-specified thick model. Second, the intention of thick modelling is to consider alternative specifications of similar quality, which the terminal models after *Gets* selection must satisfy to be mutually encompassing. The final stage of choosing one terminal model using some pre-specified criteria can be arbitrary and thick modelling will avoid this reliance on the chosen criterion.

The advantage of applying the procedure within the *Gets* framework is that most uncertainty between model specifications will be captured by the range of terminal models retained. A variable with a high non-centrality will most likely be retained in all specifications so the degree of model uncertainty is small, but distinguishing between, say, lag 1 and lag 2 of a persistent variable may be much more difficult, resulting in terminal models that retain one or other of the lag length specifications. The information content of both specifications is close, and therefore averaging across both lags of the variable would be a sensible way to capture the timing uncertainty. A similar consideration applies to variables that have small non-centralities in the LDGP, where a given draw may result in the relevant variable being statistically insignificant, and vice versa, irrelevant variables may chance to have significant selection statistics in the particular data draw. This more closely reflects model uncertainty in the robust model specification sense, where alternative draws for the innovation could result in slightly different model specifications selected.

Factors affecting the selected set $\mathcal{M}_s$ for thick modelling will include the degree of correlation between regressors and possible misspecification of the GUM, as well as α.

High correlations between candidate variables contaminate the ranking of variables based on t-statistics. Assume a GUM which is orthogonal in the sample:

$$y_t = \sum_{k=1}^{N} \beta_k x_{k,t} + \epsilon_t \tag{2}$$

where $T^{-1}\sum_{t=1}^{T} x_{k,t}x_{j,t} = \lambda_k \delta_{k,j}\ \forall k,j$, where $\delta_{k,j} = 1$ if $k = j$ and zero otherwise, with $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$, independently of the $\{x_{k,t}\}$, and $T >> N$. The DGP is nested within the GUM, where $n \leq N$ of the regressors have non-zero β_k. In this setting, orthogonality enables 1-cut selection, in which the N sample t^2-statistics testing H_0: $\beta_k = 0$ are ordered from largest to smallest and variables are retained if, and only if, their associated t^2-statistic is greater than the critical value for a chosen significance level c_α:

$$\mathsf{t}^2_{(\widetilde{n})} \geq c^2_\alpha > \mathsf{t}^2_{(\widetilde{n}+1)} \tag{3}$$

where $\widetilde{n}$ is the cut-off between the number of retained and excluded variables. In this setting, $\mathcal{M}_s(\alpha)$ will have a single member and there is no need for thick modelling, other than exploring the consequences of choosing different values of α. If the regressors are correlated, a multipath search must be conducted as the orderings based on t^2-statistics no longer represent the significance of the variables in the LDGP. Making a decision to retain/exclude a variable will affect the t-statistics of other variables, thereby changing the ordering. It is the multipath search that can lead to multiple terminal models which are mutually encompassing and this is more likely the higher the degree of correlation. One solution would be to find an orthogonal representation, so transforming an Autoregressive Distributed Lag (ADL) model to its Equilibrium Correction Mechanism (EqCM) representation, for example, would reduce the correlation structure. But some correlations are unavoidable, for example if a researcher was uncertain if the consumer price index or retail price index was the relevant measure of inflation and therefore included both in the GUM. These are highly correlated measures of inflation and if each was retained in different model specifications, thick modelling would capture this uncertainty.

Misspecification of the GUM can also to lead to a larger $\mathcal{M}_s(\alpha)$ as various irrelevant variables are retained in an attempt to proxy the variables omitted from the LDGP that are the cause of misspecification. *Autometrics* (Doornik, 2009, [19]) will still undertake the search procedure even if the GUM fails diagnostic tests, with a process of tightening the significance level of the diagnostic tests if they fail. The diagnostic tests can be user-specified, but for time-series modelling we include tests for autocorrelation, autoregressive conditional heteroskedasticity, normality, functional form and parameter constancy. *Autometrics* will try to restore diagnostic validity along the path searches if the GUM initially fails diagnostic tests. If there are multiple terminal models and some of those pass the diagnostic tests at the original significance level prior to it being reduced, then only those terminals are retained, so thick modelling in this setting would average over the congruent set of terminal models.

Castle and Hendry (2011, 2014) (see [12, 14]) discuss the implications of applying model selection to an underspecified GUM, i.e. one that omits relevant variables so the unknown

LDGP is not nested. In this setting, there can be no hope of finding a set of terminal models that closely approximates the LDGP. However, model selection still yields advantages over imposing a given theory model or simple model averaging when implementing impulse-indicator saturation (IIS) within the selection procedure. IIS includes a set of saturating impulse dummies in the GUM, see Hendry, Johansen and Santos (2008) [41], Johansen and Nielsen (2009, 2016) [52, 53], and Hendry and Santos (2010) [46], which acts as a robust method when there is model misspecification by accounting for location shifts and outliers in omitted relevant variables, helping to mitigate the adverse impacts of induced location shifts on intercepts and equation standard errors.[†] Interestingly, location shifts in omitted variables do not affect slope parameters, even when correlated with included variables, so thick modelling would be still be valid on the slope parameters in the case where the GUM is underspecified and the omitted variables shift. This suggests that IIS should always be implemented to provide robustness against misspecified GUMs and to account for outliers and location shifts in-sample.

In dynamic models, locating the precise timing of shifts can be difficult which may result in multiple terminal models in which the specifications differ due to the timing of retained impulses. Thick modelling would average over these terminal models, capturing the uncertainty of the timing of possible location shifts. In many cases, outliers or shifts are due to known reasons such as a policy change in a given month/quarter. Institutional knowledge of this kind would allow the researcher to narrow down the set of terminal models to retain those in which impulses correspond to known events. However, if there are dynamic effects which make policy effects slow to feed through, or the researcher is uncertain of such events leading to outliers or shifts in the data, then thick modelling provides a way of capturing the timing uncertainty.

