Empirical Model Discovery and Theory Evaluation

David F. Hendry

Department of Economics, Oxford University

PhD Workshop, UWE, September 2010

Research jointly with Jennifer Castle, Jurgen Doornik, Søren Johansen, Grayham Mizon and Bent Nielsen

'Any sufficiently advanced technology is indistinguishable from magic.' Arthur C. Clarke, *Profiles of The Future*, 1961

Introduction

Economic theory main basis for econometric models, but: **many features of models not derivable from theory**.

Need empirical evidence on which:

[a]: **variables are actually relevant (specification),**
[b]: their lagged responses (dynamic reactions),
[c]: functional forms of relationships (non-linearities),
[d]: structural breaks & unit roots (non-stationarities),
[e]: simultaneity (or exogeneity), expectations, etc.

Almost always must be data-based on available sample: **need to discover what matters empirically**.

Theory provides an object for modelling–but:
(A) embed that object in much more **general formulation**;
(B) search for the **simplest acceptable representation**;
(C) **evaluate** the findings.
How to accomplish? And what are its properties?

Basis of approach

Data generation process (DGP):
joint density of all variables in economy

Impossible to accurately theorize about or model precisely
Too high dimensional and far too non-stationary.

Need to reduce to manageable size in 'local DGP' (LDGP):
the DGP in space of n variables $\{\mathbf{x}_t\}$ being modelled

Theory of reduction explains derivation of LDGP:
joint density $\mathsf{D}_{\mathbf{x}}(\mathbf{x}_1 \dots \mathbf{x}_T|\boldsymbol{\theta})$**.**
Acts as DGP, but 'parameter' $\boldsymbol{\theta}$ **may be time varying**

Knowing LDGP, can generate 'look alike data' for $\{\mathbf{x}_t\}$
which only deviate from actual data by unpredictable noise

Once $\{\mathbf{x}_t\}$ **chosen, cannot do better than know** $\mathsf{D}_{\mathbf{x}}(\cdot)$**–**
so the LDGP $\mathsf{D}_{\mathbf{x}}(\cdot)$ **is the target for model selection:**
need to relate theory model to that target.

Formulating a 'good' LDGP

Choice of n variables, $\{\mathbf{x}_t\}$, to analyze is fundamental: determines the modelling target LDGP, $\mathsf{D}_{\mathbf{x}}(\cdot)$, and its properties.

Prior reasoning, theoretical analysis, previous evidence, historical and institutional knowledge all important.

Should be $90+\%$ of effort in an empirical analysis.

Aim to avoid complicated and non-constant LDGPs. Crucial not to omit substantively important variables: small set $\{\mathbf{x}_t\}$ more likely to do so.

Given $\{\mathbf{x}_t\}$, have defined the target $\mathsf{D}_{\mathbf{x}}(\cdot)$.

Now embed that target in a general model formulation, which also retains, but does not impose, the theory-based variables.

Empirical implementation

Sample of T observations, $\{\mathbf{x}_t\} = \{\mathbf{y}_t, \mathbf{z}_t\}$:
but no theory specification of **unit of time**,
observations may be contaminated (**measurement errors**),
underlying processes **integrated**,
abrupt unanticipated shifts induce various forms of **breaks**.

All these aspects must be discovered empirically:
model selection is inevitable and ubiquitous.
So how to utilize economic analyses efficiently if cannot impose theory empirically?
Answer: embed theory specification in vastly more general empirical formulation.
Retain theory formulation in congruent, parsimonious encompassing model, seeking parameters invariant to relevant policies.
Approach embodied in *Autometrics*: see Doornik (2009)
So let's perform some magic.

Route map

(1) **Discovery in general**

(2) Automatic model extension

(3) Automatic model selection

(4) Automatic model evaluation

(5) Embedding theory models

(6) Excess numbers of variables $N > T$

(7) Non-invariance of NKPCs

Conclusions

Discovery in economics

Discoveries in economics mainly from theory.

But all economic theories are:

(a) incomplete; (b) incorrect; and (c) mutable.

(a) Need strong *ceteris paribus* assumptions:
inappropriate in a non-stationary, evolving world.

(b) Consider an economic analysis which suggests:

$$\mathbf{y} = \mathbf{f}(\mathbf{z}) \tag{1}$$

where (k) $\mathbf{y}$ depend on n 'explanatory' variables $\mathbf{z}$.

Form of $\mathbf{f}(\cdot)$ in (1) depends on:
utility or loss functions of agents,
constraints they face, & information they possess.

Analyses arbitrarily assume: forms for $\mathbf{f}(\cdot)$, that $\mathbf{f}(\cdot)$ is constant, that only $\mathbf{z}$ matters, & that the $\mathbf{z}$s are 'exogenous'.

Yet must aggregate across heterogeneous individuals whose endowments shift over time, often abruptly.

Theory evolves

(c) Economic analyses have changed the world, and our understanding: from the 'invisible hand' in Adam Smith's *Theory of Moral Sentiments* (1759, p.350) onwards, theory has progressed dramatically–
key insights into option pricing, auctions and contracts, principal-agent and game theories, trust and moral hazard, asymmetric information, institutions:
major impacts on market functioning, industrial,
and even political, organization.

But imagine imposing 1900's economic theory in empirical research today.

Much past applied econometrics research is forgotten: discard the economic theory that it 'quantified' and you discard the associated empirical evidence.

Hence fads & fashions, 'cycles' and 'schools' in economics.

Discovery in general

Discovery: learning something previously unknown.
Cannot know how to discover what is not known–
unlikely there is a 'best' way of doing so.

Many empirical discoveries have element of chance:
luck: Fleming–penicillin from a dirty petrie dish
serendipity: Becquerel–discovery of radioactivity
'natural experiment': Dicke–role of gluten in celiac disease
trial and error: Edison–incandescent lamp
brilliant intuition: Faraday–dynamo from electric shock
false theories: Kepler–regular solids for planetary laws
valid theories: Pasteur–germs not spontaneous generation
systematic exploration: Lavoisier–oxygen not phlogiston
careful observation: Harvey–circulation of blood
new instruments: Galileo–moons around Jupiter
self testing: Marshall–ulcers caused by *Helicobacter pylori*.

Theoretical discoveries

Theoretical discoveries also important.

Classic examples include:

uniform motion: Galileo Galilei;

universal gravitation: Issac Newton;

electro-magnetic spectrum: Clerk Maxwell;

black-body radiation: Max Planck;

relativity: Albert Einstein;

quantum theory: Niels Bohr;

positron: Paul Dirac;

quark: Murray Gell-Mann.

Some ‘evidence based’; some ‘thought experiments’.
All required later independent evaluation.

Discovery and evaluation

Science is both inductive and deductive.

Must distinguish between:
context of discovery–where 'anything goes', and
context of evaluation–rigorous attempts to refute.

However a discovery made, needs a warrant that it is 'real'.
Methods of evaluation are subject-specific:
economics requires a theoretical interpretation consistent with 'mainstream theory'.

Accumulation and consolidation of evidence crucial:
data reduction a key attribute of science (think $E = mc^2$).

Common aspects of discovery

Seven aspects in common to above examples of discovery.

First, ***theoretical context***, or framework of ideas.

Second, going ***outside*** existing state of knowledge.

Third, ***searching*** for something.

