Empirical Model Discovery and Theory Evaluation

David F. Hendry

Department of Economics, Oxford University

GRAPES workshop, Örebro, August 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 variables are actually relevant,
their lagged responses (dynamic reactions),
functional forms of connections (non-linearity),
structural breaks and unit roots (non-stationarities),
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.

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.

But the impossible is not possible

‘Why, sometimes I’ve believed as many as six impossible things before breakfast.’

Quote from the White Queen in Lewis Carroll (1899), *Through the Looking-Glass*, Macmillan.

If only it were just 6!

Empirical econometrics

To establish 'truth' requires at least these 12 assumptions:

1. correct, comprehensive, & immutable economic theory;
2. correct, complete choice of all relevant variables & lags;
3. validity & relevance of all regressors & instruments;
4. precise functional forms for all variables;
5. absence of hidden dependencies;
6. all expectations formulations correct;
7. all parameters identified, constant over time, & invariant;
8. exact data measurements on every variable;
9. errors are 'independent' & homoscedastic;
10. error distributions constant over time;
11. appropriate estimator at relevant sample sizes;
12. valid and non-distortionary method of model selection.

If 'truth' is not on offer–what is?

Data matters for empirical implementation

Sample of T observations, $\{\mathbf{x}_t\} = \{\mathbf{y}_t, \mathbf{z}_t\}$:
but no theory specification of a 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.

'Truth' is not on offer, but theory-guided, congruent, parsimonious encompassing models with parameters invariant to relevant policies may be.

Route map

(1) **Discovery in general**

(2) Automatic model extension

(3) Automatic model selection

(4) Automatic model estimation

(5) Automatic model evaluation

(6) Embedding theory models

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

(8) An empirical example: food expenditure

Conclusions

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 methods can outperform

Five key steps:

(1) define the framework–the target for modelling–LDGP;
(2) embed target in a general formulation–model extension;
(3) search for simplest acceptable representation–select;
(4) estimate parameters near unbiasedly;
(5) evaluate the findings.

[2] formulation: many candidate variables, long lag lengths, non-linearities, outliers, and parameter shifts

[3] selection: handle more variables than observations, yet deliver high success rates by multi-path search

[4] estimation: near unbiased estimates despite selection

[5] evaluation: automatically conduct a range of pertinent tests of specification and mis-specification

Approach embodied in *Autometrics*: see Doornik (2009)

it only *appears* to be magic!

Automatic empirical model discovery

Need to tackle all complications 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).

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 $\mathsf{D}_{\mathbf{x}}(\cdot)$ for (1).

Now embed that target in a general model formulation.

Extensions for discovering a 'good' model

Second of five key steps: **extensions of** $\{\mathrm{x}_t\}$ **determine how well LDGP is approximated**.

Four main groups of automatic extensions:
additional candidate variables that 'might' be relevant;
lag formulation, implementing a **sequential factorization**;
functional form transformations for non-linearity;
impulse-indicator saturation (IIS) for parameter non-constancy and data contamination.

Must also handle mapping to non-integrated data, **conditional factorizations**, and simultaneity.

'Good choices' facilitate locating a congruent parsimonious-encompassing model of LDGP.
Congruence: matches the evidence on desired criteria;
parsimonious: as small a model as viable;
encompassing: explains the results of all other models.

Selecting and evaluating the model

Extensions create the general unrestricted model (GUM). The GUM should nest the LDGP, making it a special case; **reductions commence from GUM to locate a specific model**.

Selection is step (3):
search for the simplest acceptable representation.
Will address how that selection is done, and
(4) how near unbiased estimates obtained.

Finally, step (5):
evaluate the findings–and the selection process.
Includes tests of new aspects, such as
super exogeneity (essentially causality) for policy, and
parameter invariance (constancy across regimes).

Implications for empirical modelling

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 estimation

(5) Automatic model evaluation

(6) Embedding theory models

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

(8) An empirical example: food expenditure

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 $\boldsymbol{\Sigma}$:

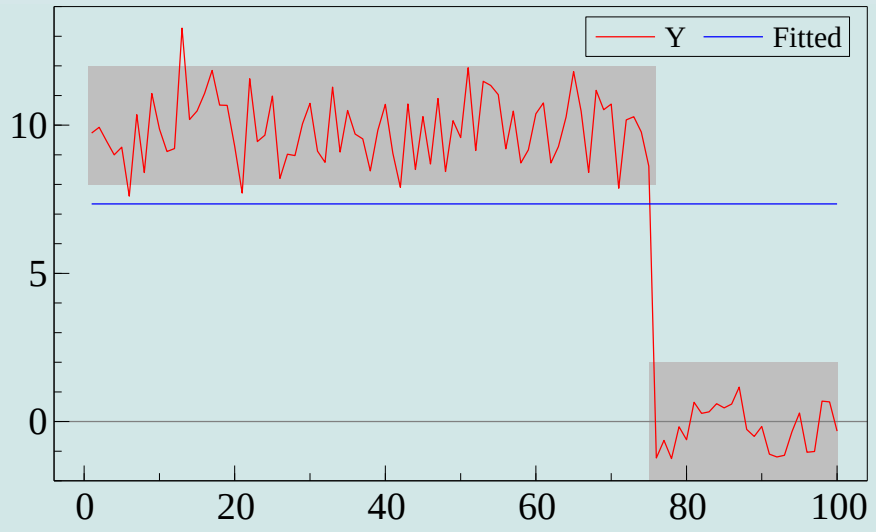
$$T^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_n\left[\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega}_\beta\right]$$

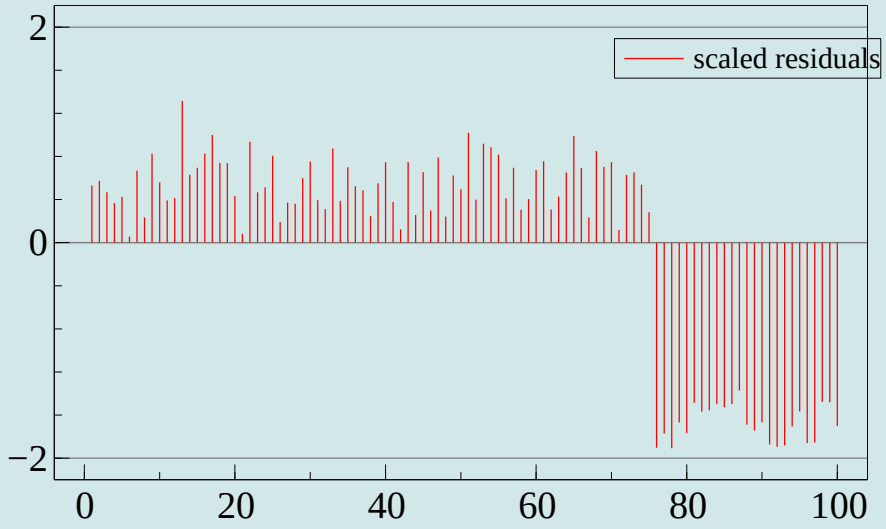
Efficiency of IIS estimator $\widetilde{\boldsymbol{\beta}}$ with respect to OLS $\widehat{\boldsymbol{\beta}}$ measured by $\boldsymbol{\Omega}_\beta$ depends on c_α and distribution

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

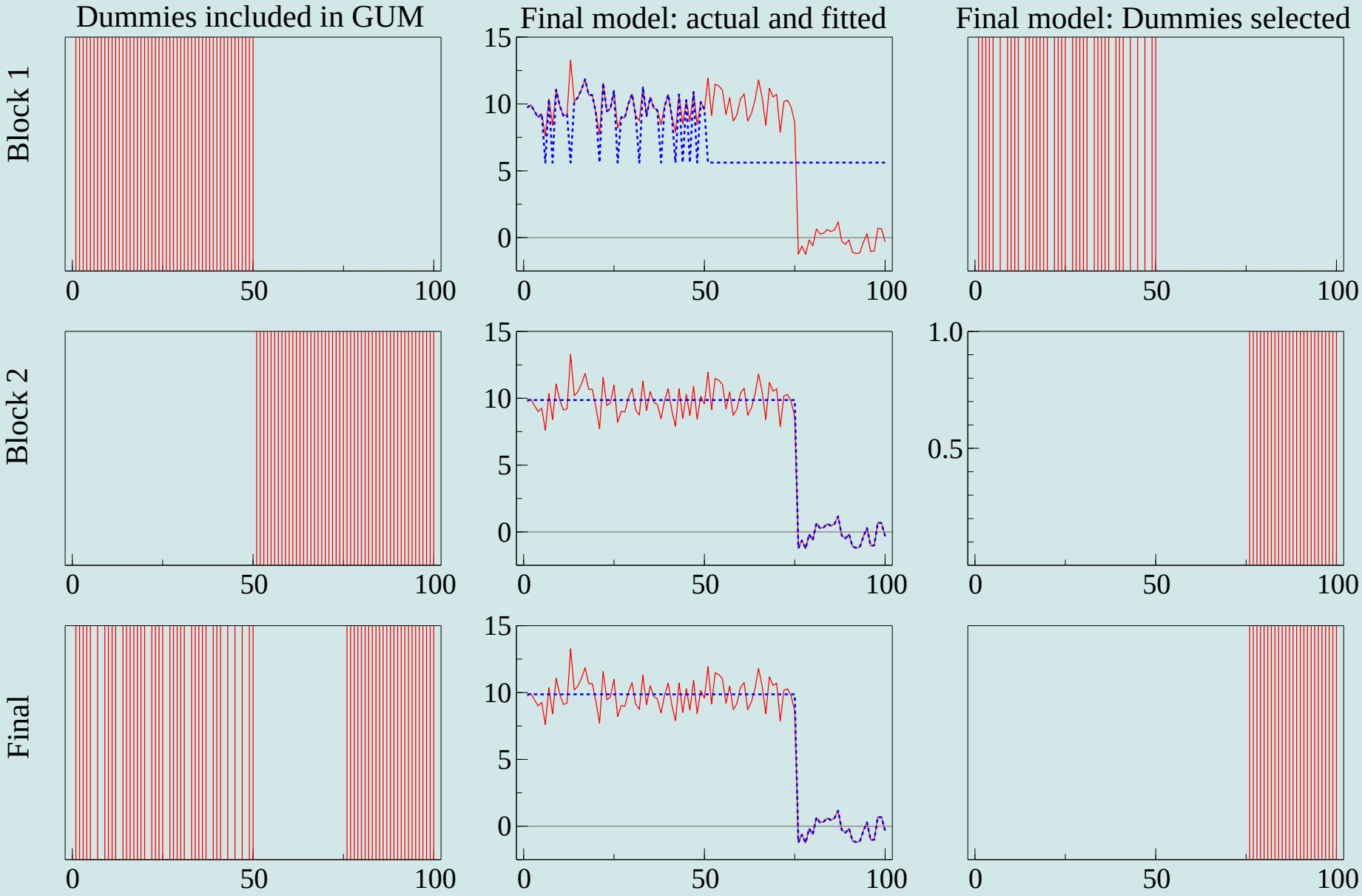




- Size of the break is **10 standard errors** at $0.75T$
- There are **no outliers** in this mis-specified model as all residuals $\in [-2, 2]$ SDs:
  outliers $\neq$ structural breaks
- step-wise regression has **zero power**

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):

$$y_t = \sum_{i=1}^{r}\sum_{j=0}^{s}\beta_{i,j}z_{i,t-j} + \sum_{i=1}^{r}\sum_{j=0}^{s}\kappa_{i,j}w_{i,t-j} + \sum_{i=1}^{r}\sum_{j=0}^{s}\theta_{i,j}w_{i,t-j}^2$$

$$+ \sum_{i=1}^{r}\sum_{j=0}^{s}\gamma_{i,j}w_{i,t-j}^3 + \sum_{j=1}^{s}\lambda_j y_{t-j} + \sum_{i=1}^{T}\delta_i 1_{\{i=t\}} + \epsilon_t$$

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

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) Automatic model estimation

(5) Automatic model evaluation

(6) Embedding theory models

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

(8) An empirical example: food expenditure

Conclusions

How to judge performance?

Many ways to judge success of selection algorithms

(A) Maximizing the goodness of fit
Traditional criterion for fitting a given model, but does not lead to useful selections

(B) Matching a theory-derived specification
Widely used, and must work well if LDGP $\simeq$ theory, but otherwise need not

(C) Frequency of discovery of the LDGP. Overly demanding–may be nearly impossible even if commenced from LDGP (eg $|\mathrm{t}| < 0.1$)

(D) Improves inference about parameters
Seek small, accurate, uncertainty regions around parameters of interest–but 'oracle principle' invalid

Operational criteria

(E) Improved forecasting over other methods
Many contenders: other selections, factors, model averages, robust devices...but forecasting is different

(F) Works for 'realistic' LDGPs
Unclear what those are–but many claimed contenders.

(G) Relative frequency of recovering LDGP starting from GUM as against starting from LDGP
Costs of search additional to commencing from LDGP

(H) Operating characteristics match theory
Nominal null rejection frequency matches actual; retained parameters of interest unbiasedly estimated

(I) Find well-specified undominated model of LDGP
'Internal criterion'–algorithm could not do better

Which criteria?

(G), (H) and (I) are main basis: aim to satisfy all three

Two costs of selection: costs of **inference** and **search**
First inevitable if tests have non-zero null and non-unit rejection frequencies under alternative
Applies even if commence from LDGP.
Measure costs of inference by RMSE of selecting or conducting inference on LDGP

When a GUM nests the LDGP, additional costs of search: calculate by increase in RMSEs for relevant variables when starting from the GUM as against the LDGP, plus those for retained irrelevant variables

Also see if *Autometrics* 'outperforms' other automatic methods:
Information Criteria, Step-wise, Lasso ....

Model selection

To successfully determine what matters and how it enters, **all main determinants must be included**:
omitting key variables adversely affects selected models.

Especially forceful issue when data are 'wide sense non-stationary'–both integrated and not time invariant

'Catch 22' – have more variables N than observations T:
so all cannot be entered from the outset.
Requires expanding as well as contracting searches
Have to select: not an option just to estimate.

To resolve conundrum, analysis proceeds in 5 steps.

Five main steps, then evaluation

1] **'1-cut' selection** for **orthogonal** designs with $N << T$.

2] Selection matters, so consider effects of **bias correction** on distributions of estimates

3] Compare '1-cut' with ***Autometrics***, which works in non-orthogonal models, still with $N << T$.

4] **More variables N than observations T follows as IIS**.

5] **Multiple breaks**, using IIS

Having resolved selection, next consider evaluation:

6] Impact of **diagnostic testing**

7] Role of **encompassing** in automatic selection

8] Testing **exogeneity** and **invariance**

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 $T \to \infty$

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.