Granger's proposal of thick modelling in which alternative specifications of similar quality are combined rather than selecting one and discarding all others is made feasible by using *Gets* which results in a set of terminal models that satisfy his criterion of similar valid models. Granger's pragmatic approach to modelling enables model averaging and model selection to complement each other. In the next section we briefly review the literature, before demonstrating the approach using artificial data.

3. Literature on thick modelling

Granger's proposal for thick modelling was motivated by his work on forecasting, where pooling of forecasts often outperformed the selected 'best' forecast. The subsequent literature using thick modelling has followed in this tradition, focussing on forecasting applications, see for example McNelis and McAdam (2004) [56] and Albacete and Espasa (2005) [3] for inflation forecasting examples, and Aiolfi and Favero (2005) [1] for an application

[†] Robustness in the statistics literature refers to methods that perform well under non-standard distributions, and in particular when there are large-sized observations or outliers. Here we use a broader definition of robustness, whereby methods are robust if they have good properties against many forms of misspecification, including outliers, location shifts, omitted variables, incorrect distributional shape, non-stationarity, misspecified dynamics and non-linearities.

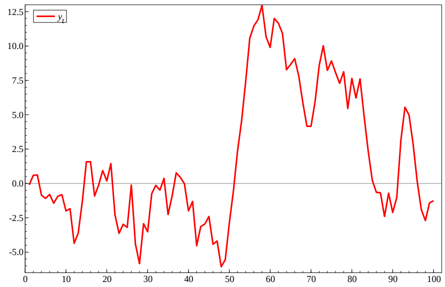

Figure 2: Dependent variable for the artificial data set

to the predictability of stock returns. Pesaran and Timmermann (2004) (see [62]) suggest thick modelling as a technique for real-time forecasting. Granger's second motivation for thick modelling came from his work on portfolio theory, and thick modelling has been applied in the multivariate volatility literature by papers such as Amendola and Storti (2015) [4], and Pesaran, Schleicher and Zaffaroni (2009) [61], who use thick modelling as a way to address model uncertainty in multivariate volatility models.

The literature has mainly focussed on the use of thick modelling in forecasting, where forecast combination has a long pedigree, but the proposal in Granger and Jeon (2004) [30] was not specific to forecasting, and thick modelling can be applied in-sample successfully as well, as we now explore.

4. Applying *Gets* with thick modelling to a data set

In this section, we apply *Autometrics* selection to a single draw from a known LDGP and then use thick modelling over the resulting terminal models.

Assume the LDGP is an $\mathsf{I}(0)$ autoregressive distributed lag model which contains two location shifts within the sample that shifts the unconditional mean but not the slope parameters:

$$y_t = \beta_0 + \beta_y y_{t-1} + \sum_{i=1}^{10} \beta_i x_{i,t} + \delta 1_{(t_L < t \leq t_U)} + \epsilon_t, \qquad \epsilon_t \sim \mathsf{IN}\left[0, 1\right] \tag{4}$$

where $1_{(t_L<t\leq t_U)}$ is an indicator function with the value zero except for unity over the interval $t_L = 50$ and $t_U = 81$, and $\mathbf{x}_t = (x_{1,t}, \ldots, x_{10,t})'$ is generated by:

$$\mathbf{x}_t = \lambda \mathbf{I} \mathbf{x}_{t-1} + \nu_t, \qquad \nu_t \sim \mathsf{IN}\left[\mathbf{0}_{10}, \boldsymbol{\Omega}\right] \tag{5}$$

We set the parameter values at $\beta_0 = 0$; $\beta_y = 0.6$; $\beta_1 = 0.2$; $\beta_2 = 0.3$; $\beta_3 = 0.4$; $\beta_4 = 0.5$; $\beta_5 = 0.6$; $\beta_6, \ldots, \beta_{10} = 0$ so only the first five variables enter the DGP, $\delta = 3$; $\lambda = 0.6$; $\boldsymbol{\Omega} = \mathbf{I}_{10}$, with $T = 100$, discarding an initial 20 observations. Figure 2 records the dependent variable.

4.1. Thick modelling with misspecification

First consider the case of selection with a misspecified model, so the DGP is not nested within the GUM:

$$y_t = \gamma_0 + \sum_{j=0}^{1} \gamma_{1j} x_{1,t-j} + \sum_{j=0}^{1} \gamma_{3j} x_{3,t-j} + \sum_{j=0}^{1} \gamma_{5j} x_{5,t-j} + \sum_{i=6}^{10} \sum_{j=0}^{1} \gamma_{ij} x_{i,t-j} + \eta_t \tag{6}$$

The GUM is heavily misspecified as it excludes the lagged dependent variable and two relevant exogenous regressors ($x_{2,t}$ and $x_{4,t}$) as well as ignoring the structural break. There are a further 5 irrelevant variables and their lags included in the GUM.

	TM1	TM2	TM3	TM4	TM5	TM6	TM7	Union	Pooled
Constant	1.976	1.923	2.028	1.939	2.048	2.009	2.008	2.138	2.009
$x_{3,t}$	.	0.524	.	.	0.256	.	.	0.174	0.119
$x_{3,t-1}$	0.921	.	0.804	0.637	0.816	0.816	0.920	0.901	0.727
$x_{5,t}$	0.992	.	.	.	0.953	.	0.862	0.883	0.461
$x_{5,t-1}$	1.325	1.638	1.868	1.754	1.227	1.852	1.333	1.351	1.543
$x_{6,t}$	.	-0.562	.	-0.532	-0.587	.	-0.568	-0.389	-0.330
$x_{6,t-1}$	-0.802	.	-0.704	.	.	-0.710	.	-0.576	-0.349
$x_{8,t}$	.	.	-1.062	.	-1.083	-1.049	-1.105	-0.578	-0.609
$x_{8,t-1}$	-1.111	-1.105	.	-1.098	.	.	.	-0.802	-0.514
$x_{9,t}$	.	.	.	.	-0.271	-0.278	.	-0.129	-0.085
$x_{10,t-1}$	.	.	0.590	.	.	.	0.503	0.491	0.198
$\mathcal{LL}$	-284.0	-288.2	-284.9	-287.5	-285.1	-286.0	-284.4	-281.3	
$\widehat{\sigma}$	4.273	4.429	4.309	4.402	4.367	4.360	4.313	4.273	
AR(2)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
ARCH(1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Normality	0.064	0.022	0.019	0.018	0.011	0.017	0.013	0.096	
Hetero	0.060	0.295	0.012	0.412	0.070	0.052	0.094	0.012	
Chow(70%)	0.206	0.762	0.415	0.776	0.322	0.488	0.274	0.275	