Fourth, ***recognition*** of significance of what is found.

Fifth, ***quantification*** of what is found.

Sixth, ***evaluating*** discovery to ascertain its 'reality'.

Seven, ***parsimoniously summarize*** information acquired.

But science perforce is simple to general–
a slow and uncertain route to new knowledge.
Econometrics discovery need not be....

Classical econometrics: covert discovery

Postulate:

$$y_t = \boldsymbol{\beta}' \mathbf{z}_t + \epsilon_t, \quad t = 1, \ldots, T \tag{2}$$

Aim to obtain 'best' estimate of the constant parameters $\boldsymbol{\beta}$, given the n correct variables, $\mathbf{z}$, 'independent' of $\{\epsilon_t\}$ and uncontaminated observations, $\mathcal{T}$, with $\epsilon_t \sim$IID$\left[0, \sigma_\epsilon^2\right]$.

Many tests to 'discover' departures from assumptions of (2), followed by recipes for 'fixing' them–
covert and unstructured empirical model discovery.

Model selection: discovering the 'best' model.
Starts from (2) assuming N 'correct' initial $\mathbf{z}$, accurate data over $\mathcal{T}$, constant $\boldsymbol{\beta}$ and valid conditioning.
Aim to 'discover' the subset of relevant variables, $\mathbf{z}_t^*$.

Selected 'best model' may be poor approximation to LDGP: almost never evaluated.

Robust statistics: discovering the best sample

Same start (2), but aim to find a **'robust' estimate** of a constant β by selecting over $\mathcal{T}$.

Worry about data contamination and outliers, so select sample, $\mathcal{T}^*$, where outliers least in evidence, given correct set of relevant variables $\mathbf{z}$.

All other difficulties still need separate tests, and must be fixed if found.

$\mathbf{z}$ rarely selected jointly with $\mathcal{T}^*$, so assumes $\mathbf{z} = \mathbf{z}^*$.

Similarly for non-parametric methods:
aim to discover 'best' functional form or distribution, assuming correct $\mathbf{z}$, no data contamination, constant β, etc., all rarely checked.

Each assumes away what the others tackle.

Automatic empirical model discovery

Need to tackle them all jointly.
Re-frame empirical modelling as discovery process: part of a progressive research strategy.

Starting from T observations on $N > n$ variables $\mathbf{z}$, aim to find $\boldsymbol{\beta}^*$ for s lagged functions $\mathbf{g}(\mathbf{z}_t^*) \dots \mathbf{g}(\mathbf{z}_{t-s}^*)$ of a subset of n variables $\mathbf{z}^*$, jointly with $\mathcal{T}^*$ and $\{\mathbf{1}_{\{t=t_i\}}\}$– indicators for breaks, outliers etc.

Embeds initial economic analysis $\mathbf{y} = \mathbf{f}(\mathbf{z})$, but in a much more general initial model.

Globally, learning must be simple to general; but locally, need not be.

General approach explained in Castle, Doornik and Hendry (2010).

Implications for automatic methods

Same seven stages as for discovery in general.

First, **theoretical derivation** of the relevant set $\mathbf{x}$.

Second, going **outside** current view by **automatic creation of a general model** from $\mathbf{x}$ embedding $\mathbf{y} = \mathbf{f}(\mathbf{z})$.

Third, **search** by **automatic selection** to find viable representations: too large for manual labor.

Fourth, criteria to **recognize** when search is completed: **congruent parsimonious-encompassing model**.

Fifth, **quantification** of the outcome: translated into **unbiasedly estimating the resulting model**.

Sixth, **evaluate** discovery to check its 'reality: **new data, new tests or new procedures**.
Can also evaluate the selection process itself.

Seventh, **summarize** vast information set in **parsimonious but undominated model**.

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) Automatic model selection

(4) Automatic model evaluation

(5) Embedding theory models

(6) Excess numbers of variables $N > T$

(7) Non-invariance of NKPCs

Conclusions

Extensions outside standard information

Extensions determine how well LDGP is approximated

Create three extensions automatically:

(i) lag formulation to implement **sequential factorization**;

(ii) functional form transformations for **non-linearity**;

(iii) impulse-indicator saturation (IIS) for **parameter non-constancy and data contamination**.

(i) Create s lags $\mathbf{x}_t \dots \mathbf{x}_{t-s}$ to formulate general linear model:

$$y_t = \beta_0 + \sum_{i=1}^{s} \lambda_i y_{t-i} + \sum_{i=1}^{r} \sum_{j=0}^{s} \beta_{i,j} z_{i,t-j} + \epsilon_t \tag{3}$$

$\mathbf{x}_t = (y_t, \mathbf{z}_t)$ could also be modelled as a system:

$$\mathbf{x}_t = \boldsymbol{\gamma} + \sum_{j=1}^{s} \boldsymbol{\Gamma}_j \mathbf{x}_{t-j} + \boldsymbol{\epsilon}_t \tag{4}$$

We focus on single equations, but systems can be handled.

Automatic non-linear extensions

Test for non-linearity in general linear model by low-dimensional portmanteau test in Castle and Hendry (2010b) (cubics of **principal components** $\mathbf{w}_t$ of the $\mathbf{z}_t$).

(ii) If reject, create $\mathbf{g}(\mathbf{w}_t)$, otherwise $\mathbf{g}(\mathbf{z}_t) = \mathbf{z}_t$: presently, implemented general cubics with exponential functions.

Number of potential regressors for cubic polynomials is:

$$M_K = K\,(K+1)\,(K+5)\,/6.$$

Explosion in number of terms as $K = r \times (s+1)$ increases:

K	1	2	3	4	5	10	15	20	30	40
M_K	3	9	19	30	55	285	679	1539	5455	12300

Quickly reach huge M_K: **but only** $3K$ **if use** $w^k_{i,t-j}$.

(Investigating **squashing functions**, to better approximate non-linearity in economics, suggested by Hal White)

Impulse-indicator saturation

(iii) To tackle multiple breaks & data contamination (outliers), add T impulse indicators to candidates for T observations.

Consider $y_i \sim$ IID $\left[\mu, \sigma_\epsilon^2\right]$ for $i = 1, \ldots, T$

μ is parameter of interest

Uncertain of outliers, so add T indicators $\mathbf{1}_{\{t=t_i\}}$

to set of candidate regressors.

First, include half of indicators, record significant:

just 'dummying out' $T/2$ observations for estimating μ

Then omit, include other half, record again.

Combine sub-sample indicators, & select significant.