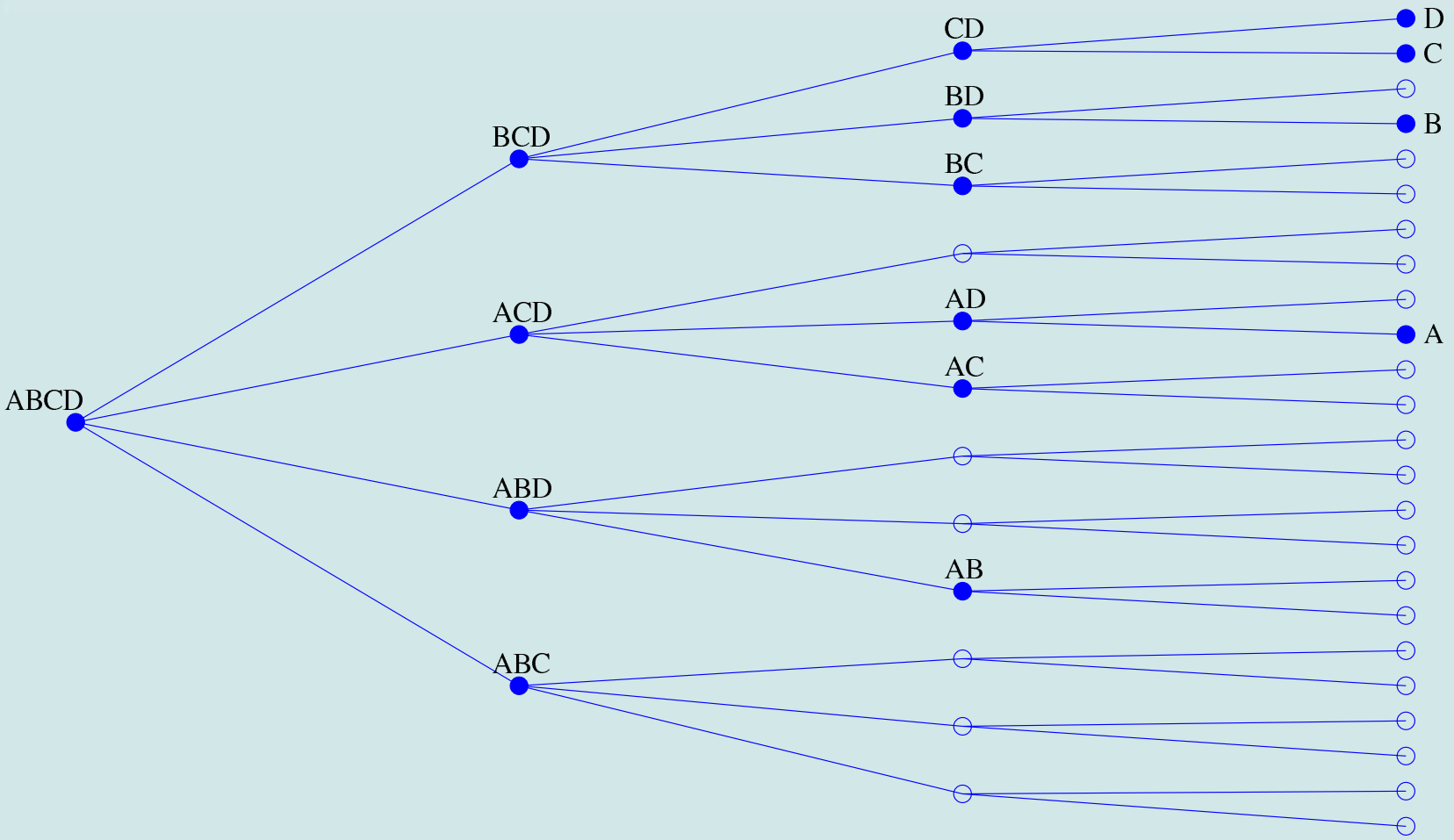
Autometrics improves on previous algorithms

- **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.

Moving from 1-cut to Autometrics

For more detail about selection outcomes, we consider 10 experiments with $N = 10$ candidate regressors and $T = 75$ based on the design in Castle, Qin and Reed (2009):

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + \epsilon_t, \tag{6}$$

$$\mathbf{x}_t \sim \mathsf{IN}_{10}\left[\mathbf{0}, \mathbf{I}_{10}\right], \tag{7}$$

$$\epsilon_t \sim \mathsf{IN}\left[0, \left(\lambda \times \sqrt{n}\right)^2\right], \quad n = 1, \ldots, 10, \; t = 1, \ldots, T \tag{8}$$

where $\mathbf{x}_t' = (x_{1,t}, \cdots, x_{10,t})$, fixed across replications.
Equations (6)–(8) specify 10 different DGPs, indexed by n, each having n relevant variables with $\beta_1 = \cdots = \beta_n = 1$ and $10 - n$ irrelevant variables ($\beta_{n+1} = \cdots = \beta_{10} = 0$).
Throughout, they set $\beta_0 = 5$.
$\lambda = 0.4$ is their $\mathsf{R}^2 = 0.9$ experiment.
Table 1 reports the non-centralities, ψ.

Experimental design

n	1	2	3	4	5	6	7	8	9	10
$\psi_1 \dots \psi_n$	21.6	15.3	12.5	10.8	9.7	8.8	8.2	7.7	7.2	6.9

Table 1: Non-centralities for simulation experiments (6)–(8).

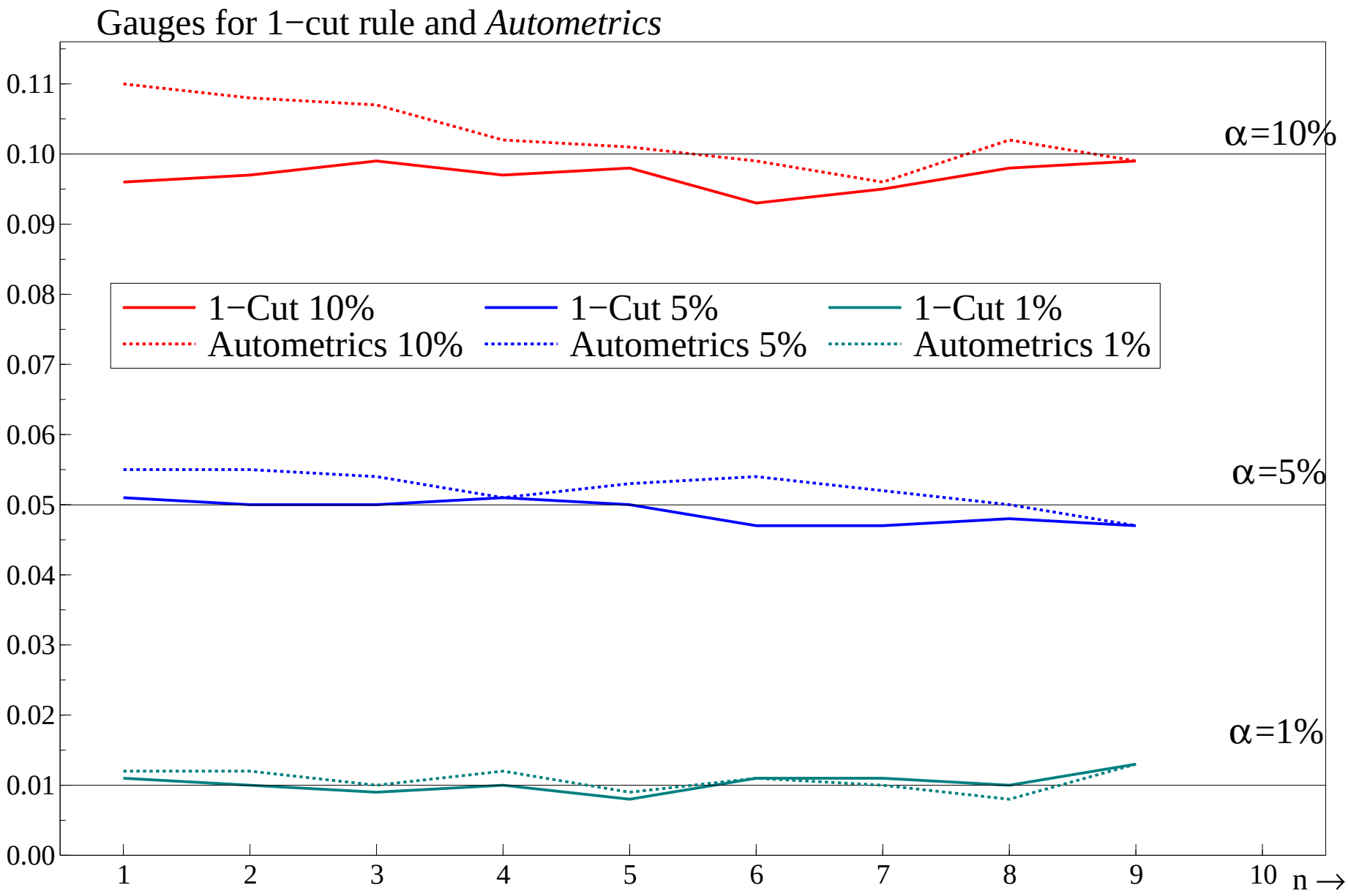
The GUM is the same for all 10 DGPs:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + u_t.$$

We first investigate how the general search algorithm *Autometrics*, without diagnostic checking, performs relative to 1-cut.

Comparative gauges are recorded in figure 43.
1-cut gauge is very accurate, and
Autometrics is quite close, especially for $\alpha = 0.01$.
Potency ratios are unity for all significance levels.

Gauges for 1-cut & Autometrics



Calculating MSEs

Also report MSEs after model selection.

$\widehat{\beta}_{k,i}$ is OLS estimate of coefficient on $x_{k,t}$ in GUM for replication i. $\widetilde{\beta}_{k,i}$ is OLS estimate after model selection

$\widetilde{\beta}_{k,i} = 0$ when $x_{k,t}$ not selected in final model.

Calculate following MSEs:

$$\begin{aligned}
\mathsf{MSE}_k &= \frac{1}{M}\sum_{i=1}^{M}\left(\widehat{\beta}_{k,i}-\beta_k\right)^2, \\
\mathsf{UMSE}_k &= \frac{1}{M}\sum_{i=1}^{M}\left(\widetilde{\beta}_{k,i}-\beta_k\right)^2, \\
\mathsf{CMSE}_k &= \frac{\sum_{i=1}^{M}\left[\left(\widetilde{\beta}_{k,i}-\beta_k\right)^2\cdot 1_{(\widetilde{\beta}_{k,i}\neq 0)}\right]}{\sum_{i=1}^{M} 1_{(\widetilde{\beta}_{k,i}\neq 0)}}, \quad \left(\beta_k^2 \text{ if } \sum_{i=1}^{M} 1_{(\widetilde{\beta}_{k,i}\neq 0)} = 0\right)
\end{aligned}$$

Unconditional MSE (UMSE):

sets $\widetilde{\beta}_{k,i} = 0$ when a variable is not selected.

Conditional MSE (CMSE) is over retained variables only.

MSEs for 1-cut & Autometrics

Figure 46 records ratios of MSEs of *Autometrics* selection to 1-cut for both unconditional and conditional distributions, but with no diagnostic tests and no bias correction.

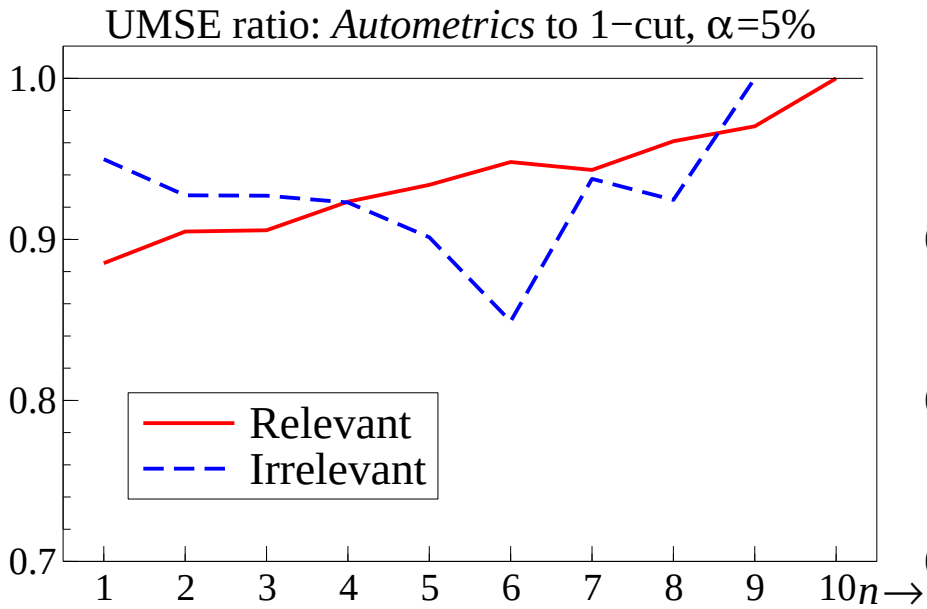
Lines labelled *Relevant* report the ratios of average MSEs over all relevant variables for a given n.
Lines labelled *Irrelevant* are based on average MSEs of irrelevant variables for each DGP (none when $n = 10$).

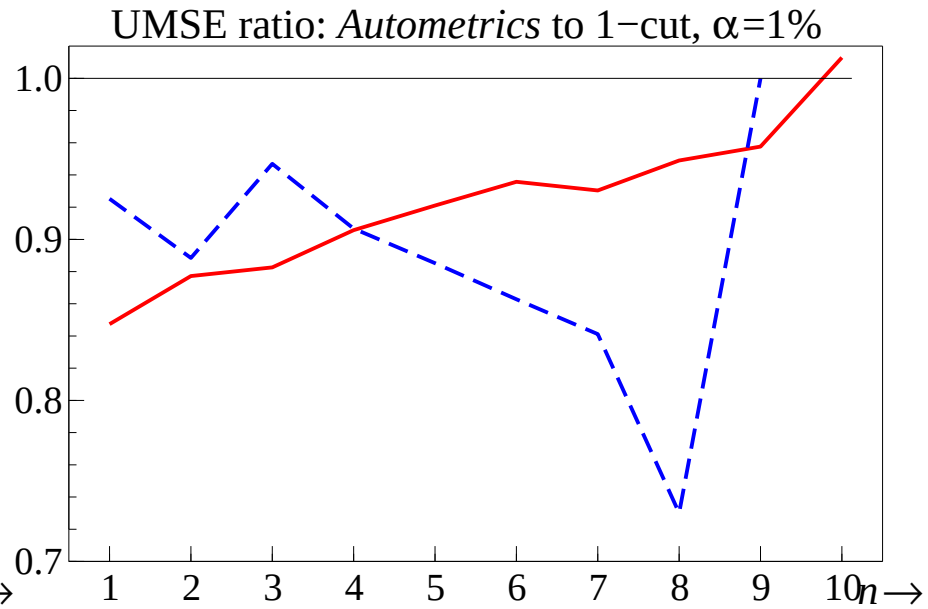
Unconditionally, ratios are close to 1 for irrelevant variables; but there is some advantage to using *Autometrics* for relevant variables, as ratios are uniformly less than 1.

