Doornik, Jurgen A.; Hendry, David F.

Article

Statistical model selection with "Big Data"

Cogent Economics & Finance

Provided in Cooperation with:
Taylor & Francis Group

Suggested Citation: Doornik, Jurgen A.; Hendry, David F. (2015) : Statistical model selection with "Big Data", Cogent Economics & Finance, ISSN 2332-2039, Taylor & Francis, Abingdon, Vol. 3, Iss. 1, pp. 1-15,
https://doi.org/10.1080/23322039.2015.1045216

Doornik & Hendry, *Cogent Economics & Finance* (2015), 3: 1045216
http://dx.doi.org/10.1080/23322039.2015.1045216

cogent economics & finance

ECONOMIC METHODOLOGY, PHILOSOPHY & HISTORY | RESEARCH ARTICLE

Statistical model selection with "Big Data"

Jurgen A. Doornik[1,2] and David F. Hendry[1,2*]

Abstract: Big Data offer potential benefits for statistical modelling, but confront problems including an excess of false positives, mistaking correlations for causes, ignoring sampling biases and selecting by inappropriate methods. We consider the many important requirements when searching for a data-based relationship using Big Data, and the possible role of Autometrics in that context. Paramount considerations include embedding relationships in general initial models, possibly restricting the number of variables to be selected over by non-statistical criteria (the formulation problem), using good quality data on all variables, analyzed with tight significance levels by a powerful selection procedure, retaining available theory insights (the selection problem) while testing for relationships being well specified and invariant to shifts in explanatory variables (the evaluation problem), using a viable approach that resolves the computational problem of immense numbers of possible models.

Subjects: Computation; Economics; Statistics; Statistics for Business, Finance & Economics

Keywords: Big Data; model selection; location shifts; Autometrics; computational problems

JEL classifications: C52; C22

Received: 11 March 2015
Accepted: 01 April 2015
Published: 22 May 2015

*Corresponding author: David F. Hendry, Institute for New Economic Thinking, Oxford Martin School, Oxford, UK; Economics Department, Oxford University, Oxford, UK
Email: david.hendry@nuffield.ox.ac.uk

Reviewing editor:
Steve Cook, Swansea University, UK

Additional information is available at the end of the article

ABOUT THE AUTHORS

Jurgen A Doornik is James Martin Fellow, Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, research fellow at Nuffield College and director of OxMetrics Technologies Ltd . He has published widely on econometric methods, modelling, software, numerical methods, computation and mathematics and developed the OxMetrics software packages, including Ox and PcGive (the latter with DF Hendry), which incorporate Autometrics, an algorithm implementing automated general-to-specific model selection.

David F Hendry is director of the Program in Economic Modeling, Institute for New Economic Thinking at the Oxford Martin School, professor of Economics and fellow of Nuffield College, Oxford University. He was knighted in 2009, and received a Lifetime Achievement Award from the Economic and Social Research Council in 2014. He has received eight Honorary Doctorates, is a Thomson Reuters Citation Laureate and has published more than 200 papers and 25 books, the latest being Empirical Model Discovery and Theory Evaluation, MIT Press, 2014, with JA Doornik.

PUBLIC INTEREST STATEMENT

Big Data offer potential benefits for discovering empirical links, but confront potentially serious problems unless modelled with care. Key dangers include finding spurious relationships, mistaking correlations for causes, ignoring sampling biases and over-stating the significance of results. We describe the requirements needed to avoid these four difficulties when seeking relationships in Big Data. Important considerations are to commence from a general initial framework that allows for all influences likely to matter (the *formulation* problem), and use high-quality data analysed by a powerful search algorithm, requiring high significance (the *selection* problem). It is also crucial not to neglect insights from prior theory-based knowledge, and to test that claimed relationships both characterize all the evidence and are not changing over time (the *evaluation* problem). Finally, one must use a method that can efficiently handle immense numbers of possible models (the *computational* problem). Our approach provides a solution to all four problems.

1. Introduction

Mining is a productive activity when efficiently conducted, carefully sifting small quantities of valuable ores from volumes of dross, but occasionally mistaking iron pyrites for gold. Likewise, data mining to discover a few substantive relationships among vast numbers of spurious connections requires an appropriate approach. Simply choosing the best fitting regressions, or trying hundreds of empirical fits and selecting a preferred one despite it being contradicted by others that are not reported, is not going to lead to a useful outcome. Nevertheless, that some approaches to a problem are bad does not entail that all are: randomly digging holes hardly competes with geology-based searches for minerals. Similarly, *when properly undertaken*, statistical modelling by mining Big Data can be productive.

Big Data differ greatly across disciplines, come in several forms, such as photographic, binary and numerical, and even the third (on which we focus here) has three basic shapes, namely "tall" (not so many variables, N, but many observations, T, with $T \gg N$), "fat" (many variables, but not so many observations, $N > T$) and "huge" (many variables and many observations, $T > N$). Then there are three main forms of numerical data: a cross-section of observations on many "individuals" at a single point in time; a time series of observations on many variables; and a panel. Each disciplines' form-shape combination of data poses different problems, partly dependent on the magnitude of, and relation between, N and T and partly on data properties and partly on the type of analysis itself. Here we only address cross-sections or time series of "fat" data in economics, when the overall data-set is not "too big", although some of the resulting considerations apply with suitable modifications to "tall" and "huge". But:

> Recall big data's four articles of faith. Uncanny accuracy is easy to overrate if we ignore false positives... The claim that causation has been 'knocked off its pedestal' is fine if we are making predictions in a stable environment but not if the world is changing .... or if we ourselves hope to change it. The promise that 'N = All', and therefore that sampling bias does not matter, is simply not true in most cases that count. As for the idea that 'with enough data, the numbers speak for themselves' – that seems hopelessly naive in data sets where spurious patterns vastly outnumber genuine discoveries.
>
> 'Big data' has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers – without making the same old statistical mistakes on a grander scale than ever. (Harford, 2014)

Researchers in biology also worry that "Big Data is not a big deal, just another tool" and echo one of Harford's worries that it "can only give associations, not causal connections or mechanisms": see http://blog.lindau-nobel.org/big-data-not-a-big-deal-just-another-tool/.

Fortunately, we can counter most of the difficulties raised by Harford. Taking them in turn, we can calculate the probabilities of false positives in advance, and set the significance level to control them at the desired magnitude (see Section 3.2). We can also test "causation" when the world changes by evaluating super exogeneity (Section 4.2). Although hidden dependence in cross-section data is a potentially serious problem that needs to be addressed, selection biases can be corrected (see Section 3.1). We will show that using automatic methods, genuine discoveries are possible even when vastly outnumbered by spurious patterns (Section 3). In his recent survey, Varian (2014) also argues that automatic methods can be productively applied to Big Data: large data-sets allow for more flexible models, in which case, many potential predictors entail the need for automatic methods to select variables. Hendry and Doornik (2014) describe how to extend the reach of statistical modelling to the discovery of new knowledge, leading to a virtuous circle of further theoretical insights and better empirical models. If it can facilitate such discovery, "Big Data" might be a big deal.