αT indicators selected on average at significance level α

Feasible 'split-sample' impulse-indicator saturation (IIS) algorithm: see Hendry, Johansen and Santos (2008)

Dynamic generalizations

Johansen and Nielsen (2009) extend IIS to both stationary and unit-root autoregressions

When distribution is symmetric, adding T impulse-indicators to a regression with n variables, coefficient $\boldsymbol{\beta}$ (not selected) and second moment $\boldsymbol{\Sigma}$:

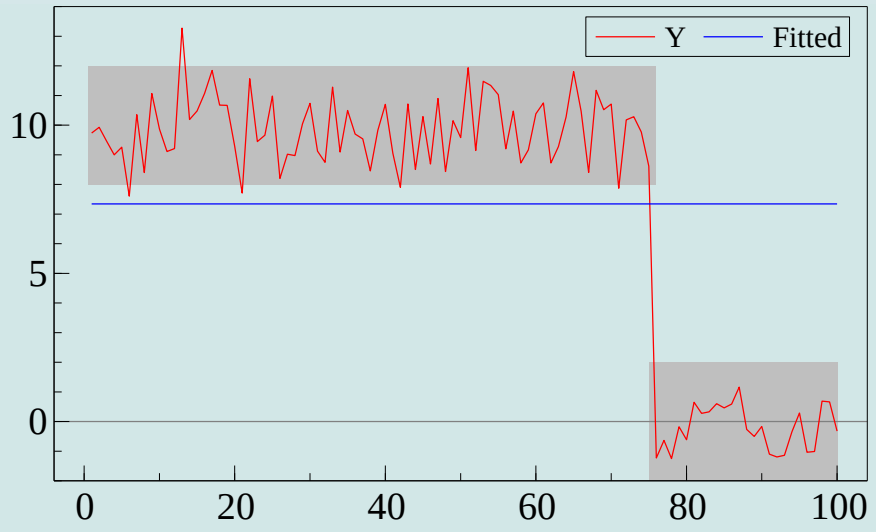
$$T^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_n\left[\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega}_\beta\right]$$

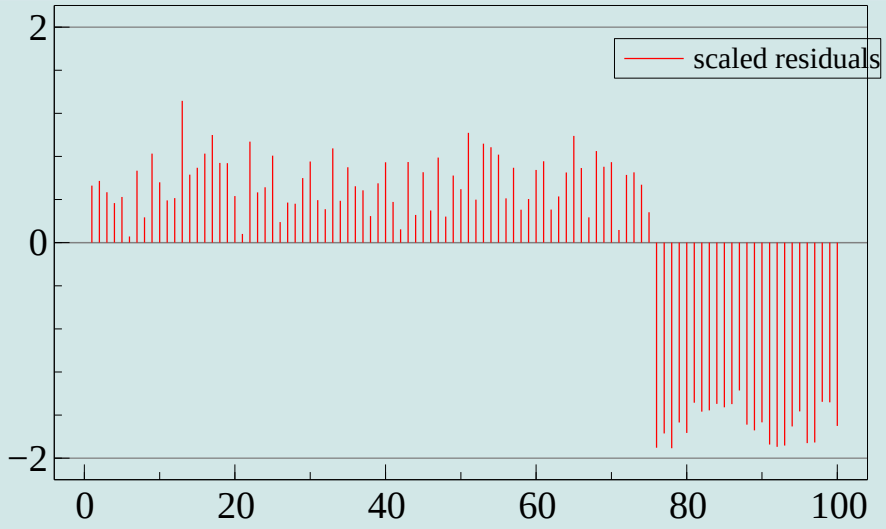
Efficiency of IIS estimator $\widetilde{\boldsymbol{\beta}}$ with respect to OLS $\widehat{\boldsymbol{\beta}}$ measured by $\boldsymbol{\Omega}_\beta$ depends on c_α and distribution

Must lose efficiency under null: but small loss αT– only 1% at $\alpha = 1/T$ if $T = 100$, despite T extra candidates.

Potential for major gain under alternatives of breaks and/or data contamination: variant of robust estimation

Structural break example

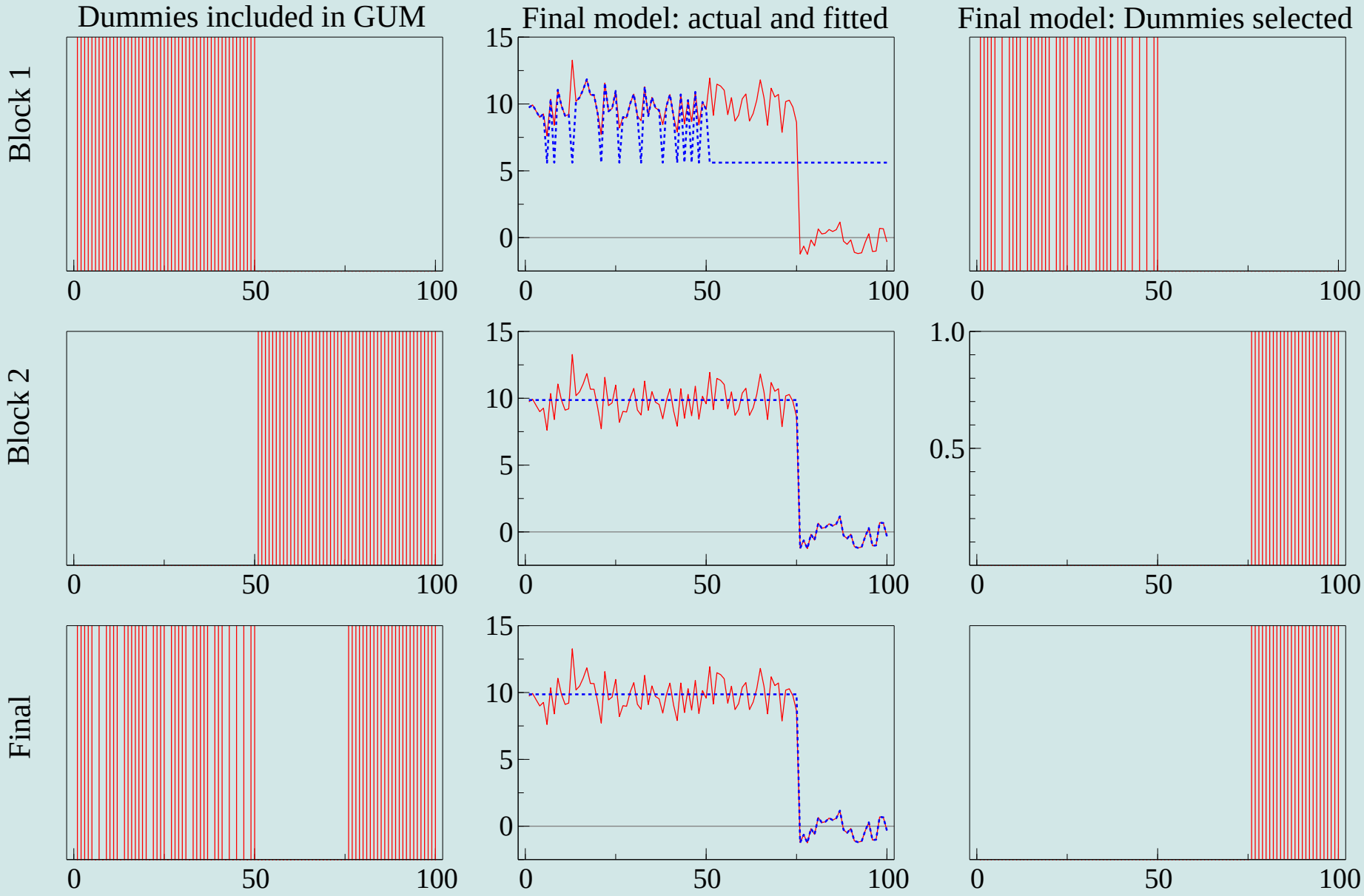




- Size of the break is **10 standard errors** at $0.75T$
- There are **no outliers** in this mis-specified model as all residuals $\in [-2, 2]$ SDs:
  outliers $\neq$ structural breaks
- step-wise regression has **zero power**

