To: Neil Ericsson

Empirical economic model discovery and theory evaluation
David Hendry, Oxford, United Kingdom

ABSTRACT:
Economies are too high dimensional and wide sense non-stationary for all features of models to be derived by either prior reasoning or data modelling alone. Selecting a viable representation intrinsically involves empirical discovery jointly with theory evaluation. Automatic methods can formulate very general initial specifications with many candidate variables, long lag lengths, and non-linearities, while allowing for outliers and location shifts at every observation, then select congruent parsimonious-encompassing models. Theory-relevant variables are retained without selection, while selecting other candidate variables. Under the null that the latter are irrelevant, by orthogonalizing with respect to the theory variables, estimator distributions of the theory-model's parameters are unaffected by selection, even for more variables than observations and for endogenous variables. Under the alternative, when the initial model nests the local data generating process, an improved outcome results from
selection, allowing rigorous evaluation of any postulated models to ascertain their validity.

EMPIRICAL MODEL DISCOVERY AND THEORY EVALUATION

David F. Hendry
Program in Economic Modelling, INET Oxford
March 2014, OxMetrics Conference, Washington

Research jointly with Jennifer Castle, Jurgen Doornik and Søren Johansen

Every decision about:

1. a theory formulation;
2. its implementation;
3. its evidential base;
4. its empirical specification; and
5. its evaluation

involves selection.

Absent omniscience,
selection is inevitable, unavoidable and ubiquitous:
issue is not whether to select, but how to select.

Data generation process (DGP)

Economic mechanism plus measurement system.
Economies are high dimensional, interdependent, heterogeneous, and evolving: a comprehensive specification of all events is impossible.
Aggregation over time, space, commodities, agents, endowments, is essential—but preclude claims to 'truth'.

Local DGP (LDGP) is DGP for n variables $\{\mathbf{x}_t\}$ under analysis:
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.

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.

Empirical models reflect LDGP, not facsimiles:
designed to satisfy–often implicit–selection criteria.

Only **congruence** is on offer in economics:
congruent models match LDGP in all measured attributes.

'True' models in class of congruent models.
Congruence is testable: necessary conditions for structure.

Encompassing: explain the results of all other models.

Theory only provides an object for modelling:
(A) embed that object in the initial **general formulation**;
(B) search for the **simplest acceptable representation**;
(C) **evaluate** the findings.

How to accomplish that? And what are its properties?

Seven categories of evidence matter jointly

(i) many candidate explanatory variables;
(ii) dynamic reactions;
(iii) parameter changes and location shifts;
(iv) relationships may be non-linear;
(v) feedbacks, exogeneity, and expectations;
(vi) evaluating congruence;
(vii) encompassing results of rival models.

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

As macroeconomic variables are highly intercorrelated,
initially need large equations to capture all these effects.

Especially forceful issue when data processes are 'wide sense non-stationary': integrated and not time invariant.

Often leads to more variables N than observations T.

'Catch 22'—if $N > T$, everything cannot be entered from the outset: necessitates iterative search algorithms to eliminate irrelevant.

To resolve conundrum, analysis proceeds in nine stages.

[A] Castle, Doornik, and Hendry (2012), *Evaluating selection***.**

[B] Hendry and Krolzig (2005), *Bias corrections***.**

[C] Doornik (2009), *Autometrics***.**

[D] Johansen and Nielsen (2009), *Impulse indicator saturation***.**

[E] Castle and Hendry (2013), *Selecting non-linearities***.**

[F] Hendry and Johansen (2014), *Embeding theory***.**

[G] Hendry and Doornik (2014), *Empirical Model Discovery and Theory Evaluation* **(forthcoming MIT Press).**

[H] Hendry and Santos (2010), *Testing super exogeneity***.**

Nine stages

1] **'1-cut' selection** for orthogonal designs with $N << T$;
establishes 'good behaviour' of selection *per se*: **[A]**.

2] Selection matters, so derive **bias corrections** for conditional distributions; improves mean-square errors (MSEs): **[B]**.

3] Compare '1-cut' with ***Autometrics*** (applicable to non-orthogonal models); shows *Autometrics* outperforms, & can handle $N > T$: **[C]**.

4] **Indicator saturation** for multiple shifts and outliers;
now $N > T$ must occur: **[D]**.

5] Selecting **non-linearities**: **[E]**.

6] Impact of **mis-specification testing**;
costs of checking congruence small compared to not testing: **[A]**.