Table 1: Terminal models (TM) from Autometrics undertaking selection at a target size of 5% for GUM (6). Coefficient estimates reported along with p-values for diagnostic tests, the log-likelihood ($\mathcal{LL}$) and estimated equation standard error ($\widehat{\sigma}$). Union model includes all variables that are retained in one or more terminal models. Final column reports average coefficients across all terminal models and union model.

Table 1 records the parameter estimates and p-values for the diagnostic tests from the terminal models and the union of the terminal models, using a target significance level of 5% with default settings for *Autometrics*, fixing the intercept.[‡] The union model includes all variables which are retained in one or more terminal models. As a result of the

[‡] It is advisable to always fix the intercept, i.e. do not select over it, as the intercept can have a substantial impact on the set of models retained, particularly if there are structural breaks present that are not modelled. See Castle, Doornik and Hendry (2012) [10] for simulation evidence on *Autometrics* with a fixed or free intercept.

GUM misspecification, additional irrelevant regressors are retained to proxy the omitted variables. As these are imperfect proxies, a range of terminal models are retained, all of which are non-congruent as evidenced by a failure of the diagnostic tests, including most notably second order serial correlation (Godfrey, 1978, [22]) and first order autoregressive conditional heteroscedasticity (Engle, 1982, [20]). The Chow (1960) test (see [16]) splits the sample in a 70%/30% split and tests for parameter constancy using a residual sum of squares (RSS) F-statistic. The lack of failure of parameter constancy across all models is due to the model fitting badly before the breakpoint, thereby inflating the RSS. Thick modelling averages across the terminal models using a simple unweighted average, and we include the union of the terminal models in the set to be averaged, given in the final column of table 1.

The intercept, averaged over the full sample, is:

$$\mathsf{E}\left(y_t\right) = \frac{\delta 1_{(50<t\leq 81)}}{1-\beta_y} = \frac{0.31 \times 3}{1-0.6} = 2.325$$

as all exogenous variables have zero means in expectation, with the equilibrium mean converging quite quickly to a mean of 7.5 during the structural break, before converging to zero again after the break. All terminal models underestimate the full sample equilibrium mean if the breaks are not modelled, with the union providing the closest estimate.

The choice of α affects the set of terminal models. Selection at $\alpha = 0.1$ results in 3 terminal models, each with 6 parameters, and selection at $\alpha = 0.01$ results in 9 terminal models with between 4 and 8 parameters. Variables are retained despite statistical insignificance in the terminal models if including them mitigates the failure of diagnostic tests. In this misspecified case, tightening α results in more terminal models being retained. A large number of terminal models points towards possible misspecification as no model mutually encompasses the others.

The conditional expectation, $\widehat{y}_t = \mathsf{E}\left[y_t|\mathbf{X}_t\right]$ where $\mathbf{X}_t$ denotes the vector of conditioning variables including lagged variables, of each terminal model is plotted in figure 3, panel a. All the retained models follow a similar pattern although with some variation, captured by panel b which records the minimum and maximum conditional expectation computed at each observation to give a range in which the conditional expectation lies along with the actual outturn. The outturns frequently lie outside of the range of terminal models, which is unsurprising given the degree of misspecification in the GUM due to omitted breaks. Note that the range is not a measure of estimation uncertainty (i.e. computing $\mathsf{E}\left[y_t|\mathbf{X}_t\right]$ for $\pm 2\sigma$ of the parameter estimates). Instead it is a reflection of the possible admissible model specifications. Confidence intervals for thick modelling could be obtained by bootstrap aggregation (bagging, see Breiman, 1996, [8]) although the time series properties would need to be carefully handled. Another note of caution is that the terminal models are assumed to represent the admissible set defined by the set of congruent models. As the GUM does not nest the LDGP, neither the GUM or any reduction thereof will be congruent.

Thick modelling does not insure against model misspecification. Commencing with a non-congruent GUM is disastrous for in-sample modelling but is detected by the diag-

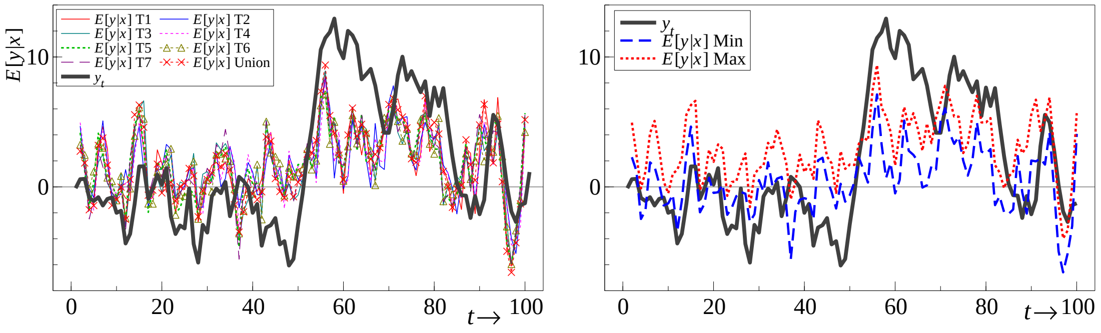

Figure 3: Selection from GUM (6); panel (a) y_t and E $\left[y|\mathbf{X}_t^1\right]$ for each terminal model; panel (b) y_t and pointwise minimum and maximum E $\left[y|\mathbf{X}_t^1\right]$.