Let's see what Autometrics reports

‘Split-sample’ search in IIS



Specification of GUM

Most major formulation decisions now made:
which r variables ($\mathbf{w}_t$, after transforming $\mathbf{z}_t$);
their lag lengths (s);
functional forms (cubics);
structural breaks (any number, anywhere).
Leads to general unrestricted model (GUM):

$$y_t = \sum_{i=1}^{r}\sum_{j=0}^{s}\beta_{i,j}z_{i,t-j} + \sum_{i=1}^{r}\sum_{j=0}^{s}\kappa_{i,j}w_{i,t-j} + \sum_{i=1}^{r}\sum_{j=0}^{s}\theta_{i,j}w^2_{i,t-j}$$
$$+ \sum_{i=1}^{r}\sum_{j=0}^{s}\gamma_{i,j}w^3_{i,t-j} + \sum_{j=1}^{s}\lambda_j y_{t-j} + \sum_{i=1}^{T}\delta_i 1_{\{i=t\}} + \epsilon_t$$

$K = 4r(s+1)+s$ potential regressors, plus T indicators:
close to what I showed live earlier.
Bound to have $N > T$: consider exogeneity later.

Understanding model selection

Consider a perfectly orthogonal regression model:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \tag{5}$$

$\mathsf{E}[z_{i,t} z_{j,t}] = \lambda_{i,i}$ for $i = j$ & $0 \; \forall i \neq j$, $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$ and $T >> N$.

Order the N sample t^2-statistics testing H_0: $\beta_j = 0$:

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

Cut-off m between included and excluded variables is:

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

Larger values retained: all others eliminated.

Only one decision needed even for $N \geq 1000$:
'repeated testing' does not occur, and
'goodness of fit' is never considered.

Maintain average false null retention at **one variable** by $\alpha \leq 1/N$, with α declining as $T \to \infty$

Repeated testing

Does repeated testing distort selection?

- (a) Severe illness:
  more tests **increase** probability of **correct diagnosis**.

Repeated testing

Does repeated testing distort selection?

- (a) Severe illness:
  more tests **increase** probability of **correct diagnosis**.
- (b) Mis-specification tests:
  if r independent tests τ_j conducted under null
  for small significance level η (critical value c_η):

$$\mathsf{P}(|\tau_j| < c_\eta \mid j = 1, \ldots, r) = (1-\eta)^r \simeq 1 - r\eta.$$

  More tests **increase** probability of **false rejection**.
  Suggests significance level η of 1% or tighter.

Repeated testing

Does repeated testing distort selection?

- (a) Severe illness:
  more tests **increase** probability of **correct diagnosis**.
- (b) Mis-specification tests:
  if r independent tests τ_j conducted under null
  for small significance level η (critical value c_η):

$$\mathsf{P}(|\tau_j| < c_\eta \mid j = 1, \ldots, r) = (1-\eta)^r \simeq 1 - r\eta.$$

  More tests **increase** probability of **false rejection**.
  Suggests significance level η of 1% or tighter.
- (c) Repeated diagnostic tests: **probabilities unaltered**.
  Conclude: no generic answer.

Interpretation

Path search gives impression of 'repeated testing'.
Confused with selecting from 2^N possible **models**
(here $2^{1000} = 10^{301}$, an impossible task).
We are selecting **variables**, not models, & only N variables.

But selection matters, as only retain 'significant' outcomes.
Sampling variation also entails retain irrelevant, or miss relevant, by chance near selection margin.

Conditional on selecting, estimates biased away from origin: **but can bias correct as know** c_α.

Small efficiency cost under null for examining many candidate regressors, even $N >> T$**.**

Almost as good as commencing from LDGP at same c_α.

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) Automatic model evaluation

(5) Embedding theory models

(6) Excess numbers of variables $N > T$

(7) Non-invariance of NKPCs

Conclusions

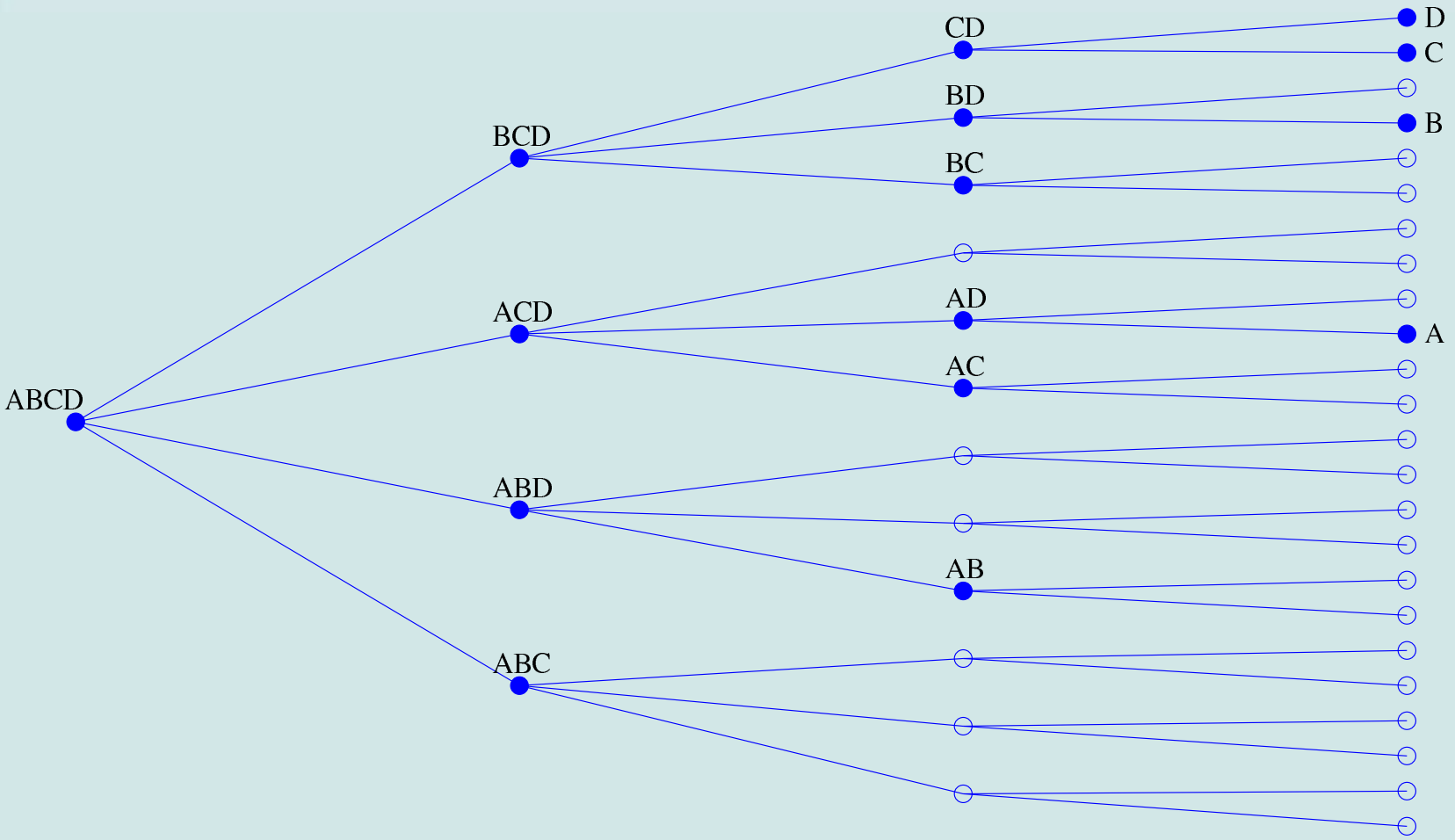
Autometrics improves on previous algorithms

- **Search paths:** Autometrics examines whole search space; discards irrelevant routes systematically.
- **Likelihood-based:** Autometrics implemented in likelihood framework.
- **Efficiency:** Autometrics improves computational efficiency: avoids repeated estimation & diagnostic testing, remembers terminal models.
- **Structured:** Autometrics separates estimation criterion, search algorithm, evaluation, & termination decision.
- **Generality:** Autometrics can handle $N > T$.