Economic data are approximate measurements of an evolving, high-dimensional, inter-correlated and probably non-linear system prone to abrupt shifts, where the observations themselves are subject to intermittent revisions. Representation of such a non-stationary process requires models that account for all the substantively important variables, their dynamic reactions, any outliers and shifts

and non-linear dependencies. Omitting key features from any selected model will result in erroneous conclusions, as other aspects of that model will proxy missing information. But investigators cannot know the complete and correct specification of an empirical model in advance: models must be data based on the available sample to *discover* what matters. Consequently, the fundamental problem facing any form of statistical data analysis is how to avoid concluding with a substantively mis-specified model or a spurious relationship at the extreme. That entails four distinct sub-problems that must be resolved successfully. The *initial formulation problem* concerns ensuring that all the relevant variables and transformations thereof are included in the set of candidates; the *selection problem* requires eliminating effects that do not matter while retaining those that do; the *evaluation problem* involves checking that a well-specified model has indeed been discovered; and the *computational problem* requires an approach that can handle selection from large numbers of candidate variables without jeopardizing the chances of locating a good model. Here, we show how the four problems can be tackled using the automatic model selection algorithm *Autometrics* (see Doornik, 2009; Doornik & Hendry, 2013). We will consider its application to "fat Big Data" when $N > T$, after allowing for all the potential determinants jointly at the outset, which is not as intractable a problem as it seems.

Section 2 addresses the initial formulation problem: lag creation in Section 2.1, non-linear extensions in Section 2.2 and multiple outliers and shifts in Section 2.3, leading to the general unrestricted model in Section 2.4. Section 3 considers the model selection problem, introducing "1-cut" selection in Section 3.1 when $T > N$, multiple testing probabilities under the null in Section 3.2 and under the alternative in Section 3.3. Then Section 3.4 describes the move from 1-cut to *Autometrics*, Section 3.5 considers how to handle more candidate variables than observations and Section 3.6 explains embedding theory insights. Section 4 turns to the model evaluation problem: mis-specification testing in Section 4.1 and testing super exogeneity in Section 4.2. Section 5 discusses some of the likely computational issues for large $N > T$ and Section 5.1 considers block searches as a possible solution. Section 6 provides an illustrative example for an orthogonal setting with $N > T$, when all variables are irrelevant in Section 6.1, when 10 are relevant in Section 6.2 and Section 6.3 extends the second example to correlated regressors, re-selected by the Lasso in Section 6.4. Section 7 concludes.

2. The initial formulation problem

The aim of the formulation step is to commence from a model that satisfies all the requirements for valid inference, so that selection decisions are reliable. The approach underlying *Autometrics* is based on the theory of reduction described in Cook and Hendry (1993) and Hendry and Doornik (2014), Ch. 6, which delineates six distinct information sets, namely the (relative) (i) past, (ii) present and (iii) future of an investigator's own data, (iv) available theoretical insights, (v) knowledge about how the data are measured and (vi) separate information that is used by alternative models. These six sets can be mapped to null hypotheses that the model has (i) homoskedastic innovation errors that are normally distributed; (ii) weakly exogenous conditioning variables as in Engle, Hendry and Richard (1983); (iii) constant and invariant parameters; (iv) with theory consistent, identified parameters in (v) a data-admissible specification using accurate observations; and (vi) that encompasses rival explanations (see e.g. Mizon & Richard, 1986). This leads to a corresponding set of mis-specification statistics to test for (i) heteroskedasticity, autocorrelation and non-normality; (ii) failures of weak (or super) exogeneity; (iii) parameter non-constancy or forecast failure; (iv) invalid restrictions or a failure of over-identification; (v) measurement errors; and (vi) non-encompassing. While the null hypotheses to be tested are easily listed, there are many possible alternatives against which they might be tested, and many forms of test for each alternative. Section 4.1 discusses the specific implementations in *Autometrics*. Models that satisfy (i)–(iii) are said to be congruent with the available information (see e.g. Hendry & Nielsen, 2007).

Consequently, a viable empirical model needs to include all the substantively relevant variables (key determinants), their lagged responses (dynamic reactions), the functional forms of their relationships (non-linearities), capture any outliers or shifts and model unit roots and cointegration (non-stationarities), and establish the validity of conditioning (exogeneity, addressed in Section 4.2). This

initial formulation problem is especially acute for Big Data as it can entail a vast number of candidate variables. Let there be n basic explanatory variables, $\{z_{1,t}, \ldots, z_{n,t}\}$ denoted by the vector $\{\mathbf{z}_t\}$, of a variable $\{y_t\}$ to be modelled, then taking a time series as the example, creating up to s lagged values leads to $K = n(s+1)+s$ linear variables. To allow for the relation being non-linear, up to a cubic polynomial say, the number of additional quadratic and cubic terms is $N_K = K(K+1)(K+5)/6$. There is an explosion in the number of terms as K increases:

K	20	40	100	1000	5000
N_K	1750	12300	176,750	167,671,000	2×10^{10}
3K	60	120	300	3000	15000

quickly reaching huge N_K. We explain in Section 2.2 below how to circumvent this infeasible problem, with a solution that includes squares, cubes and even exponentials for $3K$ rather than $N_K + K$ additional variables, as shown in the bottom row of the table: large but not huge. Allowing for possible outliers at any number of data points will add another T indicator variable in our approach, explained in Section 2.3, and hence we are bound to face $N > T$. We will focus on single equations, but systems can be handled.

The resulting huge set of functions needs automatic model *creation*. We consider automatically creating three extensions outside the standard information: the lag formulation to implement a sequential factorization (see Doob, 1953), and thereby create a martingale difference sequence (MDS) or innovation error (Section 2.1); functional form transformations to capture any non-linearities (Section 2.2); and indicator saturation for outliers and location shifts to tackle parameter non-constancy (Section 2.3). Combined, these automatic extensions create the general unrestricted model in Section 2.4 that provides the starting point for selection.

2.1. Lag creation

It is straightforward to automatically create s lags of all variables to formulate the dynamic linear model:

$$y_t = \beta_0 + \sum_{i=1}^{s} \lambda_i y_{t-i} + \sum_{j=1}^{n} \sum_{i=0}^{s} \beta_{j,i} z_{j,t-i} + \epsilon_t \tag{1}$$

Valid inference during selection requires that $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$ if the Normal critical values used for decisions are to correctly represent the uncertainty. Cross-section models need the equivalent of a sequential factorization to avoid hidden dependence.