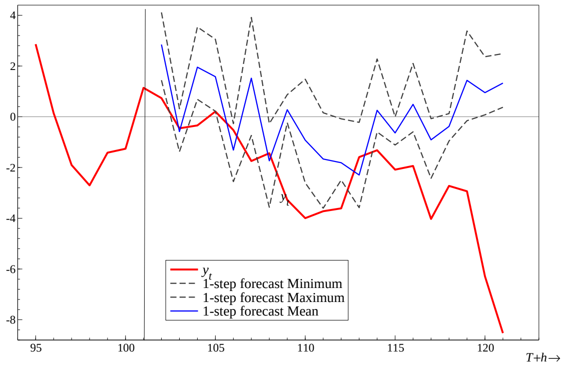

Figure 4: 1-step ahead conditional minimum, maximum and mean forecasts with known exogenous regressors from all terminal models, along with outturns.

nostic tests. Figure 4 records the 1-step ahead conditional forecasts, giving the pointwise minimum, mean and maximum forecasts. Although $\pm 2\sigma$ error bars are not included on the figure, it is clear that thick modelling does not solve the forecasting problem relative to selecting a single model as the set of models is misspecified, although model misspecification need not imply forecast failure, see *inter alia* Clements and Hendry (1998, 1999) [17, 18]. The large number of retained terminals for thick modelling could further indicate the failure of congruence, with a wide range for the pointwise conditional expectation function, but it does not 'fix' the problem of a lack of congruence or reveal anything useful about model uncertainty as all terminal models are equally poor.

The information set available should nest the LDGP for a successful modelling approach. Without this, any form of model selection or model averaging is likely to fail. With 16 variables in (6) (with a fixed intercept in every model) there are $2^{16} = 65,536$ possible models. Using an unweighted average over the full set, figure 5 records the model average fit and forecasts superimposed over the thick modelling results. The results are

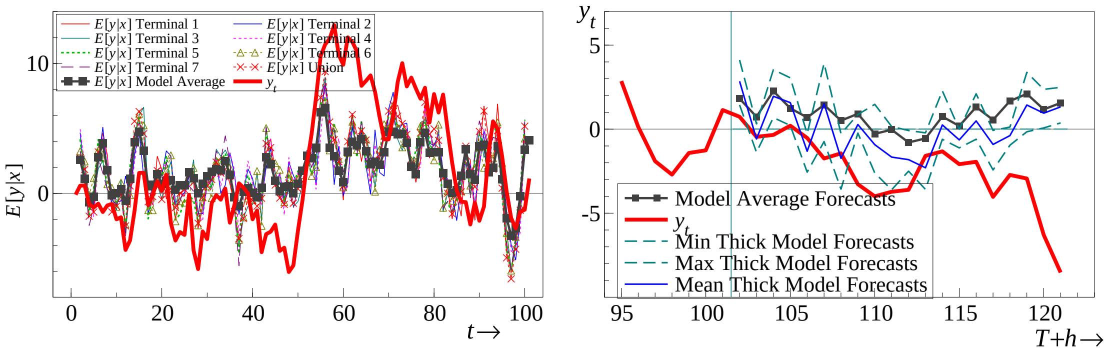

Figure 5: Panel (a): the conditional expectation for the model average and all terminal models used for thick modelling, plotted against the outturns; panel (b): forecasts from model averaging plotted with the 1-step ahead conditional minimum, maximum and mean forecasts with known exogenous regressors from all terminal models, along with outturns.

very similar to the thick model results, both in-sample and out-of-sample, both of which perform poorly. The technique, be it model averaging, thick or thin modelling, is of secondary importance to the requirement that the information set nests the LDGP. Without this, all forms of model building can fail. Methods that test for congruence to ensure the LDGP is nested within the available information set insure against this serious failure.

4.1.1. Alternative forms of misspecification

The information set in the above case did not nest the LDGP. Misspecification included omitted exogenous variables, omitting the lagged dependent variable and unmodelled structural breaks. We now extend the available information set slightly to include all relevant exogenous variables, but it still does not nest the LDGP as the GUM in equation (7) excludes the lagged dependent variable and the structural break:

$$y_t = \gamma_0 + \sum_{i=1}^{10}\sum_{j=0}^{1}\gamma_{ij}x_{i,t-j} + \eta_t \tag{7}$$

Results are similar to above, with selection at 5% resulting in 5 terminal models being retained, with 4 terminals retained at 10% and 13 terminals retained at the 1% target significance level. Figure 6, left hand panels, record the results at 5%. The pointwise minimum and maximum conditional expectation recorded in panel (a) has a slightly narrower range than in the case above, but the thick model bands still fail to fully capture the period around the location shift. Panel (c) is a cross plot of the conditional expectation against the outturn for each terminal model with fitted splines, which should be on the 45 degree line if the terminal models were well-specified. Finally, panel (e) records the minimum, maximum and mean 1-step ahead conditional forecasts with known exogenous regressors from all terminal models along with the forecast period outturns. Despite including the exogenous regressors the thick model forecasts poorly. Model averaging would require in excess of 1 million models, all of which would be misspecified.

Omitting the lagged dependent variable results in lags of the exogenous variables being retained in an attempt to proxy the dynamics. If we extend the information set to include the lagged dependent variable but not the structural break, model averaging would require more than 2 million models to be estimated, all of which are misspecified. In model selection, the GUM based on such an information set is given by:

$$y_t = \gamma_0 + \gamma_y y_{t-1} + \sum_{i=1}^{10} \sum_{j=0}^{1} \gamma_{ij} x_{i,t-j} + \eta_t \tag{8}$$

Selection from (8) at 5% results in 3 terminal models (the three identical terminals are retained at 10%) and 2 terminals are retained at 1%, corresponding to our conjecture regarding the dependence of $\mathcal{M}_s$ on α delivering fewer terminal models for a tighter selection significance level. This is in contrast to the above results where more variables were retained at tighter α, probably due to the degree of departure from congruence. The inclusion of the lagged dependent variable enables the structural break to be proxied by increased persistence ($\widehat{\beta}_y$ is biased upwards) so misspecification is less extreme.