If GUM is congruent, so are all terminals: undominated, mutually-encompassing representations.

If several terminal models, all reported: can combine, or one selected (by, e.g., Schwarz, 1978, criterion).

Autometrics *tree search*



Search follows branches till no insignificant variables; tests for congruence and parsimonious encompassing; backtracks if either fails, till first non-rejection found.

Selecting by Autometrics

Even when 1-cut applicable, little loss, and often a gain, from using path-search algorithm *Autometrics*.

***Autometrics* applicable to non-orthogonal problems, and $N > T$.**

'*Gauge*' (average retention rate of irrelevant variables) close to α.

'*Potency*' (average retention rate of relevant variables) near theory value for a 1-off test.

Goodness-of-fit not directly used to select models & no attempt to 'prove' that a given set of variables matters, but choice of c_α affects R^2 and n through retention by $|t_{(n)}| \geq c_\alpha$.

Conclude: 'repeated testing' is not a concern.

Selecting non-linear models

Transpires there are four major sub-problems:

(A) **specify general form of non-linearity**

(B) **non-normality: non-linear functions capture outliers**

(C) **excess numbers of irrelevant variables**

(D) **potentially more variables than observations**

Have solutions to all four sub-problems:

(A) **investigator's preferred general function**, simplified by encompassing tests against specific (ogive) forms

(B) remove outliers by **IIS**

(C) **super-conservative** selection strategy

(D) multi-stage **'combinatorial selection'** for $N > T$

Automatic algorithm for up to **cubic polynomials** with polynomials times exponentials in Castle and Hendry (2010a).

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model evaluation**

(5) Embedding theory models

(6) Excess numbers of variables $N > T$

(7) Non-invariance of NKPCs

Conclusions

Role of mis-specification testing

Under null of congruent GUM, Figure 35 compares 'gauges' for *Autometrics* with diagnostic checking **on** vs. **off**:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \quad \text{for} \quad \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2] \tag{6}$$

$T = 100$, $n = 1, \ldots, 10 = N$; $\beta_k = 0$ for $k > n$; $\mathsf{R}^2 = 0.9$.

'*Gauge*' is average retention rate of irrelevant variables (should be close to α).

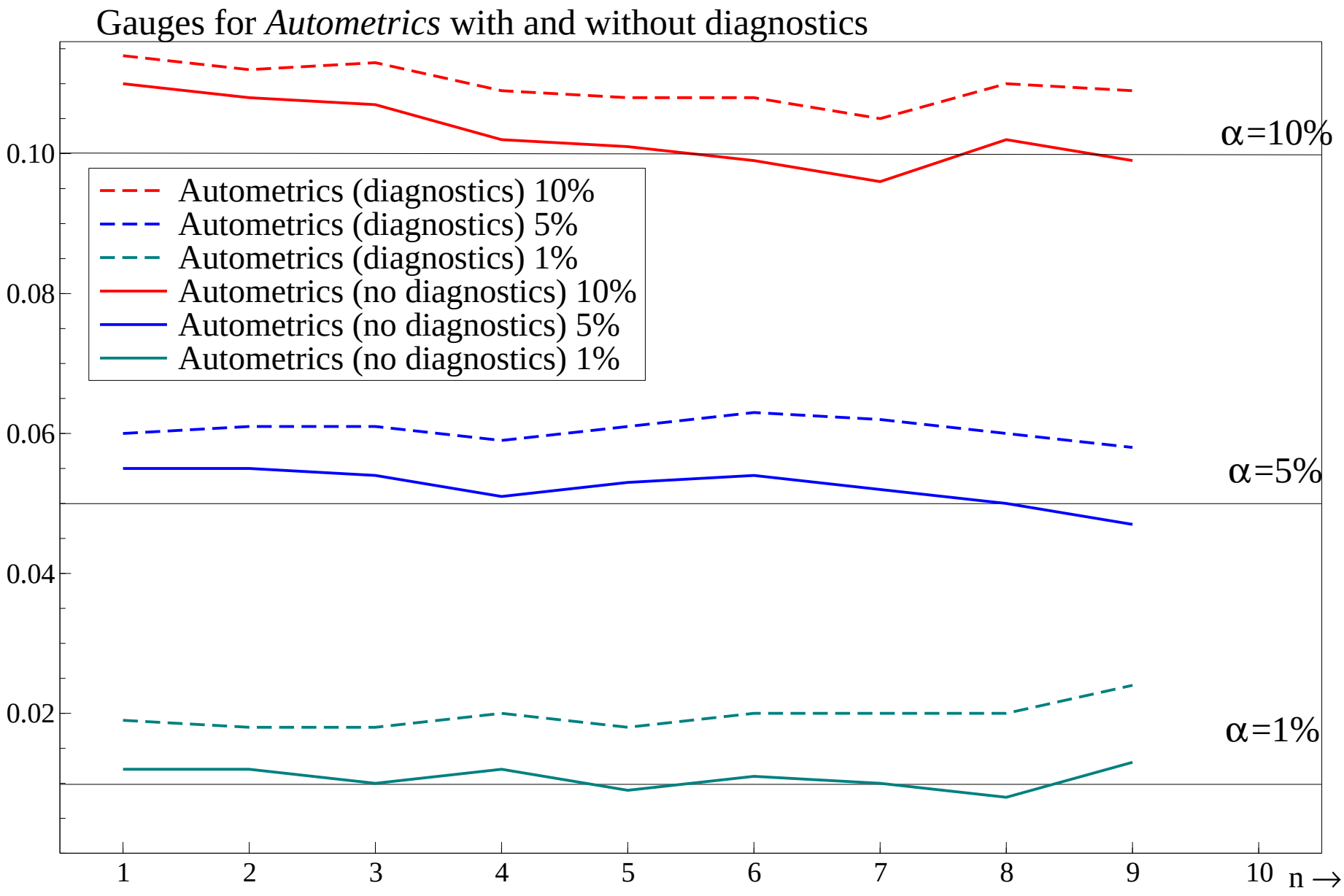
'*Potency*' is average retention rate of relevant variables (should be near theory power for a 1-off test).

Gauge is close to α if diagnostic tests **not** checked.

Gauge is larger than α with diagnostics **on**, when checking to ensure a congruent reduction.

Difference seems due to retaining insignificant irrelevant variables which proxy chance departures from null of mis-specification tests.

Gauges with diagnostic tests off & on



Role of encompassing

Variables removed only when new model is a valid reduction of GUM.