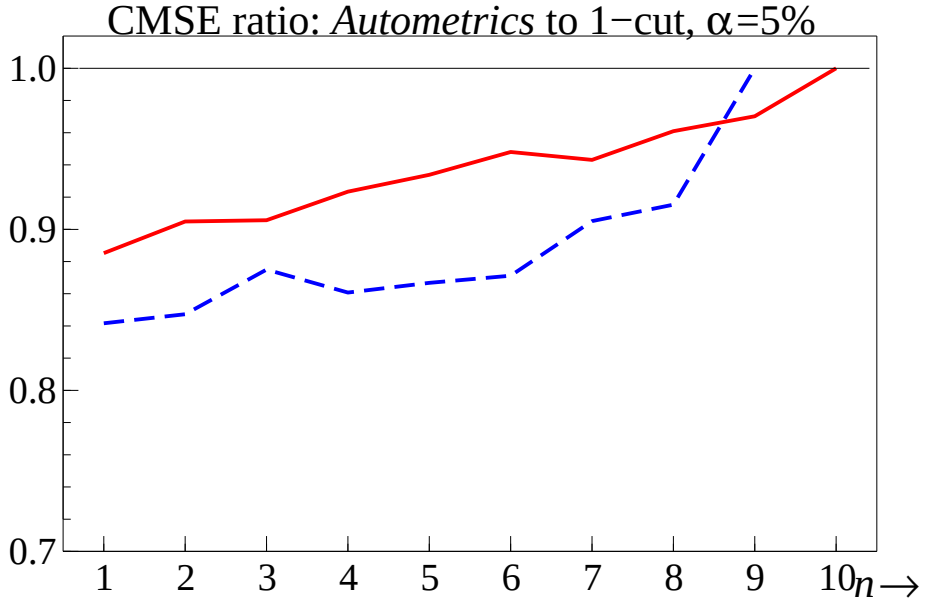
Benefits to selection are largest when there are few relevant variables that are highly significant.

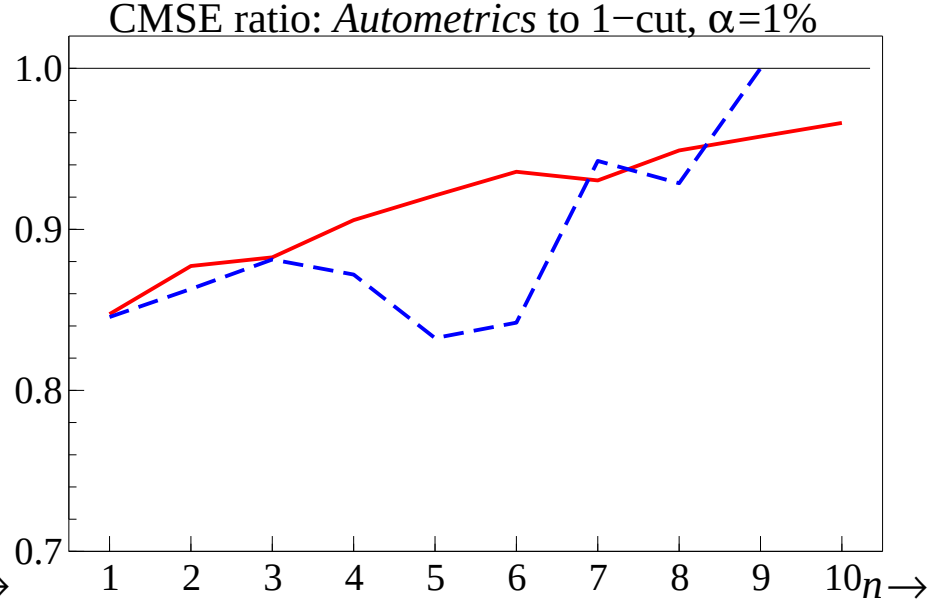
Conditionally, *Autometrics* outperforms 1-cut in most cases.

Ratios of MSEs for Autometrics to 1-cut









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).

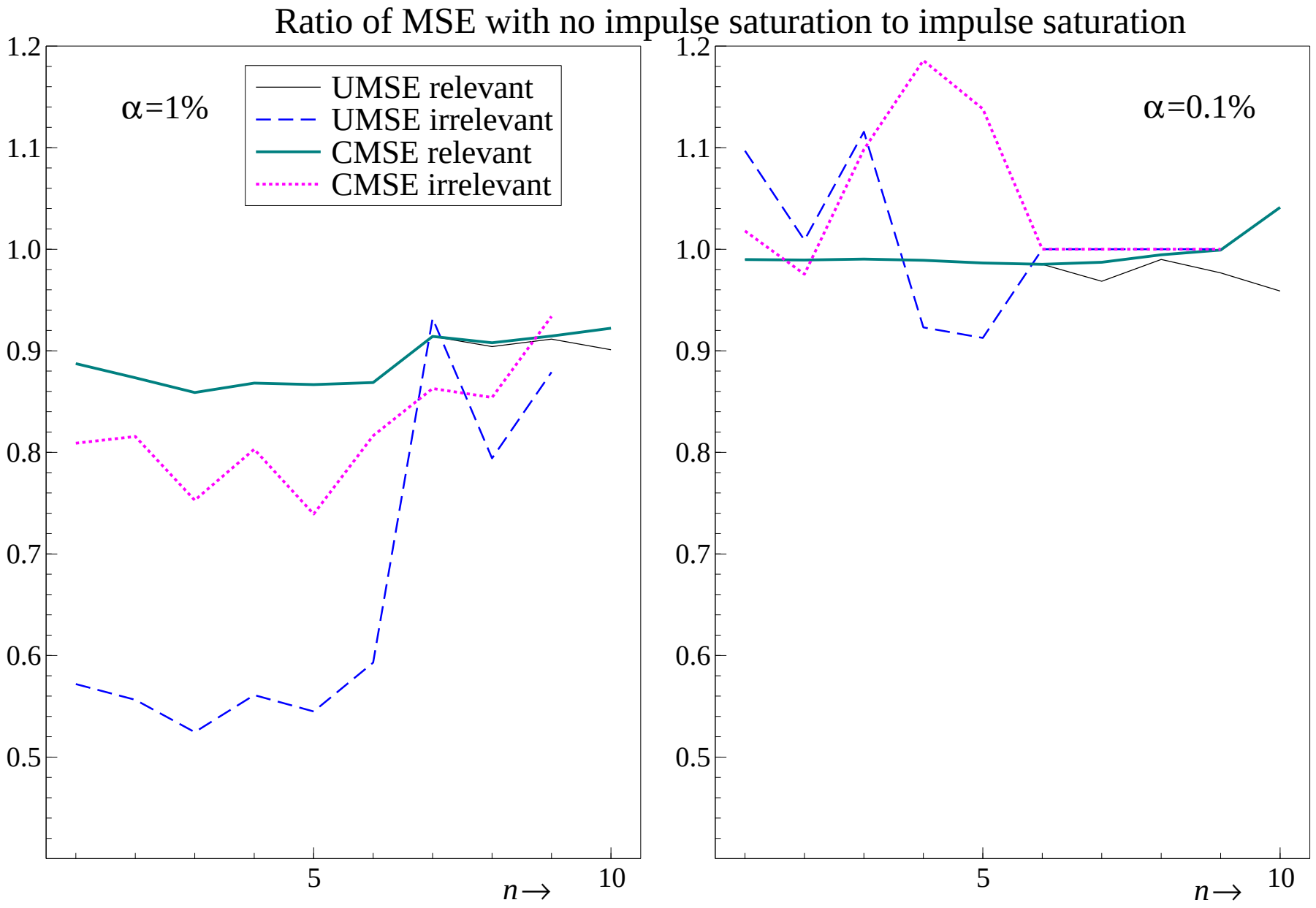
Autometrics' *IIS outcomes*

α	1%		0.1%	
	no IIS	IIS	no IIS	IIS
ave. gauge	1.22%	1.43%	0.12%	0.08%
ave. potency	99.98%	99.97%	99.87%	99.82%

Table 2: Gauge and potency averages over $n = 1, \ldots, 10$, with and without IIS.

Figure 49 records ratio of MSEs without saturation to with.
Under null, indicator saturation at tight α has small costs, but MSEs of irrelevant variables are larger than without IIS at 1%.
Correlations between dummies and retained irrelevant variables might increase MSE.
At $\alpha = 0.1\%$, little impact on MSE as so few dummies retained: impulse-indicator saturation can even improve MSE under null.

Ratios of MSEs without/with IIS



Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model estimation**

(5) Automatic model evaluation

(6) Embedding theory models

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

(8) An empirical example: food expenditure

Conclusions

Automatic bias corrections

Selection matters as only retain 'significant' variables: so correct final estimates for selection

Convenient approximation that:

$$\mathsf{t}_{\widehat{\beta}} = \frac{\widehat{\beta}}{\widehat{\sigma}_{\widehat{\beta}}} \simeq \frac{\widehat{\beta}}{\sigma_{\widehat{\beta}}} \sim \mathsf{N}\left[\frac{\beta}{\sigma_{\widehat{\beta}}}, 1\right] = \mathsf{N}\left[\psi, 1\right]$$

when non-centrality of t-test is $\psi = \beta/\sigma_{\widehat{\beta}}$

Use Gaussian approximation (IIS helps ensure):

$$\phi(w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) \quad \text{where} \quad \Phi(w) = \int_{-\infty}^{w} \phi(x)\,\mathsf{d}x$$

Doubly-truncated distribution–expected t-value is:

$$\mathsf{E}\left[|\mathsf{t}_{\widehat{\beta}}| \mid |\mathsf{t}_{\widehat{\beta}}| > c_\alpha; \psi\right] = \psi^* \tag{9}$$

so observed |t|-value is unbiased estimator for ψ^*

Truncation correction

Sample selection induces:

$$\psi^* = \psi + \frac{\phi(c_\alpha - \psi) - \phi(-c_\alpha - \psi)}{1 - \Phi(c_\alpha - \psi) + \Phi(-c_\alpha - \psi)} = \psi + r\left(\psi, c_\alpha\right) \tag{10}$$

Correct $\widetilde{\beta}$ once ψ is known: for $\beta > 0$, say:

$$\mathsf{E}\left[\widetilde{\beta} \mid \widetilde{\beta} \geq \sigma_{\widetilde{\beta}} c_\alpha\right] = \beta\left(1 + \frac{r\left(\psi, c_\alpha\right)}{\psi}\right) = \beta\left(\frac{\psi^*}{\psi}\right) \tag{11}$$

Let:

$$\widetilde{\psi} = \mathsf{t}_{\widetilde{\beta}} - r\left(\mathsf{t}_{\widetilde{\beta}}, c_\alpha\right), \quad \text{then} \quad \widetilde{\widetilde{\psi}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\widetilde{\psi}, c_\alpha\right) \tag{12}$$

leading to the bias-corrected parameter estimate:

$$\widetilde{\widetilde{\beta}} = \widetilde{\beta}\left(\widetilde{\widetilde{\psi}} / \mathsf{t}_{\widetilde{\beta}}\right) \tag{13}$$

from inverting (11).

Implementing bias correction

Bias corrects closely, not exactly, for relevant: over-corrects for some t-values.

Some increase in MSEs of relevant variables.
Correction exacerbates downward bias in unconditional estimates of relevant coefficients & increases MSEs slightly.

No impact on 'bias' of estimated parameters of irrelevant variables as their $\beta_i = 0$, so unbiased with or without selection

But **remarkable decrease in MSEs of irrelevant variables**

First 'free lunch' of new approach.
Obvious why in retrospect–most correction for $|t|$ near c_α, which occurs for retained irrelevant variables.

Simulation MSEs

Impact of bias corrections on retained irrelevant and relevant variables, for $N = 1000$ and $n = 10$ in (5).

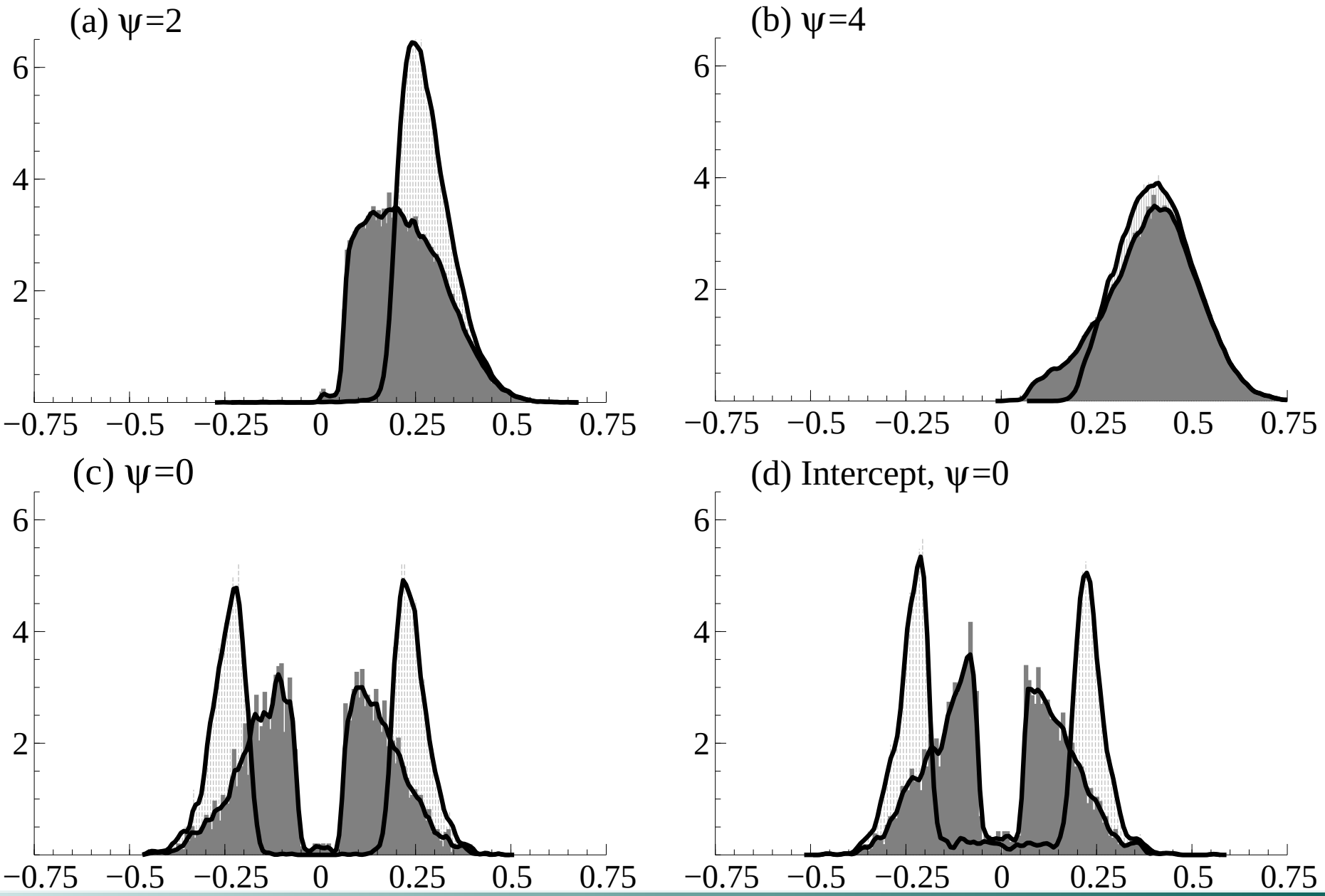
α	1%	0.1%	1%	0.1%
	average CMSE over 990 irrelevant variables		average CMSE over 10 relevant variables	
uncorrected $\widetilde{\beta}$	0.84	1.23	1.0	1.4
$\overline{\beta}$ after correction	0.38	0.60	1.2	1.3

Table 3: Average CMSEs, times 100, for retained relevant and irrelevant variables (excluding β_0), with and without bias correction.