Figure 6, right hand panels, record the results for selection at 5%. Inclusion of the lagged dependent variable results in very similar terminals as can be seen by the narrowing of the pointwise minimum and maximum conditional expectation. The differences are due to (i) one relevant variable being retained either t dated or $t-1$ dated, and (ii) one of two irrelevant variables being retained in each terminal model. With 15 irrelevant variables in the GUM we would expect 0.75 of a variable to be retained on average, so one irrelevant variable in each terminal model accords with the theory.

Despite the improved model fit of all terminal models, the failure to explicitly model the structural break results in an estimated coefficient on the lagged dependent variable that is too large:

$$\left[\widehat{\gamma}_y^{TM1} = 0.814; \widehat{\gamma}_y^{TM2} = 0.812; \widehat{\gamma}_y^{TM3} = 0.828; \widehat{\gamma}_y^{Union} = 0.809\right],$$

so the thick model estimate is $\widehat{\gamma}_y^{Thick} = 0.816$. The true parameter is $\beta_y = 0.6$. The LDV is biased upwards as more persistence is needed to capture the unmodelled structural break. Again, thick modelling serves the purpose of highlighting model uncertainty due to correlated lags or marginally significant irrelevant variables, but it does not solve misspecification arising from unmodelled breaks.

4.2. Selection from a well-specified GUM

Finally we consider an information set that nests the LDGP so the GUM and resulting terminal models are congruent. Equation (9) is the GUM specification in which we assume that the location shift is unknown and is therefore searched for using a full set of saturating

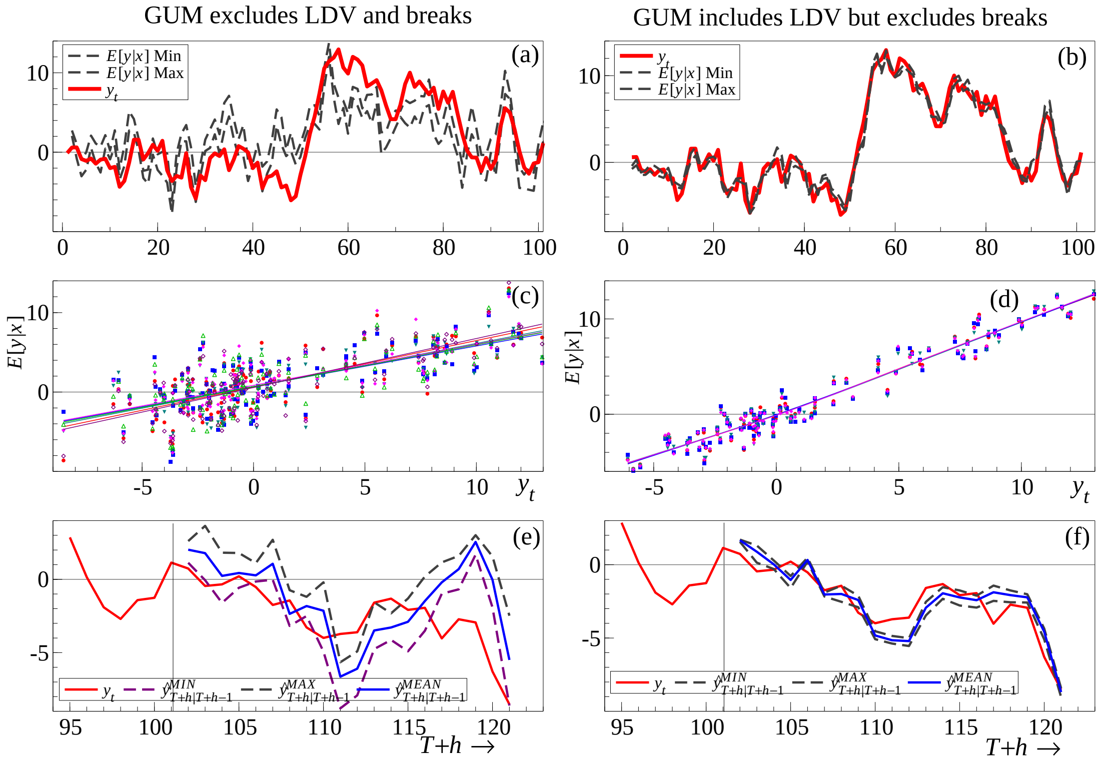

Figure 6: Left hand panels record selection from selection from (7) and right hand panels record selection from (8), using Autometrics with a 5% target size. Top panels plot pointwise minimum and maximum for conditional expectation against outturns, middle panel records cross plots of conditional expectations of terminal models against the dependent variable with splines and bottom panels record mean, minimum and maximum 1-step ahead point forecasts from pooled terminal models with known exogenous regressors.

impulse and step indicators.

$$y_t = \gamma_0 + \gamma_y y_{t-1} + \sum_{i=1}^{10}\sum_{j=0}^{1}\gamma_{ij}x_{i,t-j} + \sum_{j=1}^{T}\mu_j \mathbf{1}_{\{t=t_j\}} + \sum_{k=1}^{T-1}\kappa_j \mathbf{1}_{\{t\leq t_j\}} + \eta_t \tag{9}$$

Model averaging would estimate 2^{220} models. Even assuming the known break, so the step shift $1_{(50<t\leq 81)}$ is included in the information set, there would be over 4 million models to estimate of which all but one will be misspecified in some direction and half would exclude the known structural break. Even with tiny weights on the misspecified models, model averaging will not produce a conditional expectation close to the outturn.