Reduction fails if result does not parsimoniously encompass GUM at c_α: (see Hendry, 1995, §14.6).

If so, variable retained despite being insignificant on $\mathtt{t}$-test, as in Doornik (2008).

Autometrics **without encompassing loses both gauge and potency**:
gauge is the average retention rate of irrelevant variables;
potency is average retention rate of relevant variables

Autometrics **with encompassing is well behaved**:

gauge is close to nominal rejection frequency α.
potency is close to theory maximum of 1-off $\mathtt{t}$-test.

Simulating Autometrics on Hoover–Perez

Hoover and Perez (1999) experiments:

HP7 $y_{7,t} = 0.75y_{7,t-1} + 1.33x_{11,t} - 0.9975x_{11,t-1} + 6.44u_t \quad R^2 = 0.58$

HP8 $y_{8,t} = 0.75y_{8,t-1} - 0.046x_{3,t} + 0.0345x_{3,t-1} + 0.073\lambda u_t \quad R^2 = 0.93$

where $u_t \sim \mathsf{IN}[0, 1]$; $x_{i,t-j}$ are US macro data

The GUM has 3 DGP variables plus 37 irrelevant.

Table 1 shows results for range of values of λ and α in HP8 (they set $\lambda = 1$).

Later consider 141 irrelevant, larger than $T = 139$.

Simulations for encompassing

		Autometrics with encompassing		Autometrics no encompassing	
α	λ	Gauge	Potency	Gauge	Potency
0.1	50	0.093	0.441	0.056	0.402
0.05	50	0.055	0.405	0.021	0.364
0.01	50	0.014	0.357	0.002	0.337
0.1	10	0.096	0.940	0.062	0.904
0.05	10	0.057	0.935	0.031	0.832
0.01	10	0.017	0.895	0.002	0.630
0.1	1	0.093	1.000	0.050	1.000
0.05	1	0.055	1.000	0.019	1.000
0.01	1	0.014	1.000	0.002	0.999

Table 1: HP8 with $M = 10000$ and $T = 139$.

Testing super exogeneity

Parameter invariance essential in policy models:
else mis-predict under regime shifts.
Super exogeneity combines parameter invariance with valid conditioning so crucial for economic policy.

New automatic test in Hendry and Santos (2010):
impulse-indicator saturation in marginal models,
retain all significant outcomes and
test their relevance in conditional model

No *ex ante* knowledge of timing or magnitudes of breaks:
need not know DGP of marginal variables

Test has correct size under null of super exogeneity
for a range of sizes of marginal-model saturation tests

Power to detect failures of super exogeneity when location shifts in marginal models

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model evaluation**

(5) **Embedding theory models**

(6) Excess numbers of variables $N > T$

(7) Non-invariance of NKPCs

Conclusions

Retaining economic theory insights

Approach is not atheoretic.

Theory formulations should be embedded in GUM, can be retained without selection.

Call such imposition 'forcing' variables–ensures they are retained, but does not guarantee they will be significant.

Can also ensure theory-derived **signs** of long-run relation maintained, if not significantly rejected by the evidence.

But much observed data variability in economics is due to features absent from most economic theories: which empirical models must handle.

Extension of LDGP candidates, $\mathbf{x}_t$, in GUM allows theory formulation as special case, yet protects against contaminating influences (like outliers) absent from theory.

'Extras' can be selected at tight significance levels.

Four possible economic theory outcomes

1] **Theory exactly correct**:
all aspects significant with anticipated signs,
no other variables kept.

2] **Theory only part of explanation**:
all aspects significant with anticipated signs,
but other variables also kept as substantively relevant.

3] **Theory partially correct**:
only some aspects significant with anticipated signs,
and other variables also kept as substantively relevant.

4] **Theory not correct**:
no aspects significant and
other variables do all explanation.

Consider these in turn: see Hendry and Johansen (2010).

Theory exactly correct

Theory specifies correct set of n relevant variables, $\mathbf{z}_t$, with parameters $\boldsymbol{\beta}$:

$$y_t = \boldsymbol{\beta}' \mathbf{z}_t + \epsilon_t \tag{7}$$

where $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$, independently of $\mathbf{z}_t$. Then:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \sum_{t=1}^{T} \mathbf{z}_t \epsilon_t \sim \mathsf{N}_n \left[\mathbf{0}, \sigma_\epsilon^2 \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \right] \tag{8}$$

Next, $\mathbf{z}_t$ 'forced' to be retained during model selection over second set of k irrelevant candidate variables, $\mathbf{w}_t$, with coefficients $\boldsymbol{\gamma} = \mathbf{0}$ when $(k+n) << T$, so GUM is:

$$y_t = \boldsymbol{\beta}' \mathbf{z}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \nu_t \tag{9}$$

Orthogonalize $\mathbf{z}_t$ and $\mathbf{w}_t$ by:

$$\mathbf{w}_t = \widehat{\boldsymbol{\Gamma}} \mathbf{z}_t + \mathbf{u}_t \tag{10}$$

Then as $\boldsymbol{\gamma} = \mathbf{0}$:

$$y_t = \boldsymbol{\beta}' \mathbf{z}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \nu_t = \boldsymbol{\beta}' \mathbf{z}_t + \boldsymbol{\gamma}' \mathbf{u}_t + \nu_t \tag{11}$$

Distributions of forced estimates

Consequently:

$$\begin{pmatrix} \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widetilde{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' & \sum_{t=1}^{T} \mathbf{z}_t \mathbf{u}_t' \\ \sum_{t=1}^{T} \mathbf{u}_t \mathbf{z}_t' & \sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^{T} \mathbf{z}_t \nu_t \\ \sum_{t=1}^{T} \mathbf{u}_t \nu_t \end{pmatrix}$$

$$\sim \mathsf{N}_{n+k}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \right)^{-1} \end{pmatrix} \right] \quad (12)$$

as $\sum_{t=1}^{T} \mathbf{z}_t \mathbf{u}_t' \simeq 0$, so distribution of $\widetilde{\boldsymbol{\beta}}$ in (12) **identical** to that of $\widehat{\boldsymbol{\beta}}$ in (8): **unaffected** by model selection.

Only costs of selection are:
(a) chance retentions of some $\mathbf{u}_t$ from selection; and
(b) impact on **estimated** distribution of $\widetilde{\boldsymbol{\beta}}$ through $\widetilde{\sigma}_\epsilon^2$.
Can be offset by bias correction.

Theory only part of explanation

Different when **theory model is only part of explanation**: defined as all aspects significant with anticipated signs, but other variables also kept as substantively relevant.

Two distinct forms of under-specification:
a] omitting relevant functions or lags of variables in LDGP; avoided by sufficiently general initial model:
b] omitting relevant variables, $\mathbf{w}_t$, from the DGP; induces less useful LDGP–hard to avoid if $\mathbf{w}_t$ unknown.
In a], $\boldsymbol{\gamma} \neq \mathbf{0}$, as $\mathbf{z}_t$ and $\mathbf{u}_t$ orthogonal in (13), coefficient of former is $\boldsymbol{\beta} + \boldsymbol{\gamma}' \widehat{\boldsymbol{\Gamma}}$, which is estimated if (7) is simply fitted to the data: but may be significant with anticipated signs.