2.2. Non-linear extensions

Our approach to automatic non-linear extensions is based on Castle and Hendry (2010) who propose squares, cubics and exponentials of principal components $\mathbf{u}_t$ of the scaled $\mathbf{z}_t$. Let $\mathbf{z}_t \sim \mathsf{D}_n[\boldsymbol{\mu}, \boldsymbol{\Omega}]$, where $\boldsymbol{\Omega} = \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}'$ with $\mathbf{H}'\mathbf{H} = \mathbf{I}_n$. Empirically, let $\widehat{\boldsymbol{\Omega}} = T^{-1}\sum_{t=1}^{T}(\mathbf{z}_t - \bar{\mathbf{z}})(\mathbf{z}_t - \bar{\mathbf{z}})' = \widehat{\mathbf{H}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{H}}'$ so that $\mathbf{u}_t = \widehat{\mathbf{H}}'(\mathbf{z}_t - \bar{\mathbf{z}})$, which leads to $\mathbf{u}_t \underset{\text{app}}{\sim} \mathsf{D}_n[\mathbf{0}, \mathbf{I}]$. Now create squares, cubics and exponential functions of the individual $u_{i,t}$, namely $u_{i,t}^2$, $u_{i,t}^3$ and $u_{i,t}\exp(-|u_{i,t}|)$. When $\widehat{\boldsymbol{\Omega}}$ is non-diagonal, each $u_{i,t}$ is a linear combination of almost every $z_{i,t}$, so, for example, $u_{i,t}^2$ involves squares and cross-products of every $z_{i,t}$ etc. Then there are just $3n$ (or approximately $3K$ with lags) non-linear terms, so this achieves a low-dimensional representation of most of the important sources of departure from linearity, including asymmetry and sign-preserving cubics, with no collinearity between elements of $\mathbf{u}_t$, once the non-linear functions are demeaned.

Non-linear functions of y_t raise more problematic modelling issues, and are not addressed here (see e.g. Castle & Hendry, 2014; Granger & Teräsvirta, 1993).

2.3. Multiple outliers and shifts

To tackle multiple outliers and shifts for T observations, we add T impulse indicators, $1_{\{i=t\}}, t = 1, \dots, T$, to the set of candidate variables. Hendry, Johansen, and Santos (2008) call this impulse–indicator saturation (IIS), and use a "split-half" analysis to explain its feasibility. Set a critical value c_α for selection. Include the first half of the indicators, which just "dummies out" $T/2$ observations for estimation, record the significant ones, omit all first-half indicators and repeat for the second half. Then combine the recorded sub-sample indicators to select the significant ones overall. Under the null, αT indicators will be significant by chance. Johansen and Nielsen (2009) extend IIS to both stationary and unit-root autoregressions. When the error distribution is symmetric, adding T impulse indicators to a regression with $r < T/2$ variables, which are not subject to selection, coefficient $\boldsymbol{\beta}$ and second moment $\boldsymbol{\Sigma}$, they show:

$$T^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathsf{N}_r\left[\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega}_\alpha\right] \tag{2}$$

Thus, the usual $\sqrt{T}$ rate of convergence occurs to an asymptotic normal distribution centred on the population parameter, despite allowing for T irrelevant indicators. The efficiency of the IIS estimator $\widetilde{\boldsymbol{\beta}}$ with respect to OLS $\widehat{\boldsymbol{\beta}}$ is measured by $\boldsymbol{\Omega}_\alpha$, which depends on c_α and the underlying distribution, but is close to $(1-\alpha)^{-1}\mathbf{I}_r$. IIS must lose efficiency under the null of no outliers or shifts, but that loss is surprisingly small at roughly αT, so is (e.g.) only 1% at $\alpha = 1/T$ if $T = 100$. The potential benefit is for a major gain under the alternative of outliers and/or shifts, in a procedure that can be undertaken jointly with all other selections. Castle, Doornik, and Hendry (2012) show that IIS can handle "fat-tailed" distributions to keep normality as a good approximation for determining critical values during selection. Section 3.5 discusses the general approach in *Autometrics* when $N > T$.

Many well-known procedures are variants of IIS. Recursive estimation is IIS over future observations, reducing indicators one at a time, but not examining the information in the unused data. Moving windows uses IIS on pre and post data, and "hold back" is equivalent to IIS over excluded data points. The Chow (1960) test is IIS on a specific sub-set of data as shown by Salkever (1976). Arbitrarily excluding data (e.g. "war years" or shortening a sample early because of a known shift, etc.) implicitly uses IIS, ignoring the information in the excluded data. However, seeking to remove large residuals after a preliminary estimate can yield an outcome that is very different from IIS. When there is an unmodelled location shift, such that the mean of the model is shifted for a period, there need be no outliers in the mis-specified model as scaled residuals are judged relative to the inflated estimate of the equation-error standard deviation. Single-step expanding searches (e.g. step-wise regression) then have zero power to detect such a shift unless it is known. Step-indicator saturation (SIS) analysed in Castle, Doornik, Hendry, and Pretis (2015) provides a more powerful approach to capturing location shifts, but the non-orthogonality of successive steps could prove challenging computationally for Big Data. However, same-signed, similar magnitude successive impulse indicators could be combined almost costlessly to form steps.

2.4. The general unrestricted model

Combining the three extensions in Sections 2.1–2.3 creates the general unrestricted model (with the acronym GUM), where we replace the original n variables $\mathbf{z}_t$ by their orthogonal transformations $\mathbf{u}_t$, including all linear and non-linear terms with lag length s, with IIS:

$$\begin{aligned} y_t = &\sum_{i=1}^{n}\sum_{j=0}^{s}\beta_{i,j}u_{i,t-j} + \sum_{i=1}^{n}\sum_{j=0}^{s}\theta_{i,j}u_{i,t-j}^2 + \sum_{i=1}^{n}\sum_{j=0}^{s}\gamma_{i,j}u_{i,t-j}^3 \\ &+ \sum_{i=1}^{n}\sum_{j=0}^{s}\kappa_{i,j}u_{i,t-j}e^{-|u_{i,t-j}|} + \sum_{j=1}^{s}\lambda_j y_{t-j} + \sum_{i=1}^{T}\delta_i 1_{\{i=t\}} + \epsilon_t \end{aligned} \tag{3}$$

where $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$. With $N = 4n(s+1) + s$ potential regressors, plus T impulse indicators, $N > T$ will always occur, and is likely to be very large for Big Data. Special cases of (3) include:

(i) linear regressions with no lags, no shifts and no non-linearities;

(ii) factor models including just the principal component transforms $\mathbf{u}_t$ of $\mathbf{z}_t$;

(iii) dynamic linear models with lag length $s > 0$, but no shifts and no non-linearities;

(iv) non-linear models, here using polynomials, but squashing functions and threshold specifications, etc. could be tested by encompassing after selection from (3) as in Castle and Hendry (2014);

(v) models with outliers or location shifts.

Possible combinations of any or all of these are feasible depending on the data type, but viable models will result only if the GUM allows for all substantively relevant effects, as otherwise omitted influences will contaminate estimates of coefficients of included variables, usually resulting in biased and non-constant representations.