The GUM includes 212 irrelevant regressors (excluding the intercept which is fixed), and so $N > T$, but selection is perfectly feasible using *Autometrics* within a few seconds, see Castle and Hendry (2014) [13]. Applying selection at the tighter target significance level of 2.5% results in two terminal models being retained, see table 2. We would expect approximately 5 irrelevant variables to be retained on average at this significance level but this is not the case for the given draw. At 10% 7 terminal models are retained, with the only difference in terminals due to the impulse and step indicators. At this loose significance level *Autometrics* heavily overfits, retaining between 68 and 74 indicator variables due to the downward bias in the equation standard error by removing many observations. Eliminating sequentially marginally significant indicators results in the significance of other indicators collapsing, which we term the 'house of cards' problem which can be present when the selection significance level is too loose. At 1% a unique terminal model is retained so no thick modelling is required.

	TM1	TM2	Union	Pooled
Constant	0.089	0.126	0.081	0.099
y_{t-1}	0.640	0.666	0.641	0.649
$x_{1,t}$	0.357	0.342	0.354	0.351
$x_{2,t-1}$	0.428	0.415	0.423	0.422
$x_{3,t}$	0.302	0.275	0.297	0.291
$x_{4,t}$	0.512	0.477	0.503	0.497
$x_{5,t}$	0.766	0.757	0.761	0.761
$\mathbf{1}_{t\leq 51}$	-2.528	-2.324	-2.547	-2.426*
$\mathbf{1}_{t\leq 82}$	.	2.338	0.611	1.169*
$\mathbf{1}_{t\leq 83}$	2.555	.	1.967	1.278*
$\mathcal{LL}$	-149.0	-150.3	-148.9	
$\widehat{\sigma}$	1.126	1.140	1.131	
AR(2)	0.230	0.158	0.229	
ARCH(1)	0.118	0.254	0.118	
Normality	0.553	0.470	0.557	
Hetero	0.141	0.146	0.132	
Chow(70%)	0.718	0.623	0.745	

Table 2: Terminal models from Autometrics undertaking selection at a target size of 2.5% for GUM (9). Coefficient estimates reported, with the unweighted average of TM1, TM2 and Union in the Pooled column. p-values for diagnostic tests reported, along with the log-likelihood ($\mathcal{LL}$) and estimated equation standard error ($\widehat{\sigma}$). * denotes averaging across TM1 and TM2 but excluding the union model.

The benefits of thick modelling are most apparent in this well-specified case where the exogenous regressors are consistently selected across terminals and the only difference between the two terminal models highlights the difficulty in pinpointing the precise date of the structural break. Both terminals are mutually encompassing so thick modelling reflects the uncertainty of the break timing. Thick modelling proposes to average the coefficient estimates if a point estimate is required, although just reporting the range would be in keeping with the thick modelling philosophy. Mean coefficient estimates for step indicators should be treated carefully if averaging with the union of the terminal models. TM1 and TM2 retain different break dates, but the union combines these resulting in a joint effect of the step shift given by the sum of the coefficients in the union model. Therefore we pool the step shift coefficients by excluding the union model.

In order to judge how well the thick modelling performs we should evaluate the pooled model against the estimated LDGP. Even if a researcher knew the LDGP specification

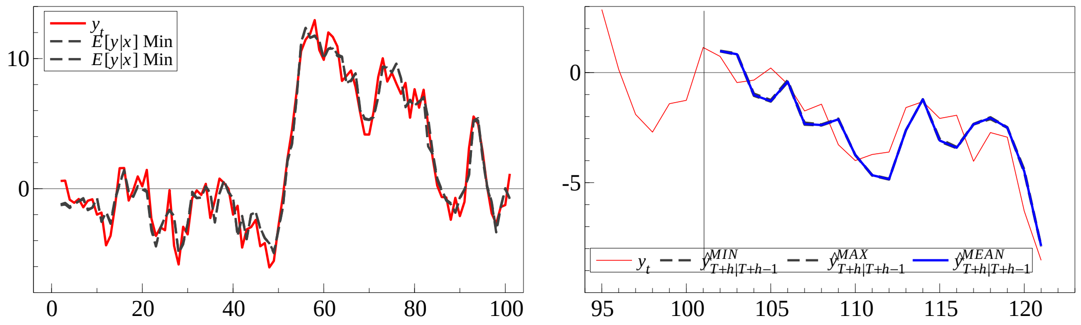

Figure 7: In-sample model fit and forecasts for selection from congruent GUM

she would still need to estimate the model, and as such could make mistakes based on inference from the estimated LDGP. The estimated LDGP is given in equation (10):

$$\begin{aligned} y_t &= \underset{(0.041)}{0.704}y_{t-1} + \underset{(0.087)}{0.265}x_{1,t} + \underset{(0.086)}{0.313}x_{2,t} + \underset{(0.093)}{0.289}x_{3,t} + \underset{(0.098)}{0.441}x_{4,t} \\ &\quad + \underset{(0.113)}{0.732}x_{5,t} + \underset{(0.400)}{2.141}\mathbf{1}_{\{50<t\leq 81\}} \\ \widehat{\sigma} &= 1.176 \end{aligned} \tag{10}$$

Both terminal models retain an irrelevant variable, $x_{2,t-1}$, instead of the relevant variable $x_{2,t}$, which has $\widehat{\mathsf{t}} = 3.64$. The other pooled coefficient estimates are close to the estimated LDGP, with the lagged dependent variable closer to the true parameter value than the estimated LDGP. Indeed, the terminals and the pooled model have smaller equation standard errors than the estimated LDGP, although fit is not a criterion used for selection.

Figure 7 records the thick model bands given by the pointwise minimum and maximum conditional expectation along with the outturns. The range has narrowed, even around the uncertain break date, so the class of mutually encompassing congruent models all result in very similar in-sample predictions. The purpose of thick modelling here is to demonstrate that there is very little model uncertainty. Panel b records the 1-step ahead conditional forecasts taken as the pointwise minimum, mean and maximum forecasts from the thick modelling. Again, the range is virtually indistinguishable from a point forecast. Applying forecast error bands to the thick model range would demonstrate the good forecast performance of the thick model.