Greatly reduces MSEs of irrelevant variables in both unconditional and conditional distributions.

Coefficients of retained variables with $|\mathrm{t}| \leq c_\alpha$ are not bias corrected–insignificant estimates set to zero.

Bias correcting conditional distributions at 5%



Implications

If regress y_t on exogenous $\{z_{j,t}\}$ $(j = 1, \ldots, N)$ over $t = 1, \ldots, T$ & select $\{z_{j,t}\}$ $(j = 1, \ldots, n)$ by *Gets*, then in:

$$\underset{(\mathsf{SE})}{\widehat{y}_t = \sum_{j=1}^{n} \widehat{\beta}_j \; z_{j,t} + \underset{\widehat{\sigma}_\epsilon}{\widehat{\epsilon}_t}} \tag{14}$$

a. **Estimates near unbiased**, $\mathsf{E}[\widehat{\beta}_j] \simeq \beta_j$, for constant parameter β_j in LDGP when bias-corrected;

b. **SEs accurate for SDs** of estimated LDGP equation: $\mathsf{V}[\widehat{\beta}_j] \simeq \mathsf{V}[\widetilde{\beta}_j]$ for that $\widetilde{\beta}_j$ estimated in LDGP equation;

c. **Estimated equation standard error nearly unbiased** ($\mathsf{E}[\widehat{\sigma}_\epsilon] \simeq \sigma_\epsilon$);

d. **Relevant variables retained with almost same probabilities as commencing from LDGP**; and

e. **Irrelevant variables eliminated at rate** $(1 - \alpha)(N - n)$

Resurrected conventional econometrics

Already amazing advances–but:

1] **Assumed all $\beta_1 \dots \beta_n, \sigma_\epsilon$ were constant in LDGP;**

2] **Assumed $\{\epsilon_t\}$ was approximately normal;**

3] **Assumed $N << T$;**

4] **Assumed $\{z_{j,t}\}$ were (weakly) exogenous;**

5] **Assumed (14) was linear in $\{z_{j,t}\}$;**

6] **Assumed $z_{j,t}$ rather than $\mathsf{E}[z_{j,t}]$ mattered;**

7] **Assumed (14) was identified within LDGP system.**

All need to be tested–and all have been solved:
first examine impact of automatic evaluation

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model estimation**

(5) **Automatic model evaluation**

(6) Embedding theory models

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

(8) An empirical example: food expenditure

Conclusions

Model evaluation criteria

Given transformed system version of (3):

$$\mathbf{\Lambda}(L)\,\mathbf{h}(\mathbf{y})_t = \mathbf{B}(L)\,\mathbf{g}(\mathbf{z})_t + \mathbf{\Gamma}\mathbf{d}_t + \mathbf{v}_t \tag{15}$$

where $\mathbf{v}_t \sim \mathsf{D}_{n_1}[\mathbf{0}, \mathbf{\Sigma}_v]$ when $\mathbf{x}_t' = (\mathbf{y}_t' : \mathbf{z}_t'),\ (n_1 : n_2)$

[1] homoscedastic innovation $\mathbf{v}_t$

[2] weak exogeneity of $\mathbf{z}_t$ for parameters of interest $\boldsymbol{\mu}$

[3] constant, invariant parameter $\boldsymbol{\mu}$

[4] data-admissible formulations on accurate observations

[5] theory consistent, identifiable structures

[6] encompass rival models

Exhaustive nulls to test–but many alternatives

Models which satisfy first four are congruent

Encompassing, congruent, theory-consistent model satisfies all six criteria

Integrated data

Autometrics conducts inferences for I(0)
Most selection tests remain valid:
see Sims, Stock and Watson (1990)
Only tests for a unit root need non-standard critical values

Implemented PcGive cointegration test in *PcGets* 'Quick Modeler'

Most diagnostic tests also valid for integrated series:
see Wooldridge (1999)

Heteroscedasticity tests an exception:
powers of variables then behave oddly
see Caceres (2007)

Role of mis-specification testing

Under null of congruent GUM, Figure 62 compares gauges for *Autometrics* with diagnostic checking **on** vs. **off**:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \quad \text{for} \quad \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2] \tag{16}$$

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

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

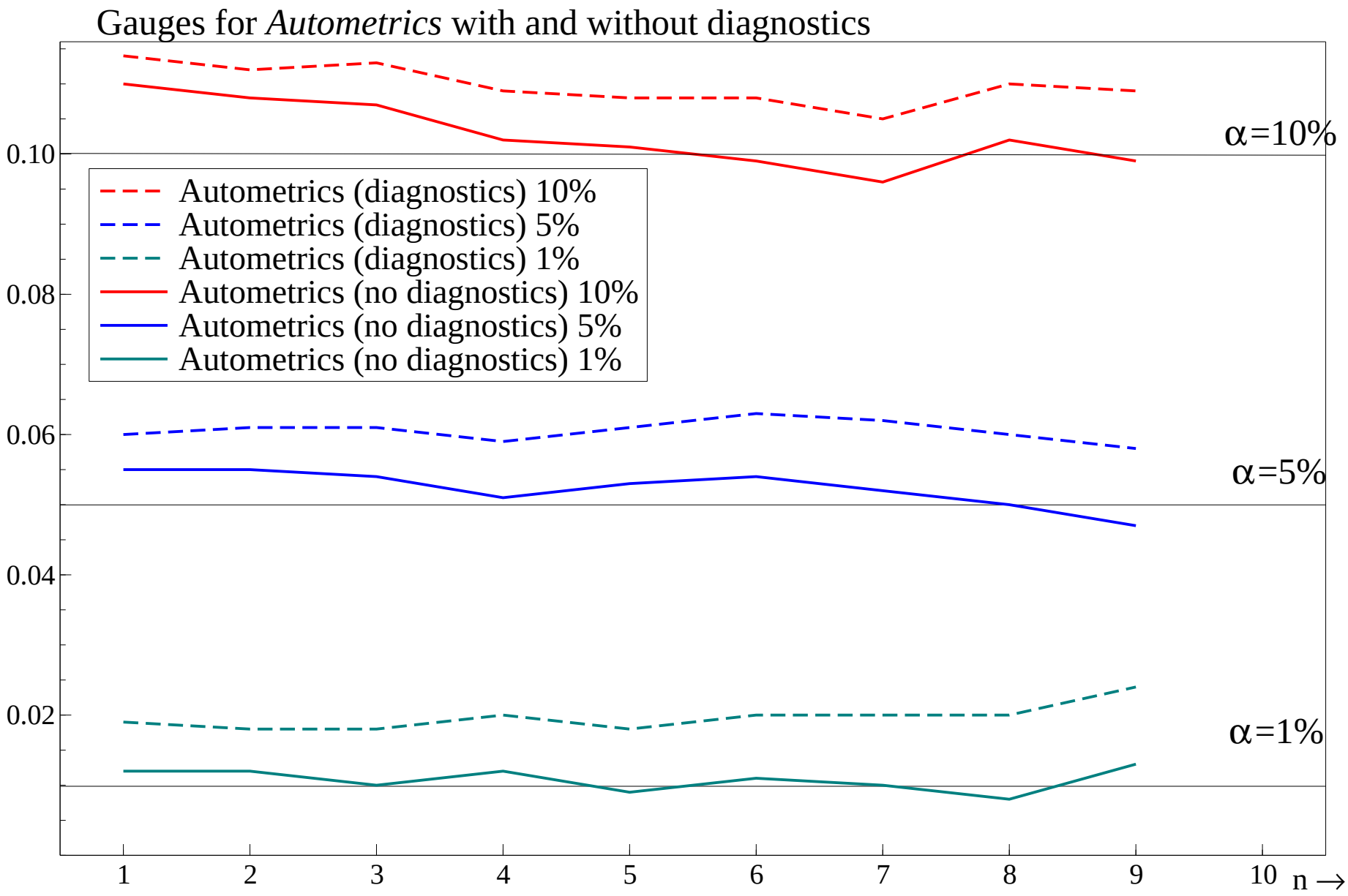
'*Potency*' is average retention rate of relevant variables (should be near theory power for a 1-off test).

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



Impact of mis-specification testing on MSEs

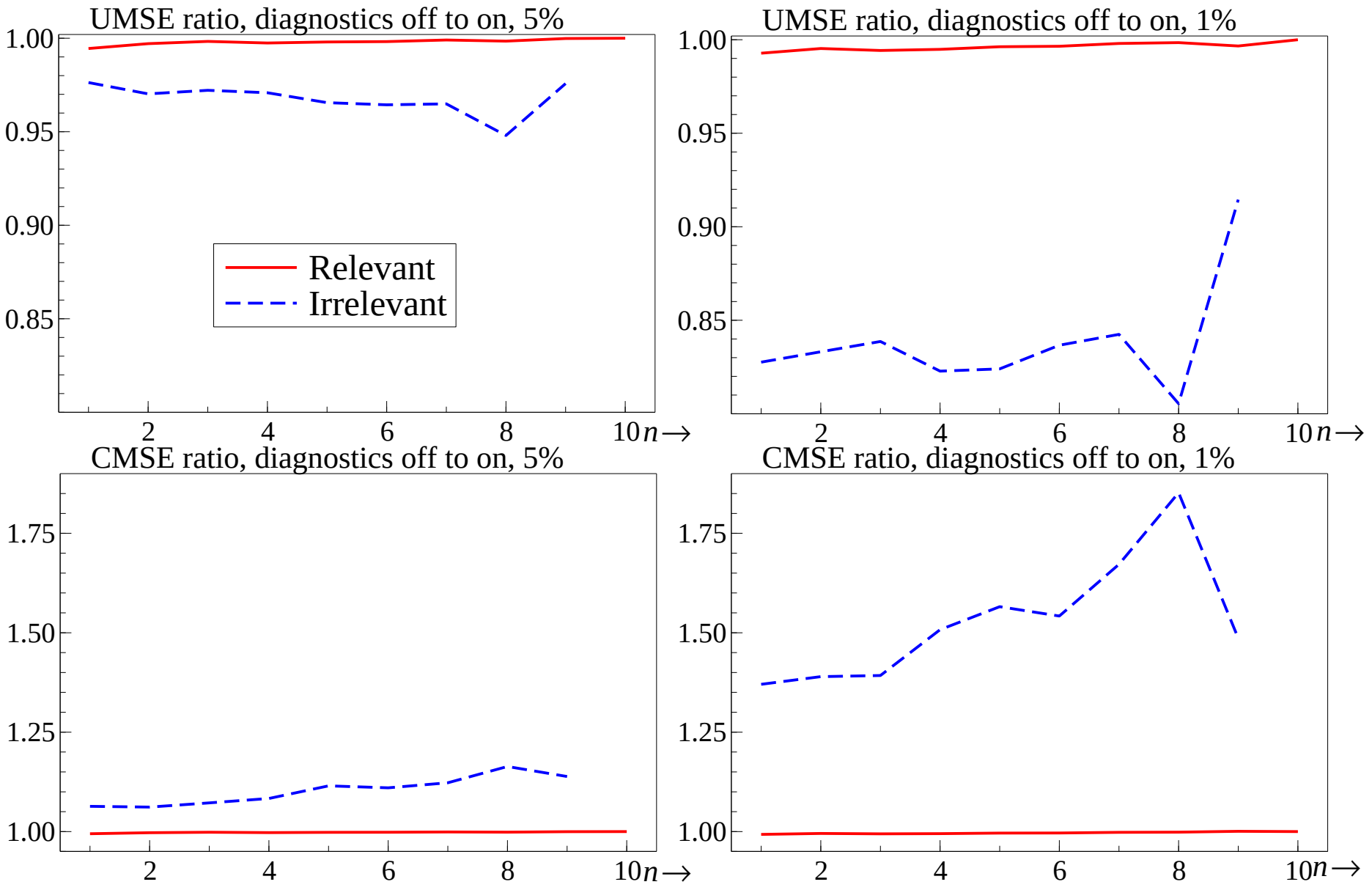
Figure 64 records ratios of MSEs in the unconditional distribution (UMSEs) and conditional (CMSEs) when diagnostic tests switched off to on, averaging within relevant and irrelevant variables.

Switching diagnostics off generally improves UMSEs, but worsens results conditionally, with the impact coming through the irrelevant variables.

Switching diagnostics off leads to fewer irrelevant regressors being retained overall, improving UMSEs, but retained irrelevant variables are now more significant than with diagnostics on.

Impact is largest at tight significance levels: at 10%, MSE ratios so close to unity they are not plotted.

Ratios of MSE*s for 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 4 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 4: 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

Implications of selection

Despite selecting from $N = 1000$ potential variables when only $n = 10$ are relevant:

(1) nearly unbiased estimates of coefficients and equation standard errors can be obtained;

(2) little loss of efficiency from checking many irrelevant variables;

(3) some loss from not retaining relevant variables at large c_α;

(4) huge gain by not commencing from an under-specified model;

(5) even works well for 'fat-tailed' errors at tight α when IIS used–see below.

Now embed theory models for non-orthogonal data sets with $N > T$.

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model estimation**

(5) **Automatic model evaluation**

(6) **Embedding theory models**

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

(8) An empirical example: food expenditure

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.