When (3) is a comprehensive specification such that a sub-set m of substantively relevant variables characterizes the empirical evidence, then the selected model should also be congruent, delivering residuals that are $\widehat{\epsilon}_t \underset{\text{app}}{\sim} \mathsf{IN}[0, \sigma_\epsilon^2]$ to a close approximation. The final model should also be able to encompass the GUM, but that is infeasible when $N > T$, and is discussed in Section 3.5.

3. The model selection problem

Once a viable GUM has been formulated, the *selection problem* requires eliminating effects that do not matter while retaining those that do. Before addressing how to undertake model selection when $N > T$, we explain three aspects in the context of $N < T$, namely "1-cut" selection in Section 3.1, multiple testing probabilities under the null in Section 3.2 and under the alternative in Section 3.3. Section 3.4 moves from 1-cut to *Autometrics*, first with $N < T$ then when $N > T$ in Section 3.5. Finally, Section 3.6 considers how theory insights can be retained costlessly while selecting over competing variables.

3.1. 1-cut model selection

Consider a correctly specified linear regression model with N accurately measured exogenous orthogonal regressors independent of the error, and constant parameters, when $T \gg N$:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2] \tag{4}$$

with $T^{-1} \sum_{t=1}^{T} z_{i,t} z_{j,t} = \lambda_{i,i}$ for $i = j$ and 0 $\forall i \neq j$. After estimation, order the N sample t^2-statistics testing H_0: $\beta_j = 0$ as:

$$\mathrm{t}^2_{(N)} \geq \mathrm{t}^2_{(N-1)} \geq \cdots \geq \mathrm{t}^2_{(1)}$$

The cut-off m between included and excluded variables is based on:

$$\mathrm{t}^2_{(m)} \geq c^2_\alpha > \mathrm{t}^2_{(m-1)}$$

Thus, variables with larger t^2-values are retained and all others eliminated. Only one decision is needed to select the model, even for $N \geq 10,000$: "repeated testing" does not occur and "goodness of fit" is never considered. The average false null retention rate can be maintained at one variable by setting $\alpha \leq 1/N$, also letting α decline as $T \to \infty$ to ensure that only relevant variables are retained asymptotically.

Nevertheless, selection *per se* matters, as only "significant" outcomes are retained. Sampling variation then entails that some irrelevant variables will be retained, and some relevant missed by chance near the selection margin. Moreover, conditional on being selected, estimates are biased away from the origin although for a normal distribution, as c_α is known, it is straightforward to bias correct (see Hendry & Krolzig, 2005). The 1-cut approach could be implemented using the $u_{i,t}$ when the interpretation of the parameters is not central. Hendry and Doornik (2014) and Castle, Doornik, and Hendry (2013), respectively, report 1-cut Monte Carlo simulations for $N = 1,000$ consistent with the above analysis, and discuss how to use principal components to detect influences with t^2-values that might be individually smaller than c_α^2, but jointly significant.

3.2. Multiple testing probabilities under the null

Selection involves many tests, and it is often claimed that this distorts inference, specifically raising the probability of false positives. When many tests are conducted and the null hypothesis is true, say N-independent t-tests at significance level α and critical value c_α, the probabilities of all 2^N null rejection outcomes for N irrelevant regressors are shown in Table 1. The first column records the events, the second their probability under the null, the third the resulting number of rejections and the final column the numerical probabilities, $p_{0.0001}$, of each outcome when N = 5,000 and $\alpha = 0.0001$.

The second column confirms that more tests increase the probability of false rejections, suggesting a need for tight significance levels; yet the average number of null variables retained on any trial is:

$$k = \sum_{i=0}^{N} i \frac{N!}{i!\,(N-i)!} \alpha^i (1-\alpha)^{N-i} = N\alpha \tag{5}$$

Despite more than 10^{1500} possible events here, 91% of the time, either none or one variable will be adventitiously retained, with $k = 0.50$. However, at $N = 500,000$ and $\alpha = 0.0001$, then $k = 50$, so spurious results will abound, suggesting the need for $\alpha \leq 0.00002$.

Fortunately, as shown in Table 2, critical values based on the normal distribution increase slowly as α decreases, and can be used for decisions during selection from a congruent GUM.

However, Table 2 relies heavily on normality. So for accurate selection based on c_α, it is important to remove outliers, asymmetry, fat tails, etc. As shown in Castle et al. (2012), IIS can do so—but would need a supercomputer for large $N > T$ using a complete search algorithm— one must exploit the mutual orthogonality of impulse indicators and principal components, selecting in sub-blocks,

Table 1. Rejection probabilities under the null

Event	Probability	Reject	$p_{0.0001}$
$\mathsf{P}\left(\lvert t_i\rvert < c_\alpha,\ \forall i = 1,\dots N\right)$	$(1-\alpha)^N$	0	0.61
$\mathsf{P}\left(\lvert t_i\rvert \geq c_\alpha \mid \lvert t_j\rvert < c_\alpha,\ \forall j \neq i\right)$	$N\alpha\,(1-\alpha)^{N-1}$	1	0.30
$\mathsf{P}\left(\lvert t_i\rvert, \lvert t_k\rvert \geq c_\alpha \mid \lvert t_j\rvert < c_\alpha,\ \forall j \neq i,k\right)$	$\frac{1}{2}N(N-1)\,\alpha^2\,(1-\alpha)^{N-2}$	2	0.08
⋮	⋮	⋮	⋮
$\mathsf{P}\left(\lvert t_i\rvert < c_\alpha \mid \lvert t_j\rvert \geq c_\alpha,\ \forall i \neq j\right)$	$N\alpha^{(N-1)}\,(1-\alpha)$	$N-1$	0
$\mathsf{P}\left(\lvert t_i\rvert \geq c_\alpha,\ \forall i = 1,\dots N\right)$	α^N	N	0

Table 2. Approximate significance levels and critical values under the null for a normal variable

α	0.05	0.01	0.005	0.0025	0.001	0.0001	2×10^{-7}
c_α	1.96	2.575	2.80	3.025	3.30	4.00	6

collecting relevant outcomes to combine at the end, possibly with a number of iterations. This generalizes the "split-half" approach of Section 2.3 to more blocks (which need not be the same size): for IIS by itself, block search is effectively a combination with "hold back" of the non-explored blocks. Section 5 explains the computational advantages for large N of reducing a 2^N general search to a series of (say) N/M searches of order 2^M using many partitions into different blocks. In principle, a different reference distribution could be used instead of normality achieved by IIS, but the non-null retention probabilities may be lower in a fat-tailed distribution.