In b], when (9) nests LDGP, but $\mathbf{w}_t$ omitted from DGP, selection can substantively improve the final model: (see Castle and Hendry, 2010c), as we will show.

Some of theory part of explanation

Next, when the theory is only partially correct:
some aspects significant with anticipated signs,
but other aspects not significant, or 'wrong' signed,
with other variables also kept as substantively relevant.

Under alternative, $\boldsymbol{\gamma} \neq \mathbf{0}$, estimating (7) will result in biased, inefficient, possibly non-constant, estimates as:

$$y_t = \boldsymbol{\beta}'\mathbf{z}_t + \boldsymbol{\gamma}'\left(\widehat{\boldsymbol{\Gamma}}\mathbf{z}_t + \mathbf{u}_t\right) + \nu_t = \left(\boldsymbol{\beta} + \boldsymbol{\gamma}'\widehat{\boldsymbol{\Gamma}}\right)'\mathbf{z}_t + \boldsymbol{\gamma}'\mathbf{u}_t + \nu_t \quad (13)$$

Now forcing $\mathbf{z}_t$ when selecting from (13) will deliver an incorrect estimate of $\boldsymbol{\beta}$, but some of the $\mathbf{u}_t$ will be correctly retained, so an implied estimate of $\boldsymbol{\beta}$ can be derived from $\boldsymbol{\beta} + \boldsymbol{\gamma}'\widehat{\boldsymbol{\Gamma}}$, $\widetilde{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$. A better estimate of $\widetilde{\sigma}^2_\nu$ should result.

Selection can also help when relevant variables, $\mathbf{w}_t$, omitted from DGP and breaks occur.

Theory not part of explanation

Finally, theory is now completely incorrect:
no aspects significant and other variables do all explanation.

Despite forcing $\mathbf{z}_t$, $\boldsymbol{\beta} = \mathbf{0}$, but interpretation awkward as coefficient of $\mathbf{z}_t$ is $\boldsymbol{\gamma}'\widehat{\boldsymbol{\Gamma}}$.

Win-win situation: theory kept if valid and complete; yet learn when it is not correct–
empirical model discovery embedding theory evaluation.

Interesting case is when $N > T$ for N candidates,
so can automatic model selection work then?

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model evaluation**

(5) **Embedding theory models**

(6) **Excess numbers of variables** $N > T$

(7) Non-invariance of NKPCs

Conclusions

As many candidate variables as observations

Analytic approach to understanding IIS applies for $N = T$ IID mutually orthogonal candidate regressors under the null.

Add first $N/2$ and select at significance level $\alpha = 1/T = 1/N$. Record which were significant, and drop all.

Now add second block of $N/2$, again select at significance level $\alpha = 1/N$, and record which are significant.

Finally, combine recorded variables from the two stages (if any), and select again at significance level $\alpha = 1/N$.

At both sub-steps, on average $\alpha N/2 = 1/2$ a variable will be retained by chance, so on average $\alpha N = 1$ from the combined stage.

Again 99% efficient under the null at eliminating irrelevant variables–lose one degree of freedom on average.

More candidate variables than observations

If also have relevant variables to be retained, and $N > T$, orthogonalize them with respect to the rest.

As $N > T$, divide in more sub-blocks, setting $\alpha = 1/N$.

Basic model retains desired sub-set of n variables at every stage, and only selects over putative irrelevant variables at stringent significance level:
under the null, has no impact on estimated coefficients of relevant variables, or their distributions.

Thus, almost costless to check even large numbers of candidate variables:
huge benefits if initial specification incorrect but enlarged GUM nests LDGP.

IIS for multiple breaks

DGP: D1: $y_{1,t} = \gamma\left(I_{T-19} + \cdots + I_T\right) + u_t, \quad u_t \sim \mathsf{N}[0,1]$
D2: $y_{3,t} = \gamma\left(I_1 + I_6 + I_{11} + \cdots\right) + u_t, u_t \sim \mathsf{N}[0,1]$

GUM: forced constant and T indicators, $T = 100$, $M = 1000$

			D1			
1% nominal size	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
Gauge %	1.5	1.2	0.9	0.3	0.7	1.1
Potency %	—	4.6	25.6	*52.6*	*86.3*	99.0
DGP found %	29.0	0.0	0.0	0.0	8.1	36.8
			D2			
1% nominal size	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
Gauge %	1.5	1.0	0.4	0.3	1.0	0.8
Potency %	—	3.5	7.9	*24.2*	*67.1*	90.2
DGP found %	29.0	0.0	0.0	0.0	3.9	24.2

Table 2: IIS for breaks in *Autometrics*

Hoover–Perez experiments

$T = 139$, **3 relevant and 37 irrelevant variables**

	Hoover–Perez		step-wise		Autometrics	
	HP7	HP8	HP7	HP8	HP7	HP8
	1% nominal size					
Gauge %	3.0*	0.9*	0.9	3.1	1.6	1.6
Potency %	94.0	99.9	100.0	**53.3**	99.2	**100.0**
DGP found %	24.6	78.0	71.6	**22.0**	68.3	**68.8**

* Only counting significant terms (but tiebreaker is best-fitting model)

$T = 139$, **3 relevant and 141 irrelevant variables**

	step-wise		Autometrics	
	HP7	HP8	HP7	HP8
	0.1% nominal size			
Gauge %	0.1	0.7	0.3	0.1
Potency %	99.7	**40.3**	97.4	**100.0**
DGP found %	87.4	**9.0**	82.9	**90.2**

Large increase in probability of locating DGP relative to $\alpha = 0.01$ not monotonic in α–so should not select by 'goodness of fit'

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model evaluation**

(5) **Embedding theory models**

(6) **Excess numbers of variables** $N > T$

(7) **Non-invariance of NKPCs**

Conclusions

Non-invariance of NKPCs

'Hybrid' NKPC given by:

$$\Delta p_t = \underset{\geq 0}{\gamma_f} \mathsf{E}_t \left[\Delta p_{t+1} | \mathcal{I}_t\right] + \underset{\geq 0}{\gamma_b} \Delta p_{t-1} + \underset{\geq 0}{\pi} s_t + u_t \tag{14}$$

Δp_t, s_t are rate of inflation & firms' real marginal costs, so:

$$\Delta p_t = \gamma_f \Delta p_{t+1} + \gamma_b \Delta p_{t-1} + \pi s_t + \epsilon_t, \quad \epsilon_t \sim \mathsf{D}\left[0, \sigma_\epsilon^2\right] \tag{15}$$

where it is claimed:

$$\mathsf{E}_t \left[\Delta p_{t+1} \mid \mathcal{I}_t\right] = \Delta p_{t+1} + \nu_{t+1} \tag{16}$$

Δp_{t+1} instrumented by k variables $\mathbf{z}_t$ implicitly postulating:

$$\Delta p_t = \boldsymbol{\kappa}' \mathbf{z}_t + v_t \tag{17}$$

Assumes a constant world. Test by IIS on (17) adding indicators to (15): significance refutes invariance.
Also, insignificance of $\widetilde{\gamma}_f$ inconsistent with forward-looking formulation.