Theory exactly correct

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

$$y_t = \boldsymbol{\beta}' \mathbf{z}_t + \epsilon_t \tag{17}$$

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

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \sum_{t=1}^{T} \mathbf{z}_t \epsilon_t \sim \mathsf{N}_n \left[\mathbf{0}, \sigma_\epsilon^2 \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \right] \tag{18}$$

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:

$$y_t = \boldsymbol{\beta}' \mathbf{z}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \nu_t \tag{19}$$

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

$$\mathbf{w}_t = \widehat{\boldsymbol{\Gamma}} \mathbf{z}_t + \mathbf{u}_t \tag{20}$$

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

$$y_t = \boldsymbol{\beta}' \mathbf{z}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \nu_t = \boldsymbol{\beta}' \mathbf{z}_t + \boldsymbol{\gamma}' \mathbf{u}_t + \nu_t \tag{21}$$

Distributions of forced estimates

Consequently:

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

$$\sim \mathsf{N}_{n+k}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \right)^{-1} \end{pmatrix} \right] \quad (22)$$

as $\sum_{t=1}^{T} \mathbf{z}_t \mathbf{u}_t' \simeq 0$, so distribution of $\widetilde{\boldsymbol{\beta}}$ in (22) **identical** to that of $\widehat{\boldsymbol{\beta}}$ in (18): **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, $\mathbf{x}_t$, from the DGP; induces less useful LDGP–hard to avoid if $\mathbf{x}_t$ unknown.
In a], $\boldsymbol{\gamma} \neq \mathbf{0}$, as $\mathbf{z}_t$ and $\mathbf{u}_t$ orthogonal in (23), coefficient of former is $\boldsymbol{\beta} + \boldsymbol{\gamma}' \widehat{\boldsymbol{\Gamma}}$, which is estimated if (17) is simply fitted to the data: but may be significant with anticipated signs.

In b], when (19) nests LDGP, but $\mathbf{x}_t$ 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 (17) will result in biased, inefficient, possibly non-constant, estimates as:

$$y_t = \boldsymbol{\beta}'\mathbf{z}_t + \boldsymbol{\gamma}'\left(\widehat{\boldsymbol{\Gamma}}\mathbf{z}_t + \mathbf{u}_t\right) + \nu_t = \left(\boldsymbol{\beta} + \boldsymbol{\gamma}'\widehat{\boldsymbol{\Gamma}}\right)'\mathbf{z}_t + \boldsymbol{\gamma}'\mathbf{u}_t + \nu_t \quad (23)$$

Now forcing $\mathbf{z}_t$ when selecting from (23) 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, $\mathbf{x}_t$, omitted from DGP and breaks occur.

Breaks in included and excluded variables

DGP:

$$y_t = \boldsymbol{\beta}_1' \mathbf{z}_t + \boldsymbol{\beta}_2' \mathbf{x}_t + \epsilon_t, \qquad \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right] \tag{24}$$

$$\begin{pmatrix} \mathbf{z}_t \\ \mathbf{x}_t \end{pmatrix} \sim \mathsf{IN}_{k_1+k_2}\left[\begin{pmatrix} \boldsymbol{\mu}_t \\ \boldsymbol{\delta}_t \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right] \tag{25}$$

Both sets of variables have one-off location shifts:

$$\boldsymbol{\mu}_t = \begin{cases} \boldsymbol{\mu}_1 & t < T^0 \\ \boldsymbol{\mu}_2 & t \geq T^0 \end{cases} \quad \text{and} \quad \boldsymbol{\delta}_t = \begin{cases} \boldsymbol{\delta}_1 & t < T^* \\ \boldsymbol{\delta}_2 & t \geq T^* \end{cases} \tag{26}$$

The LDGP is mis-specified as:

$$y_t = \gamma_0 + \boldsymbol{\gamma}_1' \mathbf{z}_t + e_t \tag{27}$$

so $\mathbf{x}_t$ is unknowingly omitted, and (27) is also the model.

Breaks in included and excluded variables

Relationship between $\mathbf{z}_t$ and $\mathbf{x}_t$ is:

$$\mathbf{x}_t = (\boldsymbol{\delta}_t - \boldsymbol{\Psi}\boldsymbol{\mu}_t) + \boldsymbol{\Psi}\mathbf{z}_t + \mathbf{u}_t \tag{28}$$

where $\mathsf{E}[\mathbf{z}_t\mathbf{u}_t'] = \mathbf{0}$ and $\boldsymbol{\Psi} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$, giving a reduced LDGP:

$$y_t = \boldsymbol{\beta}_2' (\boldsymbol{\delta}_t - \boldsymbol{\Psi}\boldsymbol{\mu}_t) + \left(\boldsymbol{\beta}_1' + \boldsymbol{\beta}_2'\boldsymbol{\Psi}\right)\mathbf{z}_t + \boldsymbol{\beta}_2'\mathbf{u}_t + \epsilon_t. \tag{29}$$

Full sample estimation of (27) yields:

$$\begin{pmatrix} \widetilde{\gamma}_0 \\ \widetilde{\boldsymbol{\gamma}}_1 \end{pmatrix} = \begin{pmatrix} T & \sum_{t=1}^T \mathbf{z}_t' \\ \sum_{t=1}^T \mathbf{z}_t & \sum_{t=1}^T \mathbf{z}_t\mathbf{z}_t' \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^T y_t \\ \sum_{t=1}^T \mathbf{z}_t y_t \end{pmatrix}$$

Breaks in included and excluded variables

$$\mathsf{E}\left[\begin{pmatrix}\widetilde{\gamma}_0\\ \widetilde{\gamma}_1\end{pmatrix}\right] \simeq \begin{pmatrix}\left(\mathbf{s}' - \mathbf{r}'\mathbf{H}^{-1}\left(\boldsymbol{\Sigma}_{12} + \lambda\left(1-\kappa\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\right)'\right)\right)\boldsymbol{\beta}_2\\ \boldsymbol{\beta}_1 + \mathbf{H}^{-1}\left(\boldsymbol{\Sigma}_{12} + \lambda\left(1-\kappa\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\right)'\right)\boldsymbol{\beta}_2\end{pmatrix}$$

$$= \begin{pmatrix}\gamma_{0,p}\\ \boldsymbol{\gamma}_{1,p}\end{pmatrix}$$

where

$\lambda = \left(T^0 - 1\right)/T$;

$\kappa = \left(T^* - 1\right)/T$;

$\mathbf{r} = \left(\lambda\boldsymbol{\mu}_1 + \left(1-\lambda\right)\boldsymbol{\mu}_2\right)$;

$\mathbf{s} = \left(\kappa\boldsymbol{\delta}_1 + \left(1-\kappa\right)\boldsymbol{\delta}_2\right)$;

$\mathbf{M} - \mathbf{r}\mathbf{s}' = \lambda\left(1-\kappa\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\delta}_1' - \boldsymbol{\delta}_2'\right)$;

$\mathbf{H} = \boldsymbol{\Sigma}_{11} + \lambda\left(1-\lambda\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)'$.

Implications

- Constant $\mathbf{z}_t$, break in $\mathbf{x}_t$

$$\left(\begin{array}{c} \gamma_{0,p} \\ \boldsymbol{\gamma}_{1,p} \end{array} \right)_{\boldsymbol{\mu}_1=\boldsymbol{\mu}_2} = \left(\begin{array}{c} \left(\left(\kappa \boldsymbol{\delta}_1 + (1-\kappa)\, \boldsymbol{\delta}_2 \right)' - \boldsymbol{\mu}' \boldsymbol{\Psi}' \right) \boldsymbol{\beta}_2 \\ \left(\boldsymbol{\beta}_1 + \boldsymbol{\Psi}' \boldsymbol{\beta}_2 \right) \end{array} \right) \tag{30}$$

  - Slope coefficient is constant;
  - Intercept shifts whenever omitted variables shift;
  - Residual var. inflated $\kappa\,(1-\kappa)\,\boldsymbol{\beta}_2'\,(\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1)\,(\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2)'\,\boldsymbol{\beta}_2$

- Break in $\mathbf{z}_t$, constant $\mathbf{x}_t$

$$\left(\begin{array}{c} \gamma_{0,p} \\ \boldsymbol{\gamma}_{1,p} \end{array} \right)_{\boldsymbol{\delta}_1=\boldsymbol{\delta}_2} = \left(\begin{array}{c} \left(\boldsymbol{\delta}' - \mathbf{r}' \mathbf{H}^{-1} \boldsymbol{\Sigma}_{12} \right) \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_1 + \mathbf{H}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}_2 \end{array} \right) \tag{31}$$

  - Slope and intercept shifts – biases in estimated coefficients lead to induced non-constancy

Impulse-indicator saturation

IIS removes location-shift induced non-constancies in intercepts and equation standard errors.

$$y_t = \boldsymbol{\beta}_2' (\boldsymbol{\delta}_t - \boldsymbol{\Psi}\boldsymbol{\mu}_t) + \left(\boldsymbol{\beta}_1' + \boldsymbol{\beta}_2'\boldsymbol{\Psi}\right)\mathbf{z}_t + \boldsymbol{\beta}_2'\mathbf{u}_t + \epsilon_t$$

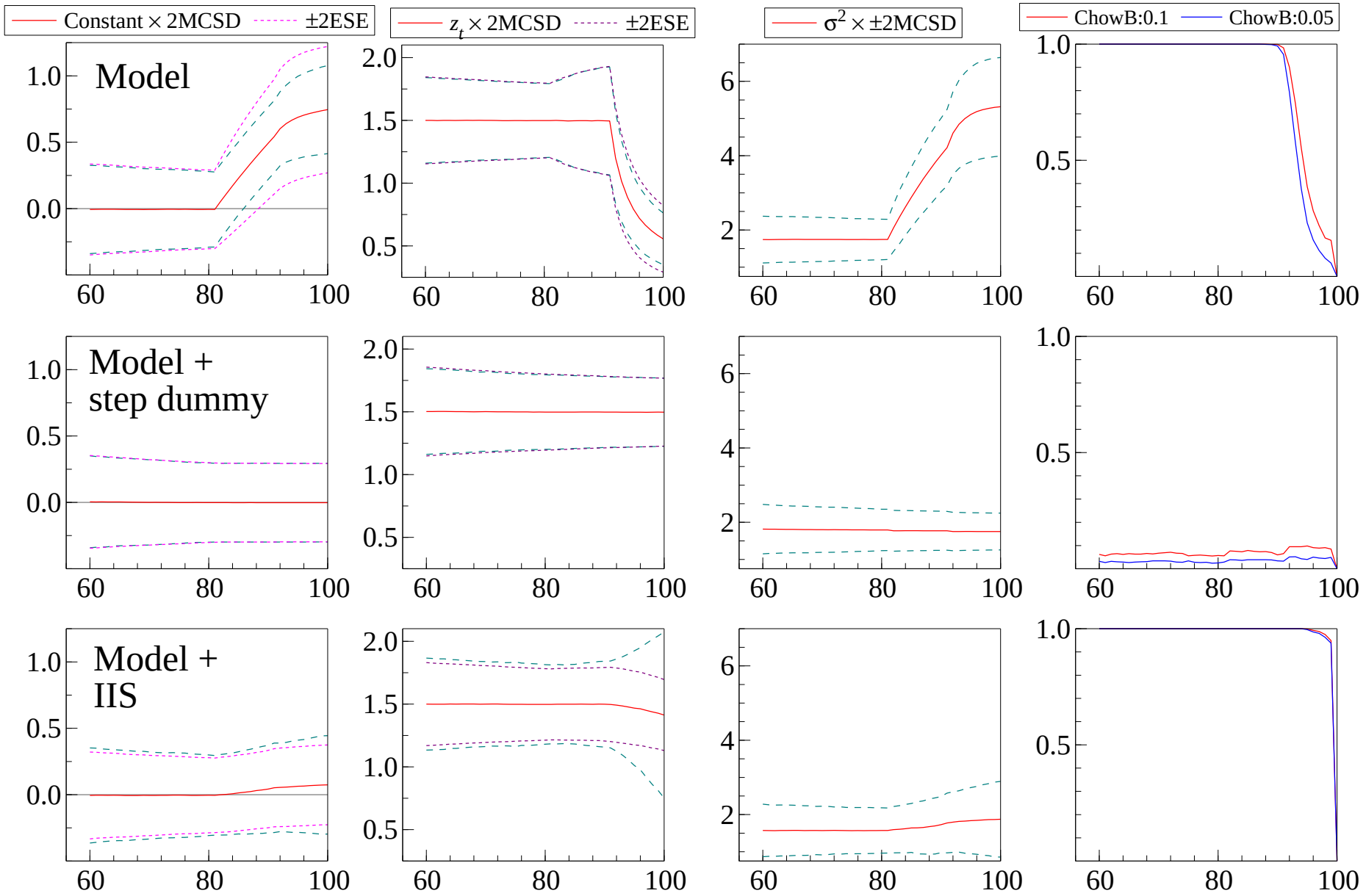
- Inconsistent coefficient of $\left(\boldsymbol{\beta}_1' + \boldsymbol{\beta}_2'\boldsymbol{\Psi}\right)$ on $\mathbf{z}_t$: 'classical' omitted-variables bias problem.
- IIS can correct non-constancy of intercept, and hence changes in *estimated* slope and goodness of fit.

'Optimal' solution to intercept shift is step dummy, but infeasible – not known that $\mathbf{x}_t$ is relevant.

Simulation of scalar case with break in z_t and x_t of $5\sigma_\epsilon$ at $T^0 = 81$, $T^* = 91$ shown in Figure 82.

IIS mimics step-shift dummy for induced shifts.

Non-constant z_t and x_t, $\delta_2 = 5$, $\mu_2 = -5$.



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 estimation**

(5) **Automatic model evaluation**

(6) **Embedding theory models**

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

(8) An empirical example: food expenditure

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 5: 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'

Model selection in ADL models

The DGP is given by:

$$y_t = 1.5y_{t-1} - 0.8y_{t-2} + \sum_{j=1}^{6} \left(\beta_j x_{j,t} - \beta_j x_{j,t-1}\right) + \epsilon_t \qquad (32)$$

where $\epsilon_t \sim \mathsf{IN}\left[0, 1\right]$ and $\mathbf{x}_t = (x_{1,t}, \ldots, x_{6,t})'$ is generated by:

$$\mathbf{x}_t = \rho \mathbf{x}_{t-1} + \mathbf{v}_t \text{ where } \mathbf{v}_t \sim \mathsf{IN}_6\left[\mathbf{0}, \mathbf{\Omega}\right] \qquad (33)$$

with $\rho = 0.5$, $\omega_{kk} = 1$, and $\omega_{kj} = 0.5, \forall k \neq j$.
$n = 0, 1, 2, 4, 6, 8, 10, 12, 14$ relevant regressors, with $\beta_k = \psi_k/\sqrt{T} = 8/\sqrt{12T}$, so is just over 2.