3.3. Multiple testing probabilities under the alternative

For $N < 10,000$, large c_α can control "false positives" despite the vast number of combinations of possible test outcomes, but the cost is lower power. Under the alternative, the relationship of the non-centrality, ψ, of a t-test to c_α determines the power of the test. When a false null hypothesis with non-centrality ψ is only tested once at c_α, the approximate power is shown in the following table. Thus, there is approximately a 50–50 chance of correctly rejecting the null when $\psi = c_\alpha$, but a relatively small chance of rejecting on 4 independent tests until $\psi \gg c_\alpha$.

t-test powers				
ψ	α	c_α	$\mathsf{P}\left(\lvert t\rvert \geq c_\alpha\right)$	$\left[\mathsf{P}\left(\lvert t\rvert \geq c_\alpha\right)\right]^4$
2	0.05	2.00	0.50	0.063
2	0.01	2.61	0.26	0.005
3	0.0025	3.08	0.50	0.063
3	0.0001	4.00	0.16	0.001
4	0.01	2.61	0.91	0.686
4	0.0001	4.00	0.50	0.063
6	0.001	3.35	0.996	0.984
6	0.0001	4.00	0.976	0.907

3.4. From 1-cut to Autometrics

Autometrics is an automatic model selection programme based on a tree-search structure commencing from the GUM. Section 2 addressed the formulation of the GUM from the investigator's candidate set of variables, using automatic creation of lags, non-linear functions and indicators leading to (3), but we consider $T > N$ in this section. First, we need to introduce two concepts relevant to selection. The *Gauge*, g_α, of a selection procedure is its empirical null retention frequency, namely how often irrelevant variables are retained by search when the significance level is α. A well-calibrated procedure will have a gauge, $g_\alpha \simeq \alpha$, so by setting α appropriately, with αN small, few false positives will result. The *Potency*, p_α, is the average non-null retention frequency when selecting at significance level α, namely how often relevant variables are retained when the critical value of the test for retention is c_α. Tighter α entails a larger c_α and so leads not only to a smaller gauge, but also a lower potency. The *gauge* of a selection procedure is not the same as the size of a similar test, as insignificant irrelevant variables can sometimes be retained to offset the potential failure of a mis-specification test, so decisions may depend on "nuisance" parameters. Similarly, *potency* is not the same as power, as insignificant relevant variables are also sometimes retained and potency is the average retention rate over all relevant variables.

When $T > N$, it is feasible to test a GUM like (4) for mis-specifications, discussed in Section 4.1 below as part of resolving the model evaluation problem. If satisfied, the algorithm proceeds; if not, the significance level is tightened and selection decisions are made "as if" the models were well specified, while *Autometrics* seeks to move to a final choice that is congruent. Starting from the GUM, the variable with the smallest absolute t-ratio is removed to create a "branch" of the tree down which

each successively smallest |t| designates the next elimination. Branches are followed till all remaining variables are significant at α, then that specification is tested for congruence, and against the starting GUM using an F-test for parsimonious encompassing (see Govaerts, Hendry, & Richard 1994), also called back-testing (see Doornik, 2008). If both checks are satisfied, that model is considered to be terminal. If back-testing fails, the algorithm back-tracks till an earlier step did not fail. Returning to the GUM, the next least significant variable is eliminated and a new branch followed, till there are no initially insignificant paths to follow. Variables can be retained in the final model despite being insignificant at level α if their removal leads to either diagnostic tests or encompassing failing, emphasizing that the gauge of selection is not the same as the nominal size of the tests used. The union of all terminal models found is called the terminal GUM, from which a further tree search can be undertaken. A unique model can be selected from the resulting set of terminal undominated congruent models by a tiebreaker, such as the Schwarz (1978) criterion.

The main calibration decisions in the search algorithm are the choices of significance levels α and η for selection and mis-specification tests, and the choices of those tests. Hendry and Doornik (2014) show there is little loss from using the path-search algorithm *Autometrics* even when 1-cut is applicable, which it is not for non-orthogonal data. They show that the gauge is close to the selected α for both approaches, and the potency is near the theoretical rejection frequency for a 1-off t-test with non-centrality ψ.

3.5. More candidate variables than observations

However, the GUM in (3) is not estimable when $N > T$, so how can selection proceed? The explanation of split-half IIS in Section 2.3 points to a solution (see Doornik & Hendry, 2013), improving on a proposal for handling $N > T$ in Hendry and Krolzig (2005). Divide the complete set of variables and indicators into sub-blocks smaller than $T/2$, still setting $\alpha = 1/N$. Now select within each block, where many variables are included in each block, record which variables are relevant in each multiple block search, and collect information across blocks on which matter when others that are significant are also included.

At each stage, *Autometrics* groups variables into those already selected and those not yet selected. Variables that are not currently selected are divided into sub-blocks and the search switches between an expansion step, selecting within the not-selected sub-blocks to find significant omitted variables. Then a simplification step is performed in which the newly added set is re-selected together with those variables already included. This process is repeated until the current model is small enough for the usual algorithm, and further searches do not locate any additional significant omitted variables. It is infeasible to test the GUM for mis-specifications if $N > T$, so the programme selects as if the estimable sub-models were well specified, and seeks to move to a final choice that is congruent.

Section 5 considers some of the resulting computational issues, and suggests some shortcuts, as it is infeasible for very large N (or T) to conduct a complete tree search. Selection can be difficult when there is substantial collinearity between the variables, hence the emphasis above on formulating models with orthogonal regressors, which makes selection easier: in particular, 1-cut could be used within blocks. Because of shortcuts, the block selection algorithm is not invariant to the initial specification: adding or dropping irrelevant variables from the initial GUM can alter the block partitioning, which may change the terminal model selections.

Nevertheless, there is an alternative approach when there is some prior knowledge as to which variables might matter most, that enables a theory model to be retained and evaluated, while investigating a wide range of potential explanations.

3.6. Embedding theory insights

Hendry and Johansen (2015) propose an approach that can avoid inference costs for theory parameters by embedding the theory without search in a much more general GUM. When there are $n \ll T$ theory-relevant variables, $\mathbf{f}_t$ say, orthogonalize all the other candidate variables, $\mathbf{w}_t$, with respect to

the $\mathbf{f}_t$. This can be done in sub-groups of $w_{i,t}$ smaller than $T/2$. Now retain the $\mathbf{f}_t$ without selection, while only selecting over (possibly blocks of) the putative irrelevant variables at a stringent significance level. Under the null, because the $\mathbf{w}_t$ are orthogonal to the $\mathbf{f}_t$, whether or not they are included (or selected) has no impact on the estimated coefficients of the retained variables or their distributions.

Thus, the basic model retains the desired sub-set of $\mathbf{f}_t$ variables at every stage, and even when $N > T$, it is almost costless to check large numbers of candidate variables. In effect, every conceivable seminar question can be answered in advance while maintaining the correct theory, yet controlling the chances of false positives–which could not be done if a random number of hypotheses were tested individually.

Moreover, there is a huge benefit when the initial specification is incomplete or incorrect, but the enlarged GUM contains all the substantively relevant variables. Then an improved model will be discovered. Consequently, in this setting, the Hendry and Johansen (2015) approach provides a win–win outcome: keeping the theory model estimates unaltered when that theory model is complete and correct, and finding an improved model otherwise.