Euro-area NKPC

Euro-area hybrid NKPC with IV estimation, Δp_{t+1} and s_t endogenous, using instruments: five lags of inflation, two lags of s_t, detrended output and wage inflation; sample $T = 102$ (1972(2) to 1998(1)):

$$\begin{aligned} \widehat{\Delta p}_t &= \underset{(0.135)}{0.655}\ \widehat{\Delta p}_{t+1} + \underset{(0.117)}{0.280}\ \Delta p_{t-1} + \underset{(0.014)}{0.012}\ s_t + \underset{(0.010)}{0.009} \\ \chi^2_S(6) &= 11.88 \end{aligned} \tag{18}$$

Elasticities sum to 0.94 and $\widehat{\gamma}_f$ comparable to reported GMM estimates.

Adding indicators to forecasting equation

Forecasting equation for Δp_t uses instrument set for NKPC estimation with IIS in *Autometrics*
For $\alpha = 0.025$, finds 11 indicators.

When hybrid NKPC augmented by these, non-congruent, with $\chi^2_S(6) = 17.83^{**}$.

Some of instruments have explanatory power for Δp_t, consistent with (earlier) standard models of inflation.

Adding gap_{t-1} and the 11 indicators makes NKPC congruent: $\chi^2_S(4) = 2.42$.

No significant tests of residual mis-specification.

Euro-area NKPC with IIS

Selecting by *Autometrics* with $\alpha = 0.05$:

$$\begin{aligned}\widehat{\Delta p}_t = &\underset{(0.270)}{-0.325}\,\widehat{\Delta p}_{t+1} + \underset{(0.029)}{0.117}\,s_t + \underset{(0.129)}{0.515}\,\Delta p_{t-1} + \underset{(0.022)}{0.088} + \underset{(0.0005)}{0.0016}\,gap_{t-1} \\ &+\underset{(0.30)}{1.10}\,I_{73(1),t} + \underset{(0.39)}{1.12}\,I_{73(3),t} + \underset{(0.32)}{0.74}\,I_{73(4),t} + \underset{(0.35)}{0.87}\,I_{74(2),t} \\ &+\underset{(0.34)}{0.82}\,I_{74(3),t} + \underset{(0.38)}{1.01}\,I_{76(2),t} + \underset{(0.29)}{0.56}\,I_{76(3),t} - \underset{(0.28)}{0.67}\,I_{78(4),t} + \underset{(0.29)}{0.69}\,I_{83(1)}\end{aligned}$$

$$\chi^2_S(6) = 5.06 \quad \mathsf{F}_{\mathsf{ar}}(5, 85) = 1.55 \quad \mathsf{F}_{\mathsf{arch}}(4, 82) = 1.27$$
$$\mathsf{F}_{\mathsf{het}}(17, 72) = 0.91 \quad \chi^2_{\mathsf{nd}}(2) = 1.04$$

(coefficients of dummies multiplied by 100)

$\mathsf{F}_{\mathsf{name}}$ denotes an approximate F-test:
F_{ar} for k^{th}-order serial correlation;
$\mathsf{F}_{\mathsf{het}}$ for heteroskedasticity;
$\mathsf{F}_{\mathsf{reset}}$ for functional form;
$\mathsf{F}_{\mathsf{arch}}$ for k^{th}-order ARCH; and
$\chi^2_{\mathsf{nd}}(2)$ for normality.

Findings for Euro-area NKPC with IIS

Nine of the 11 'reduced form' indicators retained:
clear evidence for lack of invariance in feed-forward NKPC.

Coefficient of Δp_{t+1} is negative and insignificantly different from zero.

Coefficient of wage-share is sizeable,
serves as an important equilibrating mechanism.

A failure to model breaks induces **spurious significance** of feed-forward terms proxying expectations

Inflation 'persistence' seems an artifact of mis-specified NKPC models

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model evaluation**

(5) **Embedding theory models**

(6) **Excess numbers of variables** $N > T$

(7) **Non-invariance of NKPCs**

Conclusions

Conclusions

All essential steps feasible once target LDGP defined:

1. automatically create general model from investigator's $\mathbf{x}_t$: extra variables, lags, non-linearity, & impulse indicators– **ensures congruent GUM**;

2. embed theory-model as a 'forced' specification– **ensures theory insights retained**;

3. select most parsimonious encompassing model– **ensures undominated representation**;

4. compute near-unbiased parameter estimates– **ensures appropriate policy analyses**; and

5. stringently evaluate results– **ensures selected model valid**.

Generalizes to $N > T$ with expanding and contracting searches: see HP8 when $N = 145$, $T = 139$ at $\alpha = 0.001$.

Overall conclusions

Little difficulty in eliminating almost all irrelevant variables from the GUM (a small cost of search). **Avoids huge costs from under-specified models.**

When the LDGP would be retained by *Autometrics* if commenced from it, then a close approximation is generally selected when starting from a GUM which nests that LDGP.

Model selection by *Autometrics* with tight significance levels and bias correction is a successful approach which allows multiple breaks to be tackled.

Applied to NKPC shows lack of invariance, insignificance of feed-forward term

All the ingredients for empirical model discovery jointly with theory evaluation are in place.

The way ahead

Host of developments in automatic empirical model discovery already achieved

Theory of many stages still to be formalized

Now implementing *automatic*:

modelling of simultaneous systems

selecting cointegration vectors

testing expectations models for invariance

model averaging across terminals for forecasting.

The future is bright:

the future is *Autometrics*

References

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2010). Evaluating automatic model selection. *Journal of Time Series Econometrics*, forthcoming.

Castle, J. L., and Hendry, D. F. (2010a). Automatic selection of non-linear models. In Wang, L., Garnier, H., and Jackman, T. (eds.), *System Identification, Environmental Modelling and Control*, forthcoming. New York: Springer.

——(2010b). A low-dimension, portmanteau test for non-linearity. *JEcts*, **158**, 231–245.

——(2010c). Model selection in under-specified equations. Economics Department, Oxford.

Castle, J. L., and Shephard, N. (eds.)(2009). *Methodology and Practice of Econometrics*. OUP.

Doornik, J. A. (2008). Encompassing and automatic model selection. *OxBull*, **70**, 915–925.

Doornik, J. A. (2009). Autometrics. In Castle, and Shephard (2009), pp. 88–121.

Hendry, D. F. (1995). *Dynamic Econometrics*. OUP.

Hendry, D. F. (1999). An econometric analysis of US food expenditure, 1931–1989. In Magnus, J. R., and Morgan, M. S. (eds.), *Methodology and Tacit Knowledge*, pp. 341–361. Wiley.

Hendry, D. F., and Johansen, S. (2010). Model selection when forcing retention of theory variables. Economics Department, Oxford.

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

Hendry, D. F., and Mizon, G. E. (2010). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, forthcoming.

Hendry, D. F., and Santos, C. (2010). Automatic test of super exogeneity. In Watson, M. W., Bollerslev, T., and Russell, J. (eds.), *Volatility & Time Series Econometrics*. OUP.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered. *EctsJ*, **2**, 167–191.

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, and Shephard (2009), pp. 1–36.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Tobin, J. (1950). A statistical demand function for food in the U.S.A.. *Journal of the Royal Statistical Society*, *A, 113*(2), 113–141.

Retracing route

(1) Discovery in general

(2) Automatic model extension

(3) Automatic model selection

(4) Automatic model evaluation

(5) Embedding theory models

(6) Excess numbers of variables $N > T$

(7) Non-invariance of NKPCs

Conclusion