The DGP has negative relations between pairs of exogenous regressors (first differences).

Modelling an ADL

The main difficulty for an ADL is choosing the lag length. Here, there are 7 GUMs, given by $s = 0, 1, 2, 5, 10, 15, 20$:

$$y_t = \mu + \sum_{k=1}^{s} \alpha_k y_{t-k} + \sum_{j=1}^{6} \sum_{k=0}^{s} \gamma_{j,k} x_{j,t-k} + e_t. \tag{34}$$

with $N = 7, 14, 21, 42, 77, 112, 147$ regressors and $T = 100$

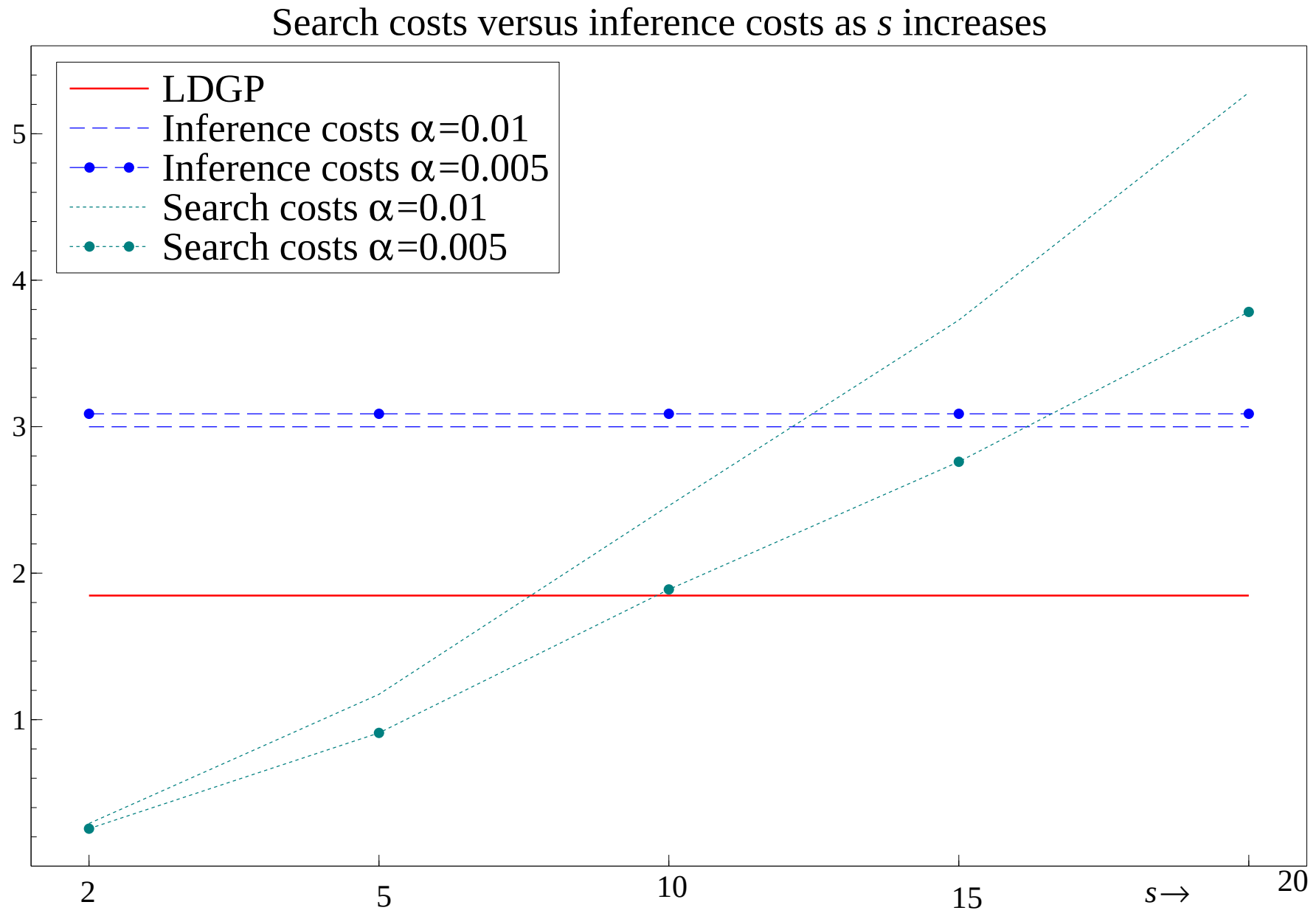
Thus, there are cases with $N < T/2$, N near T, and $N > T$, plus under-specified when $s = 0, 1$ at $\alpha = 1\%,\ 0.5\%$.

We focus on costs of search and inference, using RMSEs for exogenous regressors only, and assess the search costs for lagged y_{t-k} by deviations from the long-run solution:

$$\mathsf{USRMSE}_{\mathsf{LDV}} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} \left(\sum_{k=1}^{s} \widetilde{\beta}_{y,k,i} - \sum_{k=1}^{s} \beta_{y,k,i} \right)^2} \tag{35}$$

so the timing exact of the dynamics is not required.

Search & inference costs as s increases



Under-specification

Now consider cases where $s = 0, 1$ and extend set of DGPs. Let 6 in (32) correspond to $r - 2$ for $r = 3, \ldots, 8$.

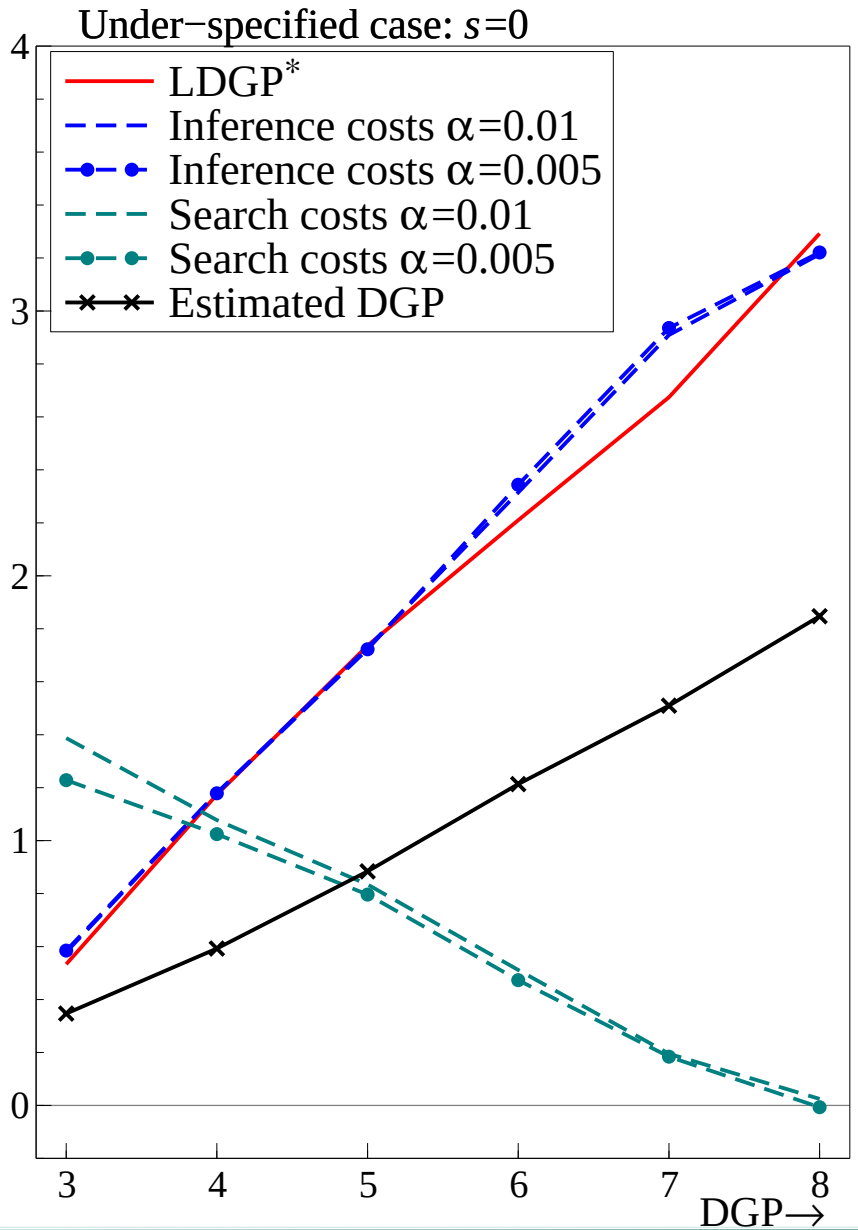
LDGP is joint density of included variables:
a subset creates a less useful reduction of the DGP denoted LDGP*.

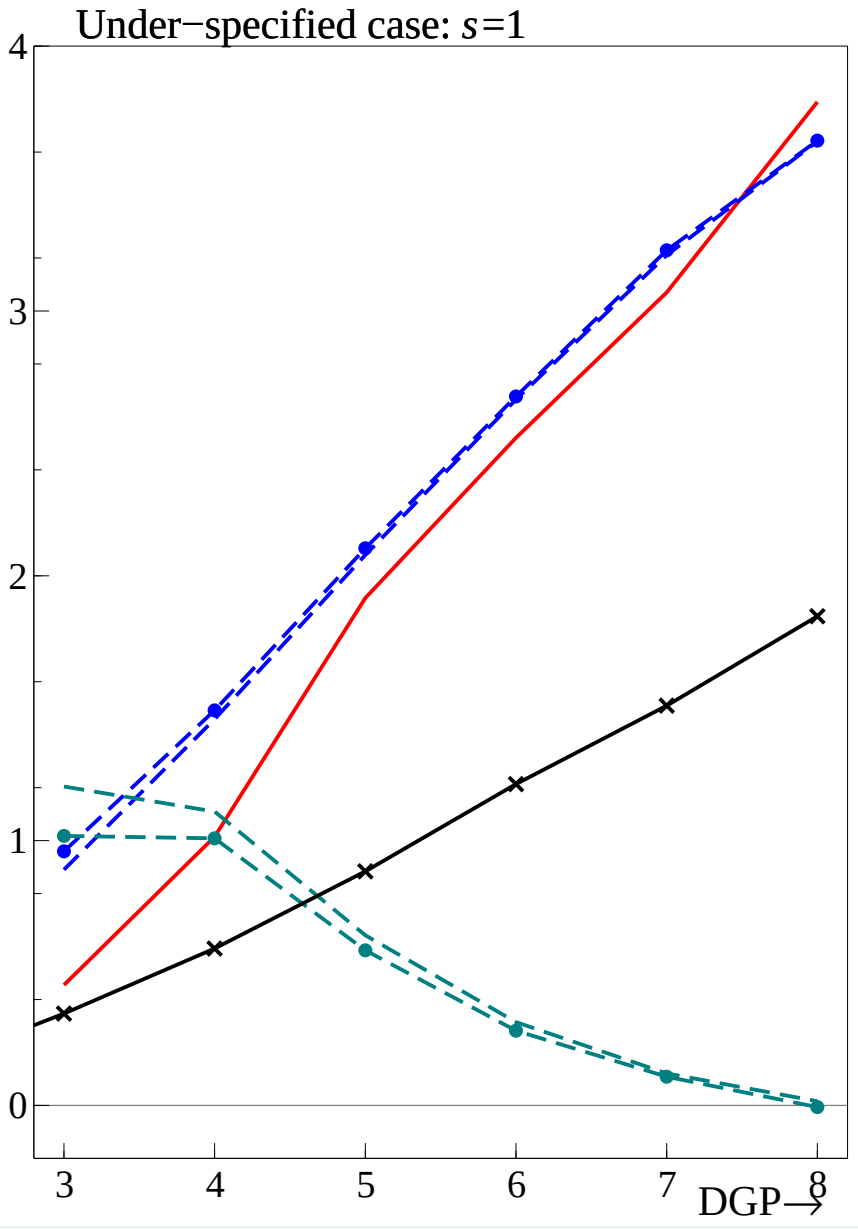
Benchmark for inferences remains DGP parameters, not the induced parameters of LDGP*.

Figure 93 plots search and inference costs at each DGP for the two GUMs with $s = 0, 1$.

As more variables relevant, inference costs increase, **but search costs decline**– and can even be negative.
Higher RMSE costs from just estimating DGP than searching from an incorrect GUM.
Choice of $\alpha = 0.01$ or 0.005 makes almost no difference.

Search & inference costs from under-specification





Implications

(A) The costs of search increase as s increases from more irrelevant variables to search over.

(B) Costs of search increase steadily–almost linearly–despite a shift from $N \ll T$ to $N > T$ between $s = 10$ to $s = 15$.

(C) A tighter significance level results in lower search costs.

(D) At $\alpha = 0.005$, the costs of search are lower than the costs of inference even when $N > T$ ($s = 15$), so there are an additional **98** irrelevant variables.

(E) The costs of inference over the LDGP with no selection are substantial for the larger DGPs.

(F) Advantages of search relatively larger with mis-specification, as costs of estimating DGP can be higher than searching from GUM.

Impulse saturation in fat-tailed distributions

Impulse saturation aims to detect outliers and location shifts: is it 'confused' by a fat-tailed distribution (e.g., t_3)?

Use design in (6) and (7), but (8) becomes:

$$\epsilon_t \sim \left(0.4 \times n^{0.5}\right) \times \mathsf{t}_3, \quad n = 1, \ldots, 10 \tag{36}$$

Autometrics checks normality:
if it rejects, p_d-value of later normality tests is reduced, but may retain irrelevant variables to retrieve original p_d.

Table 6 records average gauges and potencies over all $n = 1, \ldots, 10$ experiments using t_3:
with diagnostic testing at $\mathsf{p}_d = 1\%$,
without, and
with impulse saturation at α.