4. The model evaluation problem

Once a final selection has been made, it has to be rigorously evaluated both to check that congruence has been maintained and ensure encompassing, but also to "step outside" the information used in modelling as external validation. Section 2 discussed mis-specification testing and Section 4.1 considers the re-application of the same test statistics as diagnostic tests to check for no substantive losses of information from simplification. Finally, Section 4.2 describes an automatic test for super exogeneity that evaluates possible causal links, and uses information outside that guiding model selection.

4.1. Mis-specification testing

As noted in Section 2, a range of mis-specification tests at significance level η is applied to the feasible GUM. These include tests for normality based on skewness and kurtosis (see Doornik & Hansen, 2008), heteroskedasticity (for non-constant variance, using White, 1980), for parameter non-constancy in different sub-samples (the Chow, 1960, test), residual autocorrelation (see e.g. Godfrey, 1978), autoregressive conditional heteroskedasticity (ARCH: see Engle, 1982) and non-linearity (RESET test: see Ramsey, 1969). Parsimonious encompassing of the feasible general model by sub-models ensures no significant loss of information during reductions, and maintains the null retention frequency of *Autometrics* close to α, as shown by Doornik (2008). Both congruence and encompassing are checked by *Autometrics* when each terminal model is reached after path searches, and it backtracks to find a valid, less reduced, earlier model on that path if any test fails. This reuse of the original mis-specification tests as diagnostics to ensure congruence has been maintained and to check the validity of reductions does not affect their distributions (see Hendry & Doornik, 2014; Hendry & Krolzig, 2005).

4.2. Testing super exogeneity

Parameter invariance under regime shifts is essential to avoid mis-prediction facing policy changes. Super exogeneity combines parameter invariance with valid conditioning, so is crucial in conditional models facing shifts. The automatic IIS-based test in Hendry and Santos (2010) undertakes indicator saturation in marginal models of all the contemporaneous variables in the finally selected model of y_t, retains all the significant outcomes and tests their relevance in that conditional model. No *ex ante* knowledge of the timings or magnitudes of breaks in the marginal models is needed, and *Autometrics* is used to select other relevant variables as explained shortly. The resulting test has the correct size under the null of super exogeneity for a range of sizes of marginal-model saturation tests and has power to detect failures of super exogeneity when there are location shifts in the marginal models.

The first stage is to apply IIS to all the marginal models, which here we describe for fewer non-indicator variables than observations. Let $\mathbf{x}_t = (y_t, \mathbf{z}_t)$, then formulate:

$$\mathbf{z}_t = \boldsymbol{\pi}_0 + \sum_{j=1}^{s} \boldsymbol{\Pi}_j \mathbf{x}_{t-j} + \sum_{i=1}^{T} \boldsymbol{\rho}_{i,\alpha_1} 1_{\{t=i\}} + \mathbf{v}_t \tag{6}$$

Apply *Autometrics* to each variable in turn, recording indicators that are significant at level α_1. The second stage adds the m significant indicators to the selected conditional equation:

$$y_t = \mu_0 + \boldsymbol{\beta}' \mathbf{z}_t + \sum_{i=1}^{m} \delta_{i,\alpha_1} 1_{\{t=t_i\}} + \epsilon_t \tag{7}$$

and conducts an F-test for the joint significance of $\left(\delta_{1,\alpha_1} \ldots \delta_{m,\alpha_1}\right)$ at level α_2. Significant indicators in (7) capture shifts not explained by the included regressors, so reveal a failure of co-breaking (see Hendry & Massmann, 2007). Non-rejection shows that the selected model was not shifted by "outside" shifts, so is close to capturing causal links as $\frac{dy_t}{d\mathbf{z}_t} = \boldsymbol{\beta}$, even when $\mathbf{z}_t$ shifts substantively.

5. Computational issues

As a "proof of concept", $N > T$ could be handled by a single-step incremental search, adding variables till no further significant ones were found or a maximum model size was reached. To undertake 1-step forward searches over N variables just requires calculating N correlations, then adding variables till the next highest correlated variable is insignificant when added, so is relatively fast even for very large N. Unfortunately, such a strategy fails in many settings, most noticeably when several variables need to be added jointly for them to be significant: see Section 6.3. Thus, an improved algorithm is essential.

In comparison, *Autometrics* is a tree-search algorithm commencing from the GUM and selecting congruent, parsimonious encompassing simplifications. Implicitly, all 2^N models need checking, where N here includes T if using IIS. Even for just $N = 100$, computing 10^{10} models per second, it would take longer than the age of the universe ($\approx 2.5 \times 10^{18}$ seconds) to calculate all $2^{100} \approx 10^{30}$ possible models.

A number of features of the algorithm can affect the speed with which the tree is searched. First, relative to earlier versions like *PcGets*, speed is substantively improved by only calculating diagnostic tests for terminal models and not during path searches: the heteroskedasticity test of White (1980) is especially expensive computationally. Next, it is possible to undertake pre-search simplifications, which are single-path reductions: presently this is only implemented to remove insignificant lags, speeding up selection in dynamic equations and reducing the fraction of irrelevant lags selected. A third aspect is *bunching* groups of variables together and deleting them as a group if they are jointly insignificant (rather than eliminating one variable at a time). Next, *chopping* deletes the group from the remaining sub-paths, so is analogous to pre-search. Finally, the maximum number of terminal models is set to 20.

5.1. Block searches in Autometrics

Nevertheless, a complete tree search is infeasible for large N and $T > N$, so some shortcuts are needed, albeit at the potential risk of a sub-optimal selection. In contrast to single-step forward searches, imagine adding pairs of the two next most highly correlated variables. Extending that idea suggests adding blocks of M variables at a time. Block searches require 2^M paths, repeated N/M times to examine all the candidate variables, and if rerun R times across different sets of blocks, leads to $2^M NR/M$ searches. Thus, there is a middle way between single variable additions and adding all N variables (even when $T > N$), namely adding blocks of variables, as always occurs anyway in *Autometrics* when $N > T$. For $N = 100$, using $M = 20$ and mixing $R = 50$ times, then $2^M NR/M \approx 10^8$, which would be calculated almost instantaneously. However, for $N = 10{,}000$ and $M = 50$ repeated $R = 100$ times, $2^M NR/M \approx 10^{19}$.

Nevertheless, the crucial cost is the 2^M complete sub-tree search component. If that could be achieved instead by 1-cut applied to sets of orthogonal variables, the time taken ceases to be a serious constraint. All candidate variables are mutually orthogonal in impulse–indicator saturation, and when using $\mathbf{u}_t$ for $\mathbf{z}_t$ in (3). Then selection can be undertaken in feasible stages initially using a loose α, but tighter when sub-selections are combined for the final choice. For large T, one can exploit "hold back" ideas, so sub-samples of T are explored and later combined, with care if shifts can occur. Other possible circumventing strategies include using non-statistical criteria to restrict the variables to be selected over (e.g. delineating a sub-set of likely candidate variables), and searching sub-units separately (e.g. individuals).