5. Conclusions

Granger's pragmatic approach to econometrics led him to recognise that econometric models will never exactly capture the LDGP, but will be approximations to it, with some models closer approximations than others. Rather than discarding useful information in the process of selecting just one unique model there are advantages to retaining all specifications that are close approximations. Combining this argument with the theory of reduction, which gives measurable criteria for determining whether a model is a close approximation to the LDGP, enables thick modelling over a congruent encompassing set of models that are valid representations of the phenomenon of interest.

Thick modelling does not insure against model misspecification. The information set must nest the LDGP, otherwise any form of model selection or model averaging is likely to fail. If the information set available results in a poorly specified GUM it is unlikely that any resulting terminal model will be congruent. Thick modelling across non-congruent terminal models will lead to a wide pointwise range for the conditional expectation function. Such a measure of model uncertainty is another signal of non-congruence. Model averaging will fail in this context too.

Thick modelling is very helpful in cases where variables are highly correlated such as where two consecutive lags are indistinguishable or where different impulses are retained but are within close proximity of each other. In this setting, thick modelling flags model uncertainty which will have implications for the timing of policy implementation, for example. The resulting thick model in which parameter estimates are pooled (as opposed to variables, as in the union model, or forecasts) will enable pointwise interpretation, although statistical tests will require the standard errors to be pooled as well.

The dependence of the set of terminals for thick modelling on α suggest that different criteria may be used for in-sample modelling and forecasting. Typically a tight significance level is used such as $\alpha = min\,(1/N, 1/T, 1\%)$, particularly when applying IIS and/or SIS, to control the gauge. In forecasting it is not clear that ensuring parsimony by tight selection is advantageous. Castle, Doornik and Hendry (2016) (see [11]) suggest setting α between 15-30% to obtain a large set of undominated congruent terminal models. However, these are likely to contain many adventitiously significant variables, and so applying bias correction (Hendry and Krolzig, 2005, [43]) will downweight such effects. Averaging over the resulting class of retained terminals will reduce the problem of retaining bad models which selection aims to address, but will capture a broader range of close specifications that thick modelling advises. A further alternative would be to apply thick modelling by averaging over the terminals from a range of α, which is left for future research.

Although thick modelling does not overcome misspecification, indicator saturation techniques such as IIS and SIS help to deliver well-specified models. Such techniques are a form of robust estimation (see Johansen and Nielsen, 2016, [53]), such that the resulting terminal models are valid representations of the data. Thick modelling in conjunction with robust *Gets* selection using IIS and SIS would seem to be a pragmatic and informative approach to model building which is in keeping with Granger's views.

Acknowledgements

Financial support from the Institute for New Economic Thinking is gratefully acknowledged, as are helpful comments on an earlier draft by Michael P. Clements, David F. Hendry, Ryoko Ito and Bent Nielsen.

References

[1] M. Aiolfi and C. A. Favero. Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting*, 24(4):233–254, 2005.

[2] A. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. L. Csaki, editors, *Second International Symposium of Information Theory*, pages 267–281. Akademiai Kiado, Budapest, 1973.

[3] R. Albacete and A. Espasa. Forecasting inflation in the euro area using monthly time series models and quarterly econometric models. DES Working Paper No. 050401, Departamento de Estadistica, Universidad Carlos III de Madrid, 2005.

[4] A. Amendola and G. Storti. A thick modeling approach to multivariate volatility prediction. In M. Carpita, E. Brentari, and M. Qannari, editors, *Advances in Latent Variables: Methods, Models and Applications*, pages 207–217. Springer International Publishing, 2015.

[5] J. M. Bates and C. W. J. Granger. The combination of forecasts. *Operations Research Quarterly*, 20:451–468, 1969.

[6] C. Bontemps and G. E. Mizon. Congruence and encompassing. In B. P. Stigum, editor, *Econometrics and the Philosophy of Economics*, pages 354–378. Princeton University Press, Princeton, 2003.

[7] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995.

[8] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[9] J. L. Castle, J. A. Doornik, and D. F. Hendry. Evaluating automatic model selection. *Journal of Time Series Econometrics*, 3(1):DOI: 10.2202/1941–1928.1097, 2011.

[10] J. L. Castle, J. A. Doornik, and D. F. Hendry. Model selection when there are multiple breaks. *Journal of Econometrics*, 169(2):239–246, 2012.

[11] J. L. Castle, J. A. Doornik, and D. F. Hendry. Automatic model selection using soft thresholding. Mimeo, Economics Department, University of Oxford, 2016.

[12] J. L. Castle and D. F. Hendry. A Tale of 3 Cities: Model Selection in Over-, Exact, and Under-specified Equations. In M. Kaldor and P. Vizard, editors, *Arguing About the World*, pages 31–55. Bloomsbury Academic, London, 2011.

[13] J. L. Castle and D. F. Hendry. 'Data mining' with more variables than observations. VoxEU article, 13 August, 2014, http://www.voxeu.org/article/data-mining-more-variables-observations, 2014.

[14] J. L. Castle and D. F. Hendry. Model selection in under-specified equations facing breaks. *Journal of Econometrics*, 178(2):286–293, 2014.

[15] J. L. Castle and N. Shephard, editors. *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry.* Oxford University Press, Oxford, 2009.

[16] G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605, 1960.

[17] M. P. Clements and D. F. Hendry. *Forecasting Economic Time Series.* Cambridge University Press, Cambridge, 1998.

[18] M. P. Clements and D. F. Hendry. *Forecasting Non-stationary Economic Time Series.* MIT Press, Cambridge, Massachusetts, 1999.

[19] J. A. Doornik. Autometrics. In Castle and Shephard [15], pages 88–121.

[20] R. F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, 50:987–1008, 1982.

[21] C. Gatu and E. J. Kontoghiorghes. Branch-and-bound algorithms for computing the best subset regression models. *Journal of Computational and Graphical Statistics*, 15:139–156, 2006.