Diagnostics & impulse saturation for t_3

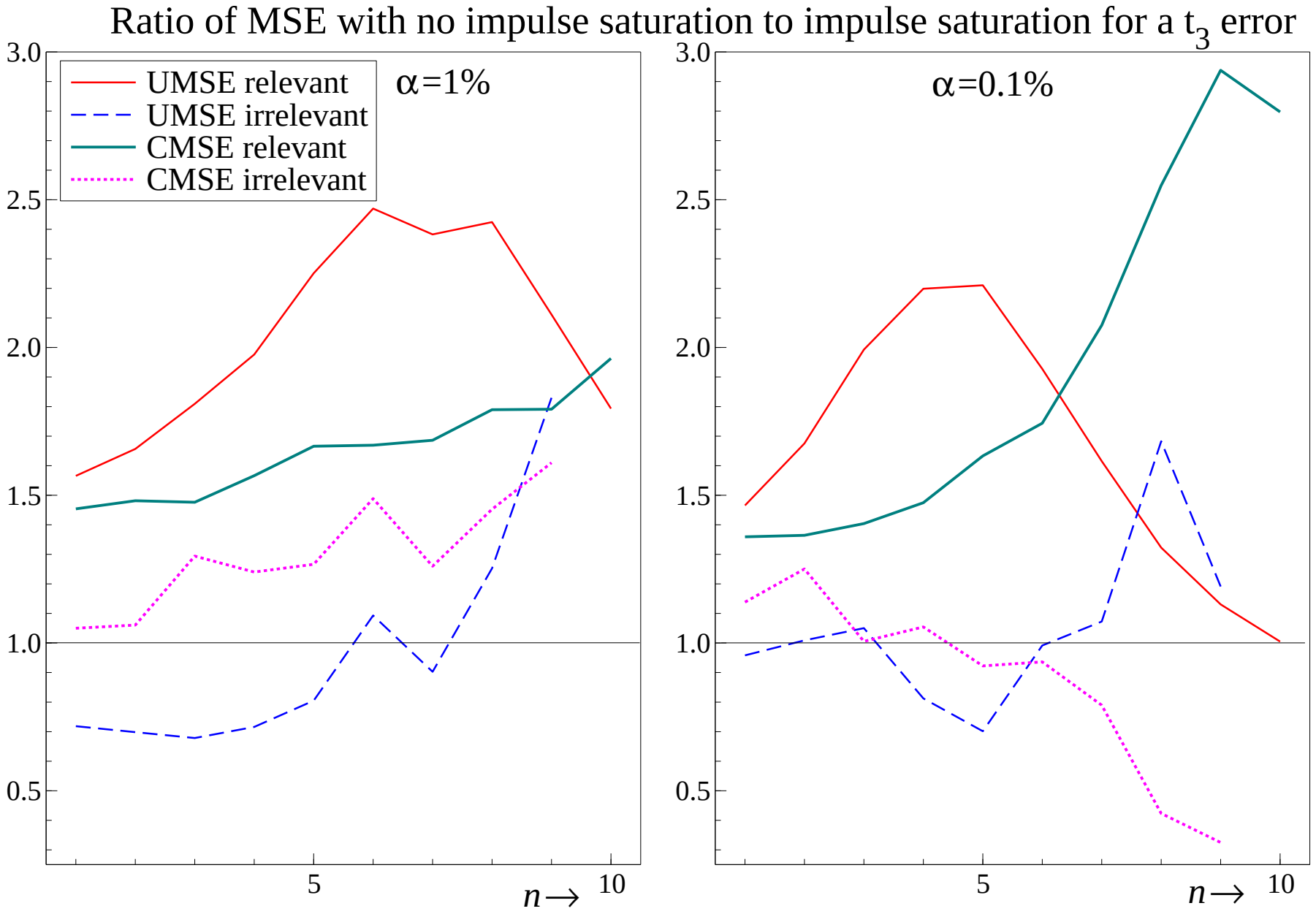
impulse saturation	no	no	yes	no	no	yes
diagnostic tracking	yes	no	no	yes	no	no
α		1%			0.1%	
gauge%	8.6	1.4	4.9	6.8	0.52	1.4
gauge% (no dummies)	–	–	1.9	–	–	0.23
potency%	96.3	96.3	98.8	92.4	92.5	91.4

Table 6: Gauge and potency for t_3 over $n = 1, \ldots, 10$ with and without impulse saturation and diagnostics.

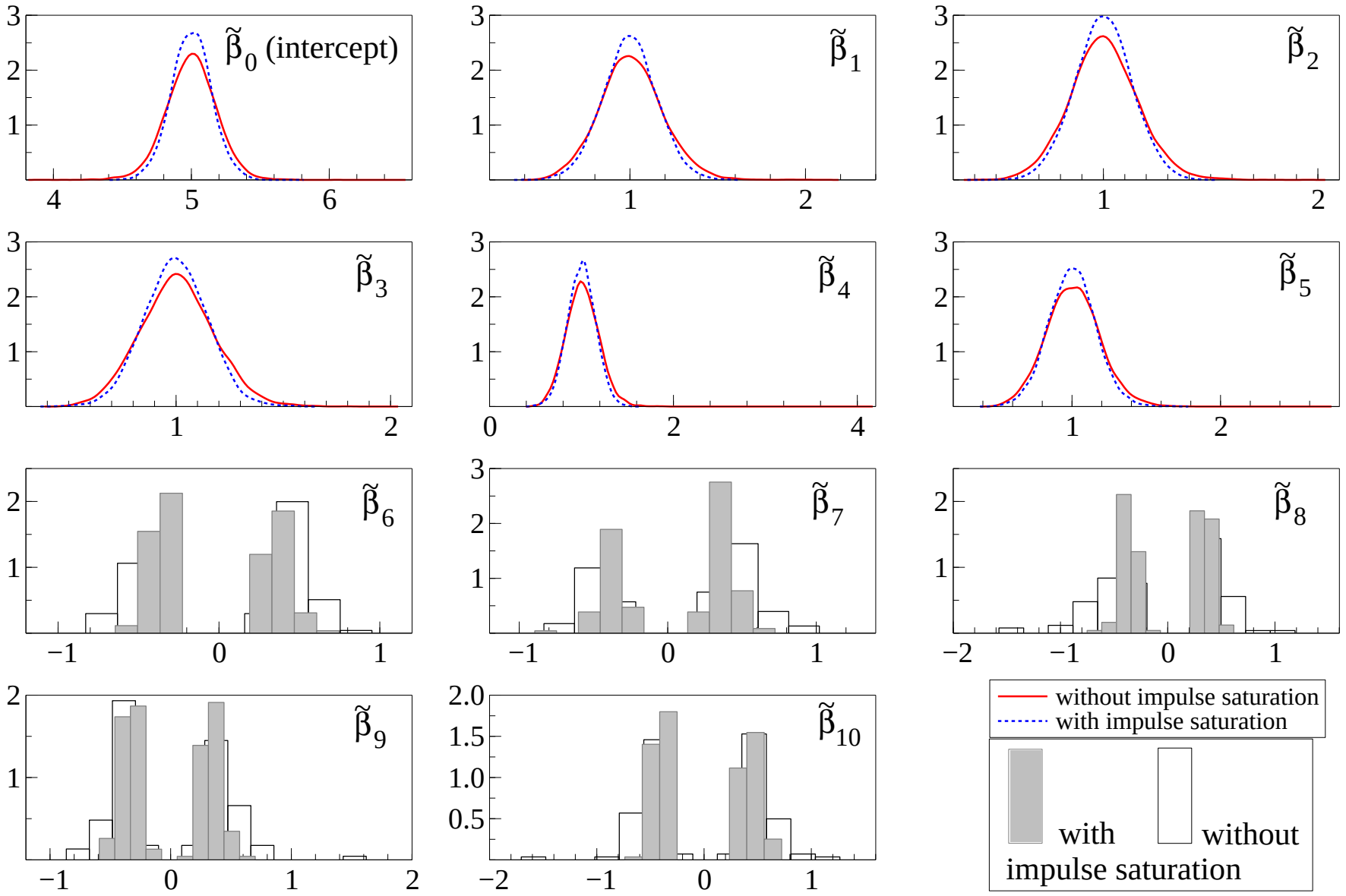
Definition of gauge is ambiguous:
are retained dummies relevant or not, & part of potency?
So, also calculate gauge not counting retained dummies.

If DGP incorrectly assumed to be normal with diagnostic testing, gauge is far too high at tight α: 7% for $\alpha = 0.1\%$.
No diagnostic testing improves gauge, but still too large.

Ratios of MSEs without to with IS for t_3



Conditional distributions with and without IS



Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model estimation**

(5) **Automatic model evaluation**

(6) **Embedding theory models**

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

(8) **An empirical example: food expenditure**

Conclusions

Modelling expenditure on food

Many correct decisions needed for successful modelling:

expenditure depends on **many** relevant variables: incomes, prices, interest rates, taxes, demography, etc.

All effects could vary with changes in 'outside factors': legislation, policy regimes, financial innovation, etc.

Dependence could be linear or non-linear

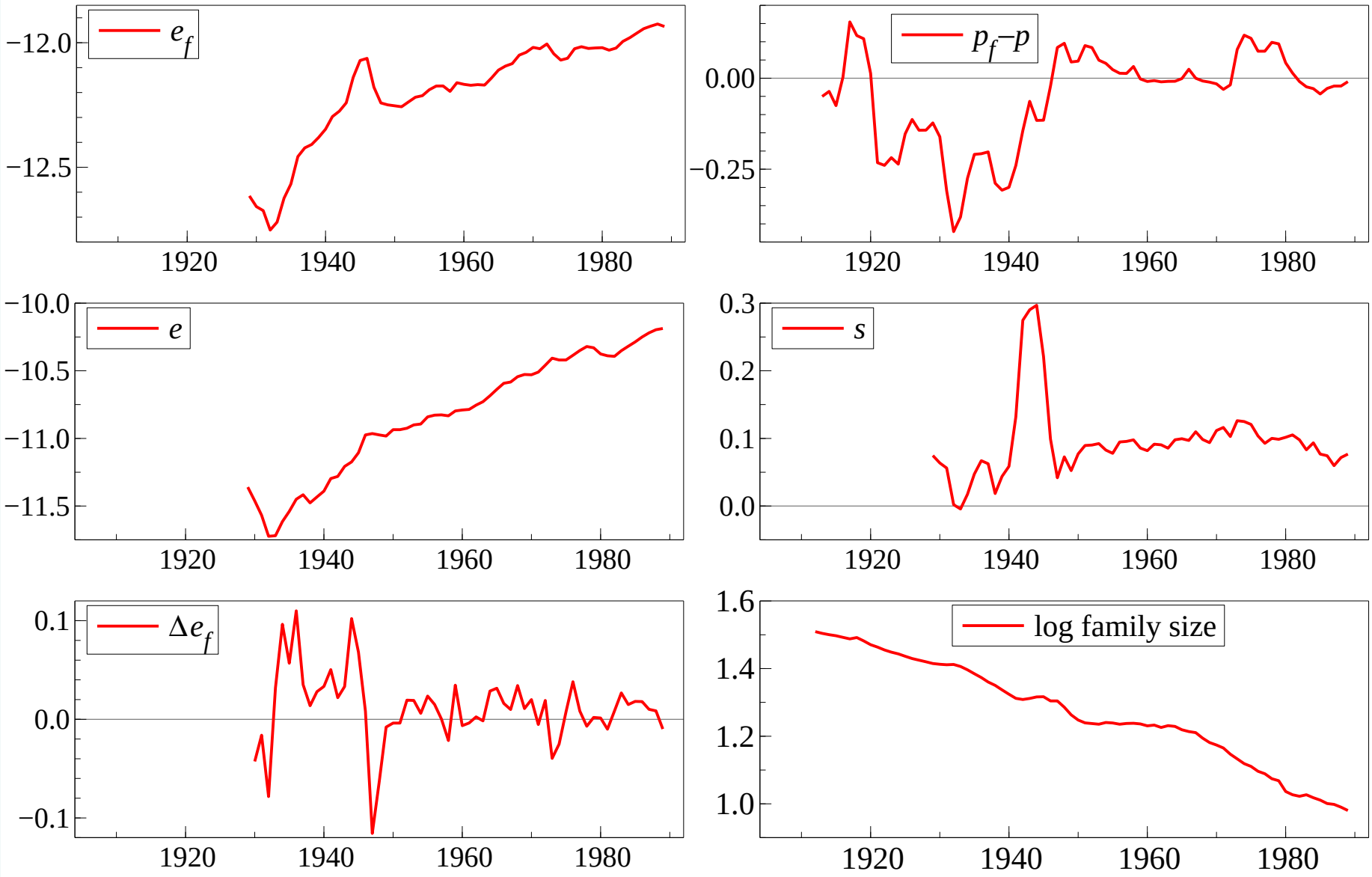
Short-run, long-run and seasonal responses may differ

Relationship may evolve over time

Level of aggregation matters: national or regional, by income levels, categories of transactions, etc.

Non-stationarities entail that any mis-specifications have deleterious effects

US real food expenditure and price data



Modelling problems

'Econometric Experiment' in Magnus & Morgan (1999)
Extension of Tobin (1950): data over 1929–1989
Per capita constant price expenditure on food, e_f, related to: constant price total final expenditure, e; real food prices, p_f; savings rate, s; family size, a; & previous values

Most participants abandoned interwar period:
perturbed by Great Depression and Food Relief.

When model reformulated to explain changes:
excellent properties – equation standard error = 0.75%,
no significant diagnostics, yet **fitted to whole sample**

Autometrics equation is similar to Hendry (1999)

Model derived in a fraction of time it took earlier:
invaluable for labour saving.

Can even 'forecast' post-war from 1952 on.
Complete analysis in Hendry and Mizon (2010).