6. An artificial data illustration

As an example comparing tree search with using sub-division in blocks as discussed in Section 5.1, an artificial data-set with $T = 5{,}000$ and $N = 20$ was created from the data generation process (DGP):

$$y_t = \mu_0 + \sum_{i=1}^{20} \beta_i z_{i,t} + \epsilon_t \tag{8}$$

where $\mu_0 = 0$, $\mathbf{z}_t \sim \mathsf{IN}_{20}[\mathbf{0}, \mathbf{\Omega}]$ and $\epsilon_t \sim \mathsf{IN}[0, 1]$. We consider three main sets of cases. First, in Section 6.1 all variables are orthogonal, so $\mathbf{\Omega} = \mathbf{I}_{20}$, and irrelevant, so $\beta_i = 0$, $i = 1, \ldots, 20$. Second, all variables are orthogonal, but 10 are relevant as described in Section 6.2. Third, in Section 6.3 all variables are inter-correlated and again 10 are relevant. In Section 6.4, we apply the Lasso to the third data-set.

Additional irrelevant variables are created by including lagged values of current-dated variables in the GUM. In all three sets of cases, selection is conducted without and with IIS. Selection uses $\alpha = 0.0001$ throughout, and the intercept is always retained during selection. Calculation times are based on a 64 bit dual-core 3.4 GHz desktop PC with 32 GB RAM under Windows 7. As a baseline for timings and outcomes by *Autometrics*, when $s = 1$ lag (which creates 42 regressors), selection took 0.2 seconds for the null DGP, and 0.35 and 1.84 seconds when 10 variables were relevant in the orthogonal and correlated variables cases, correctly locating their respective DGPs in all three cases.

6.1. All variables are orthogonal and irrelevant

In the first set, all variables are orthogonal with none relevant as $\beta_i = 0$, $i = 1, \ldots, 20$.

First, with $s = 1$ lag and IIS, there were 42 regressors and also 4999 impulse indicators, and selection took 4 min, and correctly selected the null model.

Second, N was increased to just over 1,000 by creating $s = 50$ lags on both the 20 variables in (8) and the dependent variable. Now selection was without IIS, so the GUM was estimable, which took 11 min using lag pre-search, and again correctly selected the null model.

Third, repeating the second experiment, but without pre-search, took just 15 seconds as no significant variables were found to start tree-search paths, so pruning almost immediately leapt to the final (correct) null model.

Fourth, with $N \approx 1{,}000$ and IIS (so lag pre-search would be ignored), with 50 blocks of approximately 100–128 took only 5.5 min. Thus, moving from $N \approx 1{,}000$ with pre-search to $N \approx 6{,}000$ using block searches halved the time taken.

6.2. All variables are orthogonal and 10 are relevant

Next, the same experiments were re-run when $n = 10$ variables were relevant. Because of the large sample size, the relevant variables' coefficients were chosen over the range 0.075–0.175, roughly corresponding to non-centralities in the range 5.3–12.4. A value of $\alpha = 0.0001$ maps to $c_\alpha \approx 4$, so there would be about a 60% chance of retaining the variable with the smallest coefficient.

For $s = 1$ lag and IIS, selection took 3.5 min and located the DGP, so the presence of the $n = 10$ relevant variables had little impact on the search time taken.

For $N \approx 1,000$ without IIS, using lag pre-search took 11 min, whereas without lag pre-search search effort of zero stopped after 18 seconds, in both cases finding the DGP exactly. However, a higher search effort without block searches took more than 9 hours on a much faster computer, which was reduced to 14.5 min by enforcing block search, leading to 11 blocks of about 100 variables each, but only retaining nine of the relevant variables.

Finally, with $N \approx 1,000$ and IIS, selection now took 12.5 min and selected 9 out of the 10 relevant variables but no irrelevant. Repeating that experiment retaining all 10 relevant variables as an example of a correct theory took roughly half the time at 6.5 min, and by construction kept the DGP. If instead 5 of the relevant and 5 irrelevant variables are retained, some of the theory is correct, but half of the assumed variables are irrelevant, the DGP is again located correctly after 8 min with the 5 retained irrelevant variables being insignificant at 5%. Thus, retaining even some of the valid theory has both substantive and computational advantages.

6.3. Correlated regressors where 10 are relevant

To illustrate the difficulties that can arise when regressors are correlated, the same basic experiment was rerun with the relevant variables' coefficients alternating in pairs over the range ± 0.25 to ± 0.35 by steps of 0.025, and the $\mathbf{z}_t$ were all inter-correlated with $\rho = 0.9$, so $\mathbf{\Omega} = (1-\rho)\mathbf{I}_{20} + \rho \boldsymbol{\iota}\boldsymbol{\iota}'$ where $\boldsymbol{\iota}'$ is a 20×1 row vector of 1s, producing non-centralities from 7.7 to 10.8. The resulting correlations of y_t with the 10 relevant variables in one experiment are recorded in Table 3.

These are very small and although T = 5,000, would not be significant at $\alpha = 0.005$. The largest correlation is for $z_{4,t}$, and the regression on that alone, mimicking a 1-step forward search procedure delivers:

$$y_t = \underset{(0.015)}{-0.0396}\, z_{4,t}$$

so would be retained only for $\alpha \geq 0.005$. The next largest correlation is with $z_{10,t}$, and when that is added:

$$y_t = \underset{(0.034)}{-0.0233}\, z_{4,t} \quad \underset{(0.034)}{-0.0181}\, z_{10,t}$$

so both variables cease to be significant at any level. However, all 10 are significant at $\alpha = 0.0001$ when jointly included, and the new DGP is located precisely when commencing from $s = 1$, that with IIS, s = 50, and that with IIS, which again found 9. Computational times were similar to, or only slightly longer than, the corresponding orthogonal regressor cases above.

Thus, the number of variables, the number that are relevant, their importance and the nature of the data all influence the time taken, as well as the mode of operation of the *Autometrics* algorithm. However, although there were roughly 10^{1750} possible spurious relationships, none was found in any of the illustrations, in almost all of which the DGP in (8) was correctly located.

6.4. Lasso outcomes for correlated artificial data

The same data was analysed using Lasso (see Efron, Hastie, Johnstone, & Tibshirani, 2004; Tibshirani, 1996) using cross-validation (see e.g. Efron & Gong, 1983). For $s = 0$ there were 20 regressors and an

Table 3. Correlations of y_t with the $z_{i,t}$

$z_{1,t}$	$z_{2,t}$	$z_{3,t}$	$z_{4,t}$	$z_{5,t}$	$z_{6,t}$	$z_{7,t}$	$z_{8,t}$	$z_{9,t}$	$z_{10,t}$
0.027	−0.029	0.033	−0.038	0.026	−0.030	0.031	−0.035	0.020	−0.037

intercept, where Lasso finds all 10 relevant variables, but also retains (i.e. does not shrink to 0) 8 of the irrelevant variables (out of 10); so Potency is 100%, but Gauge is 80%. Setting $s = 1$ so $N = 42$, Lasso again finds all 10 relevant variables, but retains 23 irrelevant variables (out of 31); so Potency is 100% and Gauge is still 74%. Increasing s to 50 lags, creating $N = 1071$, Lasso still finds all 10 relevant variables, but now retains 167 irrelevant variables (out of 1,060) making Potency still 100% with a Gauge of 15.7%. We did not apply Lasso to any of the IIS cases as it would have meant creating 5000 indicators (*Autometrics* handles IIS automatically). Throughout, computation was reasonably fast, with selection taking 53 seconds when $N = 1{,}071$.