[22] L. G. Godfrey. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, 46:1293–1301, 1978.

[23] C. W. J. Granger, editor. *Modelling Economic Series.* Clarendon Press, Oxford, 1990.

[24] C. W. J. Granger. *Empirical Modeling in Economics: Specification and Evaluation.* Cambridge University Press, Cambridge, 1999.

[25] C. W. J. Granger. Some comments on risk. *Journal of Applied Econometrics*, 17(5):447–456, 2002.

[26] C. W. J. Granger. Modeling, evaluation, and methodology in the new century. *Economic Inquiry*, 43(1):1–12, 2005.

[27] C. W. J. Granger. The past and future of empirical finance: some personal comments. *Journal of Econometrics*, 129:35–40, 2005.

[28] C. W. J. Granger. In praise of pragmatics in econometrics. In Castle and Shephard [15], pages 255–267.

[29] C. W. J. Granger and D. F. Hendry. A dialogue concerning a new instrument for econometric modeling. *Econometric Theory*, 21:278–297, 2005.

[30] C. W. J. Granger and Y. Jeon. Thick modeling. *Economic Modelling*, 21:323–343, 2004.

[31] C. W. J. Granger, M. L. King, and H. White. Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics*, 67:173–187, 1995.

[32] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, B, 41:190–195, 1979.

[33] B. E. Hansen. Least-squares forecast averaging. *Journal of Econometrics*, 146:342–350, 2008.

[34] D. F. Hendry. *Dynamic Econometrics.* Oxford University Press, Oxford, 1995.

[35] D. F. Hendry. Denis Sargan and the origins of LSE econometric methodology. *Econometric Theory*, 19:456–480, 2003.

[36] D. F. Hendry. The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. In T. C. Mills and K. D. Patterson, editors, *Palgrave Handbook of Econometrics*, pages 3–67. Palgrave MacMillan, Basingstoke, 2009.

[37] D. F. Hendry. Revisiting UK consumers' expenditure: Cointegration, breaks, and robust forecasts. *Applied Financial Economics*, 21:19–32, 2010.

[38] D. F. Hendry and M. P. Clements. Pooling of forecasts. *Econometrics Journal*, 7:1–31, 2004.

[39] D. F. Hendry and J. A. Doornik. *Empirical Model Discovery and Theory Evaluation.* MIT Press, Cambridge MA, 2014.

[40] D. F. Hendry and S. Johansen. Model discovery and Trygve Haavelmo's legacy. *Econometric Theory*, 31:93–114, 2015.

[41] D. F. Hendry, S. Johansen, and C. Santos. Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, 33:317–335, 2008. Erratum, 337–339.

[42] D. F. Hendry and H-M. Krolzig. Improving on 'data mining reconsidered' by K.D. Hoover and S.J. Perez. *Econometrics Journal*, 2:41–58, 1999.

[43] D. F. Hendry and H.-M. Krolzig. The properties of automatic *Gets* modelling. *Economic Journal*, 115:C32–C61, 2005.

[44] D. F. Hendry and J. J. Reade. Modelling and forecasting using model averaging. Working paper, Economics Department, Oxford University, 2008.

[45] D. F. Hendry and J.-F. Richard. On the formulation of empirical models in dynamic econometrics. In B. Cornet and H. Tulkens, editors, *Contributions to Operations Research and Economics: The Twentieth Anniversary of Core.* MIT Press, Cambridge, Massachusetts, 1989.

[46] D. F. Hendry and C. Santos. An automatic test of super exogeneity. In M. W. Watson, T. Bollerslev, and J. Russell, editors, *Volatility and Time Series Econometrics*, pages 164–193. Oxford University Press, Oxford, 2010.

[47] R. R. Hocking and R. N. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9:531–540, 1967.

[48] A. E. Hoerl and R. W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970.

[49] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[50] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999.

[51] K. D. Hoover and S. J. Perez. Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, 2:167–191, 1999.

[52] S. Johansen and B. Nielsen. An analysis of the indicator saturation estimator as a robust regression estimator. In Castle and Shephard [15], pages 1–36.

[53] S. Johansen and B. Nielsen. Outlier detection algorithms for least squares time series regression. *Scandinavian Journal of Statistics*, 43(2):321–348, 2016. With Discussion.

[54] L. R. LaMotte and R. R. Hocking. Computational efficiency in the selection of regression variables. *Technometrics*, 12:83–93, 1970.

[55] E. E. Leamer. Let's take the con out of econometrics. *American Economic Review*, 73:31–43, 1983. Reprinted in Granger (1990), pp. 29–49.

[56] P. McNelis and P. McAdam. Forecasting inflation with thick models and neural networks. Working paper 352, European Central Bank, 2004.

[57] G. E. Mizon. The encompassing approach in econometrics. In D. F. Hendry and K. F. Wallis, editors, *Econometrics and Quantitative Economics*, pages 135–172. Blackwells, Oxford, 1994.

[58] G. E. Mizon. Progressive modelling of macroeconomic time series: the LSE methodology. In K. D. Hoover, editor, *Macroeconometrics: Developments, Tensions and Prospects*, pages 107–170. Kluwer, Boston, 1995.

[59] G. E. Mizon and J.-F. Richard. The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*, 54:657–678, 1986.

[60] A. R. Pagan. Three econometric methodologies: a critical appraisal. *Journal of Economic Surveys*, 1:3–24, 1987.

[61] M. H. Pesaran, C. Schleicher, and P. Zaffaroni. Model averaging in risk management with an application to futures markets. *Journal of Empirical Finance*, 16:280–305, 2009.

[62] M. H. Pesaran and A. G. Timmermann. Real time econometrics. IZA Discussion Paper No. 1108, CESifo Working Paper Series No. 1169, 2004. Available at SSRN: http://ssrn.com/abstract=533685.

[63] P. C. B. Phillips. Reflections on econometric methodology. *Economic Record*, 64:344–359, 1988.

[64] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[65] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.