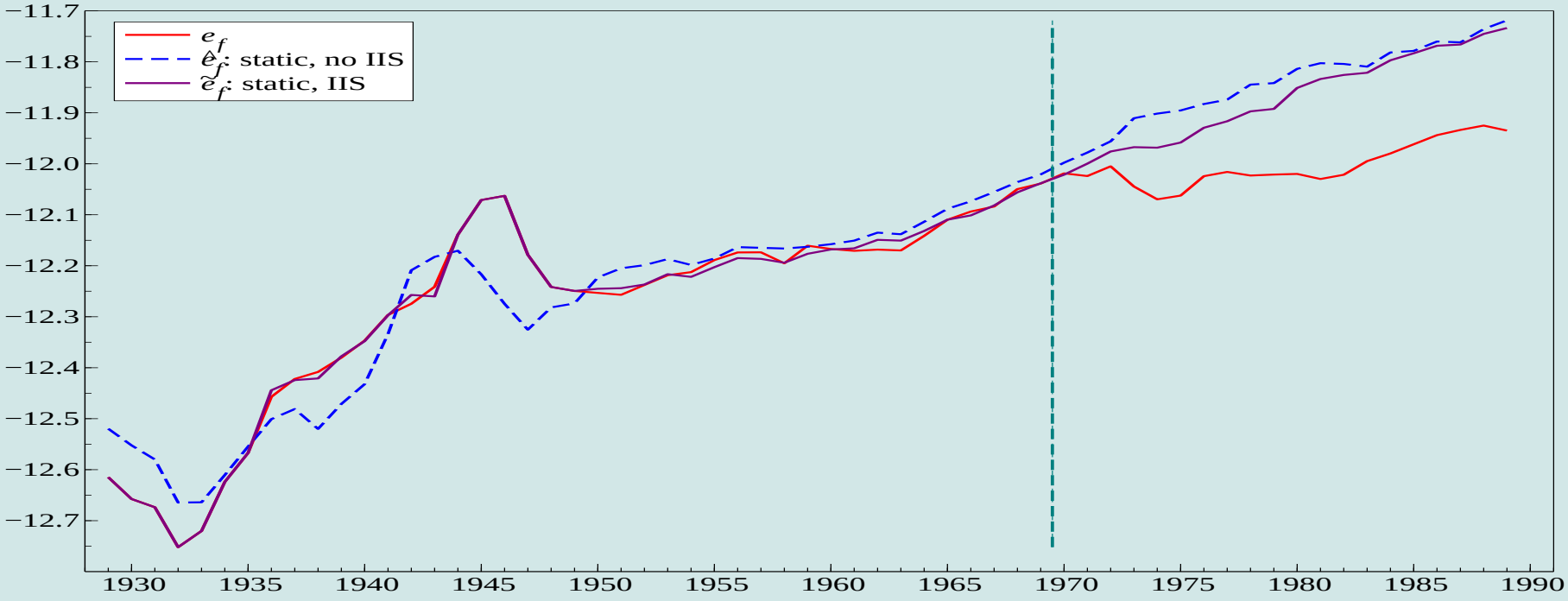
Enforcing theory

$$\underset{}{e_{f,t}} = \underset{}{-}\underset{(0.71)}{7.42} + \underset{(0.09)}{0.45}\, e_t - \underset{(0.13)}{0.034}\, p_{f,t} + \underset{(0.16)}{0.86}\, s_t + \underset{(0.27)}{0.073}\, a_t$$

$$\mathsf{R}^2 = 0.91 \;\; \widehat{\sigma} = 0.066 \;\; \mathsf{F}_{\mathsf{M}}(4,56) = 137.7^{**} \;\; \mathsf{F}_{\mathsf{ar}}(2,54) = 55.6^{**}$$

$$\chi^2(2) = 10.4^{**} \;\; \mathsf{F}_{\mathsf{arch}}(1,59) = 73.65^{**} \;\; \mathsf{F}_{\mathsf{reset}}(2,54) = 14.2^{**}$$

$$\mathsf{F}_{\mathsf{het}}(8,52) = 10.5^{**} \;\; \mathsf{F}_{\mathsf{Chow}}(20,36) = 0.58$$



Selecting

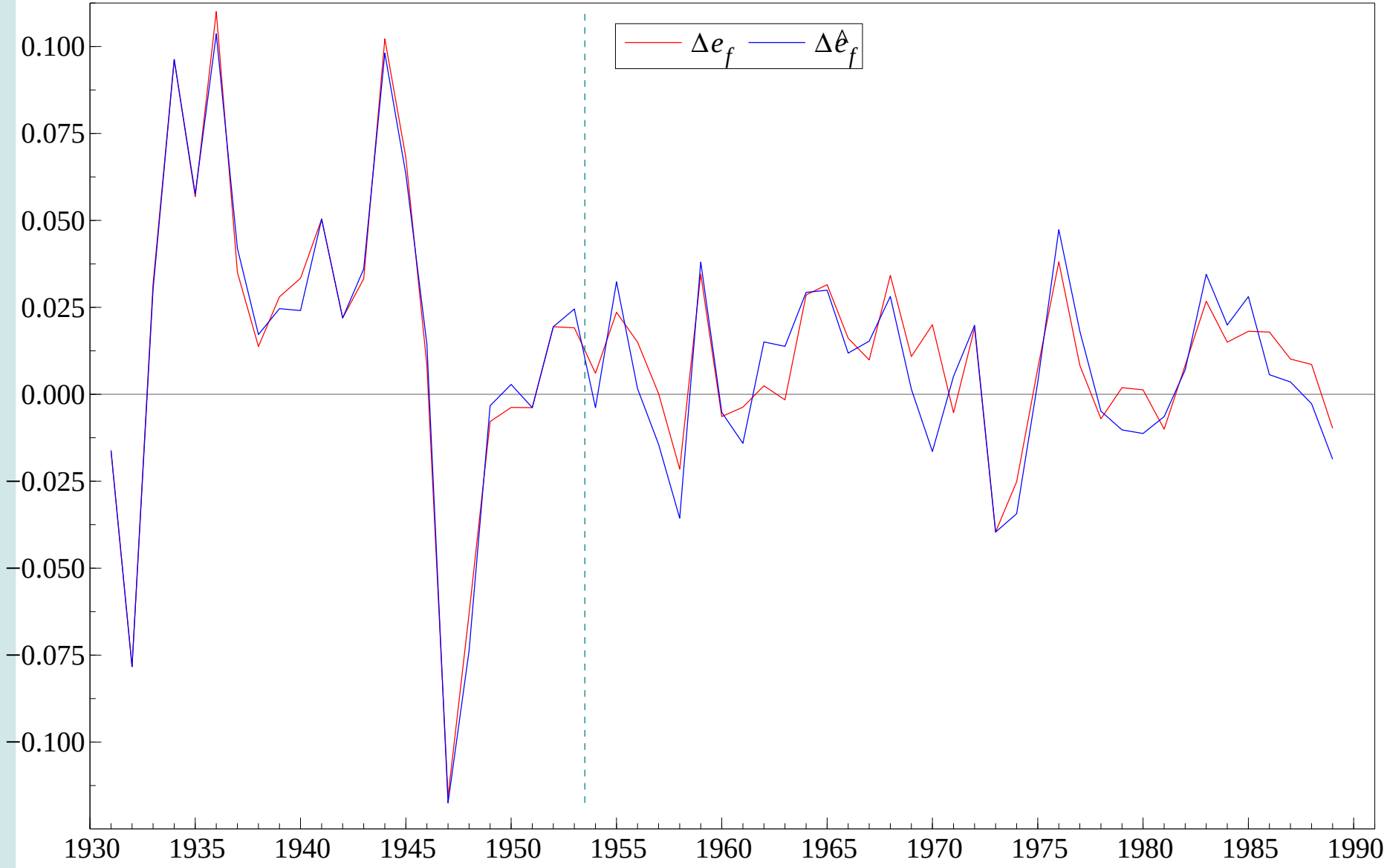
IIS removes 1929–1935 & 1944–1949 but now does reject constancy with $\mathsf{F}_{\mathsf{Chow}}(20,23) = 3.89^{**}$ whereas:

$$\begin{aligned}
\Delta e_{f,t} = {} & \underset{(0.02)}{0.33}\, s_{t-1} - \underset{(0.02)}{0.32}\, c_{0,t-1} + \underset{(0.05)}{0.77}\, \Delta e_t + \underset{(0.03)}{0.11}\, \Delta e_{t-1} \\
& - \underset{(0.04)}{0.69}\, \Delta(p_f - p)_t - \underset{(0.01)}{0.09}\, \mathsf{I}_{31} - \underset{(0.01)}{0.10}\, \mathsf{I}_{32} + \underset{(0.01)}{0.03}\, \mathsf{I}_{34} \\
& + \underset{(0.01)}{0.03}\, \mathsf{I}_{41} + \underset{(0.01)}{0.06}\, \mathsf{I}_{42} + \underset{(0.01)}{0.04}\, \mathsf{I}_{51} + \underset{(0.01)}{0.02}\, \mathsf{I}_{52}
\end{aligned}$$

$$\begin{aligned}
(\mathsf{R}^*)^2 &= 0.99 \;\; \widehat{\sigma} = 0.0067 \;\; \mathsf{F}_{\mathsf{M}}(12,10) = 108^{**} \;\; \mathsf{F}_{\mathsf{ar}}(1,9) = 0.09 \\
\chi^2(2) &= 0.72 \;\; \mathsf{F}_{\mathsf{arch}}(1,21) = 0.07 \;\; \mathsf{F}_{\mathsf{reset}}(2,8) = 2.39 \\
& \mathsf{F}_{\mathsf{Chow}}(36,10) = 1.82
\end{aligned}$$

$$c_0 = e_f + 7.99 - 0.4e + 0.36(p_f - p) \tag{37}$$

‘Forecasting’ the post-war period



Super exogeneity in food expenditure

Build 'automatic' lagged equations with IIS for e; $p_f - p$; s; a.

Finds 25 new impulse-indicators for:
(e): 1946; ($p_f - p$): 1936, 1937, 1940, 1950, 1958, 1967, 1978; (s): 1943, 1968, 1984, 1987; (a): 1947, 1954, 1957, 1961, 1963; ($p_f - p$,s): 1944, 1945, 1973; (s,a): 1949; (e,a): 1980; (e,$p_f - p$,s): 1933, 1938.
Already had (impulses in common with model of e_f in **bold**): **1931** (e,$p_f - p$,s), **1932** (e,s), **1934** ($p_f - p$,s), **1941** (s), **1942** ($p_f - p$,s), **1951** ($p_f - p$), 1952, 1970.
Adding 25 to selected model of e_f yields test:
$\mathsf{F}_{22}^{25} = 1.63$ ($\mathsf{p} = 0.13$)
Does not formally reject, but common impulses do.
However, no impulses in common in the post-1952 period yet constant across that sub-sample.

Route map

(1) **Discovery in general**

(2) **Automatic model extension**

(3) **Automatic model selection**

(4) **Automatic model estimation**

(5) **Automatic model evaluation**

(6) **Embedding theory models**

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

(8) **An empirical example: food expenditure**

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;
2. embed theory-model as a 'forced' specification;
3. select congruent, parsimonious encompassing model;
4. compute near-unbiased parameter estimates; and
5. stringently evaluate results.

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

Little difficulty in eliminating almost all irrelevant variables from the GUM (a small cost of search).

Avoids huge costs from under-specified models.

Overall conclusions

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.

Theory formulations can be embedded in the GUM, to be retained without selection, with no impact on estimator distributions, despite selecting over $N > T$ variables.

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

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

References

Caceres, C. (2007). Asymptotic properties of tests for mis-specification. Oxford.

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*, forthcoming. Springer.

—— (2010b). A low-dimension test for non-linearity. *JEcts*, DOI:10.1016/j.jeconom.2010.01.006.

—— (2010c). Model selection in under-specified equations with breaks. Oxford.

Castle, J. L., Qin, X., and Reed, W. R. (2009). How to pick the best regression. Canterbury, NZ.

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.

—— (2009). Autometrics. In Castle, and Shephard (2009), pp. 88–121.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

—— (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., 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). An automatic test of super exogeneity. In Watson, M. W. *et al*. (eds.), *Volatility and Time Series Econometrics*, pp. 164–193. OUP.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered. *Econometrics J.*, **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.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference with unit roots. *Ecta*, **58**, 113–144.

Tobin, J. (1950). A statistical demand function for food in the U.S.A. *JRSS*, *A, 113*(2), 113–141.

Wooldridge, J. M. (1999). Specification tests in linear models with integrated processes. In Engle, R. F. & White, H. (eds.), *Cointegration, Causality & Forecasting*, pp. 366–384. OUP.

Retracing route

(1) Discovery in general

(2) Automatic model extension

(3) Automatic model selection

(4) Automatic model estimation

(5) Automatic model evaluation

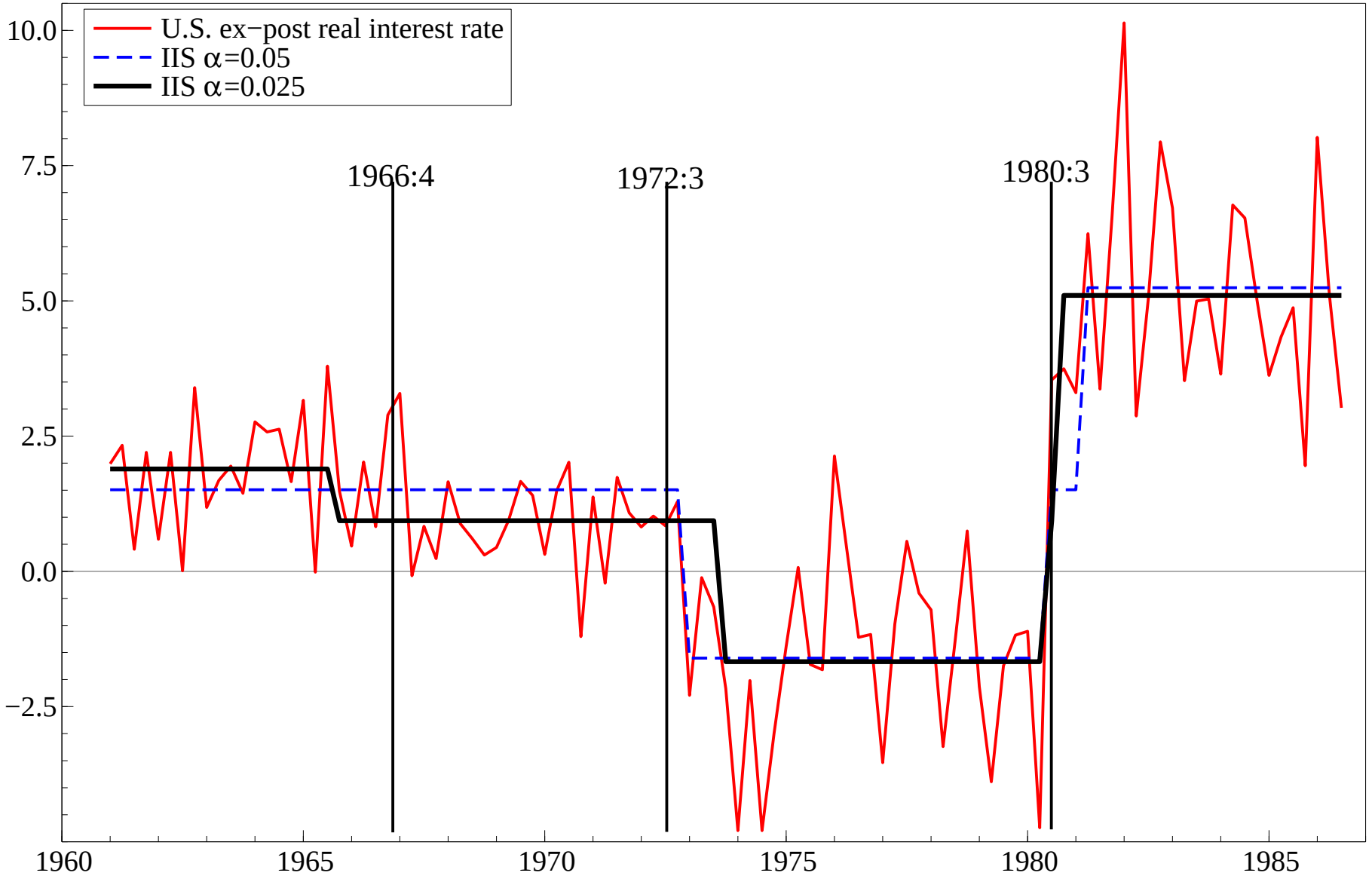
(6) Embedding theory models

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

(8) An empirical example: food expenditure

Conclusion

Original BP and IIS on US real interest rates



Extended BP and IIS on US real interest rates