7. Conclusions

There are many important requirements of any procedure searching for a data-based relationship using Big Data. Paramount considerations include:

- embedding all candidate variables in general initial models—which clearly favours Big Data;
- using high-quality data on all variables—which could be a worry for some analyses of Big Data;
- enforcing very tight significance levels to avoid an excess of spurious findings or false positives when N is large;
- applying an effective selection procedure, not distorted by the properties of the variables in some data-sets;
- restricting the number of variables to be selected over by searching sub-units separately or using non-statistical criteria; and
- testing for relationships being invariant to shifts in explanatory variables.

The approach described here for "fat" Big Data when $N > T$, based on *Autometrics*, tackled the *initial formulation problem* by automatic creation of additional lagged values to ensure a sequential factorization, squares, cubes and exponential functions of the principal components of the original variables to handle potential non-linearities, and impulse–indicator saturation (IIS) for any number and form of outliers and shifts. The *selection problem*—eliminating irrelevant effects while retaining variables that matter – was tackled by a block tree-search approach exploiting the orthogonality of the principal components of the variables, their non-linear functions and impulse indicators. The *evaluation problem* was resolved by checking that any chosen terminal models were well specified and encompassing, retaining any theory-based variables without search. Finally, solving the *computational problem* used a multi-path search across large numbers of candidate variables without jeopardizing the chances of locating a good model.

With an appropriate model formulation, control of the selection probabilities, stringent evaluation and efficient computation, the difficulties noted in the opening quote can be resolved to a considerable extent. Consequently, the power of Big Data can then facilitate the discovery of new knowledge in observational sciences, enhancing the virtuous circle of further theoretical insights and better empirical models.

Acknowledgements
Financial support from the Open Society Foundations and the Oxford Martin School is gratefully acknowledged, as are helpful comments from Jennifer L. Castle, Felix Pretis and Genaro Sucarrat. Numerical results are based on *Ox* and *Autometrics*, and we are indebted to Felix Pretis for the Lasso calculations.

Funding
This work was financially supported by Open Society Foundations and the Oxford Martin School.

Author details
Jurgen A. Doornik[1,2]
E-mail: jurgen.doornik@nuffield.ox.ac.uk
David F. Hendry[1,2]
E-mail:david.hendry@nuffield.ox.ac.uk
[1] Institute for New Economic Thinking, Oxford Martin School, Oxford, UK.
[2] Economics Department, Oxford University, Oxford, UK.

Citation information
Cite this article as: Statistical model selection with "Big Data", Jurgen A. Doornik & David F. Hendry, *Cogent Economics & Finance* (2015), 3: 1045216.

References

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

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2013). Model selection in equations with many 'small' effects. *Oxford Bulletin of Economics and Statistics, 75*, 6–22.
Castle, J. L., Doornik, J. A., Hendry, D. F., & Pretis, F. (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics, 3*, 240–264.
Castle, J. L., & Hendry, D. F. (2010). A low-dimension portmanteau test for non-linearity. *Journal of Econometrics, 158*, 231–245.
Castle, J. L., & Hendry, D. F. (2014). Semi-automatic non-linear model selection. In N. Haldrup, M. Meitz, & P. Saikkonen (Eds.), *Essays in nonlinear time series econometrics* (pp. 163–197). Oxford: Oxford University Press.
Castle, J. L., & Shephard, N. (Eds.). (2009). *The methodology and practice of econometrics*. Oxford: Oxford University Press.
Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica, 28*, 591–605.
Cook, S., & Hendry, D. F. (1993). The theory of reduction in econometrics. *Poznań Studies in the Philosophy of the Sciences and the Humanities, 38*, 71–100.
Doob, J. L. (1953). *Stochastic processes* (1990 ed.). New York, NY: John Wiley Classics Library.
Doornik, J. A. (2008). Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics, 70*, 915–925.
Doornik, J. A. (2009). Autometrics. J. L. Castle & N. Shephard (Eds.), *The methodology and practice of econometrics* (pp. 88–121). Oxford: Oxford University Press.
Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics, 70*, 927–939.
Doornik, J. A., & Hendry, D. F. (2013). *Empirical econometric modelling using PcGive* (7th ed., Vol. I). London: Timberlake Consultants Press.
Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37*, 36–48.
Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*, 407–499.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica, 50*, 987–1007.
Engle, R. F., Hendry, D. F., & Richard, J.-F. (1983). Exogeneity. *Econometrica, 51*, 277–304.
Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica, 46*, 1303–1313.
Govaerts, B., Hendry, D. F., & Richard, J.-F. (1994). Encompassing in stationary linear dynamic models. *Journal of Econometrics, 63*, 245–270.
Granger, C. W. J., & Teräsvirta, T. (1993). *Modelling nonlinear economic relationships*. Oxford: Oxford University Press.
Harford, T. (2014, April). Big data: Are we making a big mistake? *Financial Times*. Retrieved from http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#ixzz2xcdlP1zZ
Hendry, D. F., & Doornik, J. A. (2014). *Empirical model discovery and theory evaluation*. Cambridge, MA: MIT Press.
Hendry, D. F., & Johansen, S. (2015). Model discovery and Trygve Haavelmo's legacy. *Econometric Theory, 31*, 93–114.
Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics, 33*, 317–335. Erratum, 337–339.
Hendry, D. F., & Krolzig, H.-M. (2005). The properties of automatic GETS modelling. *Economic Journal, 115*, C32–C61.
Hendry, D. F., & Massmann, M. (2007). Co-breaking: Recent advances and a synopsis of the literature. *Journal of Business and Economic Statistics, 25*, 33–51.
Hendry, D. F., & Nielsen, B. (2007). *Econometric modeling: A likelihood approach*. Princeton, NJ: Princeton University Press.
Hendry, D. F., & Santos, C. (2010). An automatic test of super exogeneity. In M. W. Watson, T. Bollerslev, & J. Russell (Eds.), *Volatility and time series econometrics* (pp. 164–193). Oxford: Oxford University Press.
Johansen, S., & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In J. L. Castle & N. Shephard (Eds.), *The methodology and practice of econometrics* (pp. 1–36). Oxford: Oxford University Press.
Mizon, G. E., & Richard, J.-F. (1986). The encompassing principle and its application to non-nested hypothesis tests. *Econometrica, 54*, 657–678.
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.
Salkever, D. S. (1976). The use of dummy variables to compute predictions, prediction errors and confidence intervals. *Journal of Econometrics, 4*, 393–397.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B, 58*, 267–288.
Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*, 3–28.
White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*, 817–838.