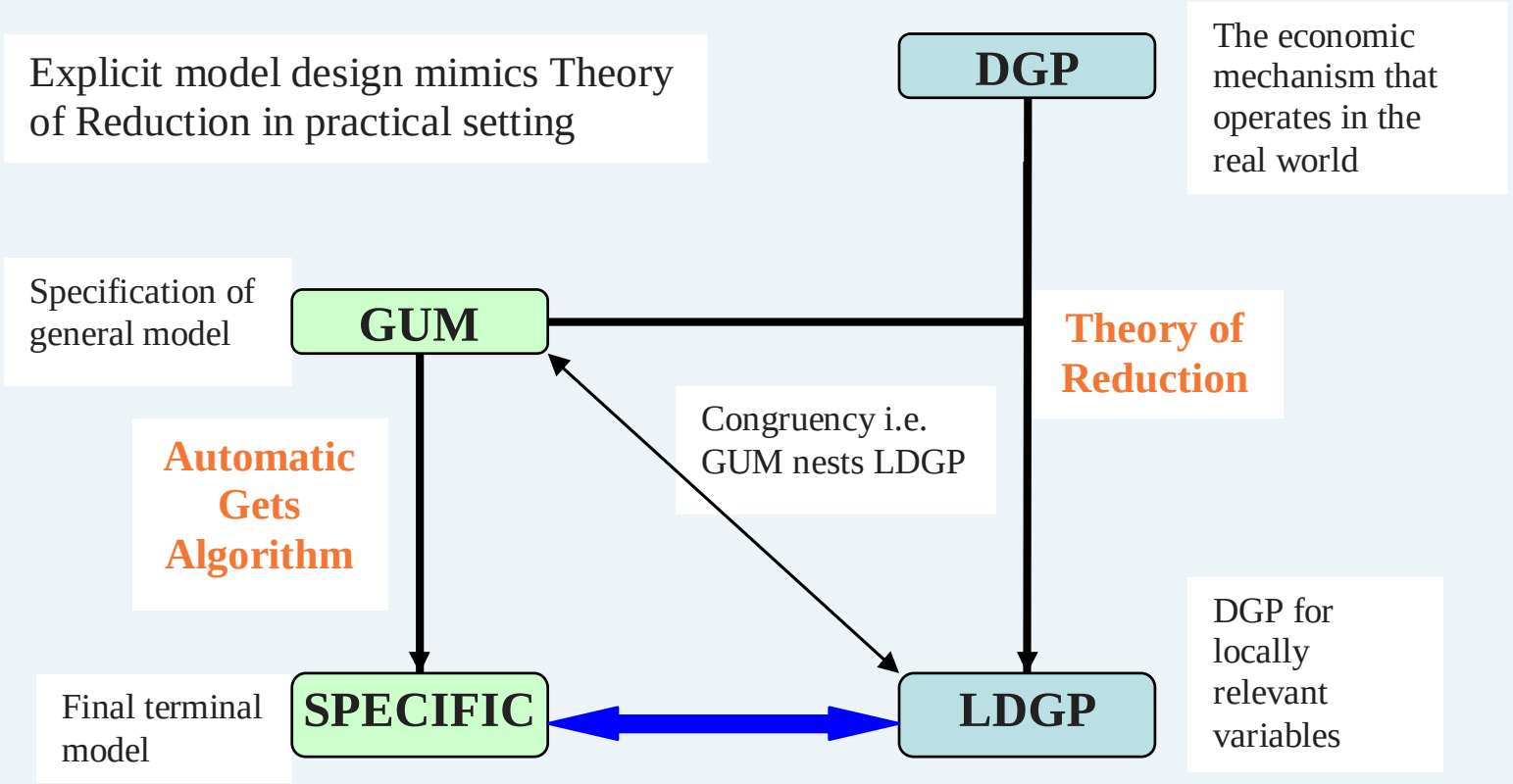
7] Role of **encompassing** in automatic selection;
controls 'good behavior' & avoids missing relevant combinations: **[G]**.

8] Empirical model discovery **jointly with theory evaluation**: **[F]**.

9] Finally, testing **exogeneity** in selected model: **[H]**.

Route map

From DGP to selected model



Aim for final selection that maintains congruence of GUM, and parsimoniously encompasses it, so is 'best' representation of LDGP. Embodied in **PcGive & Autometrics**: see Doornik and Hendry (2013).

Aim for frequency of recovering LDGP starting from GUM same as starting from LDGP.

Two costs of selection: costs of **inference** and costs of **search**.

First inevitable if tests have non-zero null retention and non-unit rejection frequencies under alternative:
applies even if commence from LDGP.

Avoid for theory parameters by embedding theory without search.

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 GUM as against LDGP, plus costs for retained irrelevant variables.

Route map

(1) Selecting empirical models

(2) **Simulating '1-cut' selection**

(3) Automatic model extensions: *Autometrics*

(4) Detecting and modelling multiple location shifts

(5) Mis-specification testing and encompassing

(6) Empirical model discovery and theory evaluation

(7) Modelling UK real wages over the last 150 years

(8) Conclusions

'*gauge*' is null retention frequency of selection statistics.

'*potency*' is average non-null retention frequency.

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

$\widetilde{\beta}_{k,i}$ is OLS estimate after selection ($\widetilde{\beta}_{k,i} = 0$ if $z_{k,t}$ not selected).

$$\begin{aligned} \text{retention rate: } \widetilde{p}_k &= \frac{1}{M}\sum_{i=1}^{M} 1_{(\widetilde{\beta}_{k,i}\neq 0)}, \quad k = 0,\ldots,N, \\ \text{potency:} &= \frac{1}{n}\sum_{k=1}^{n} \widetilde{p}_k, \\ \text{gauge:} &= \frac{1}{N-n+1}\left(\widetilde{p}_0 + \sum_{k=n+1}^{N} \widetilde{p}_k\right). \end{aligned}$$

CMSE is conditional MSE:

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

GUM includes all N variables (1001 here with intercept):

$$y_t = \beta_0 + \beta_1 z_{1,t} + \cdots + \beta_{1000} z_{1000,t} + \nu_t \qquad (1)$$

DGP is given by:

$$y_t = \beta_1 z_{1,t} + \cdots + \beta_{10} z_{10,t} + \epsilon_t, \quad (2)$$

$$z_t \sim \mathsf{IN}_{1000}[0, \Omega], \quad (3)$$

$$\epsilon_t \sim \mathsf{IN}[0, 1], \quad (4)$$

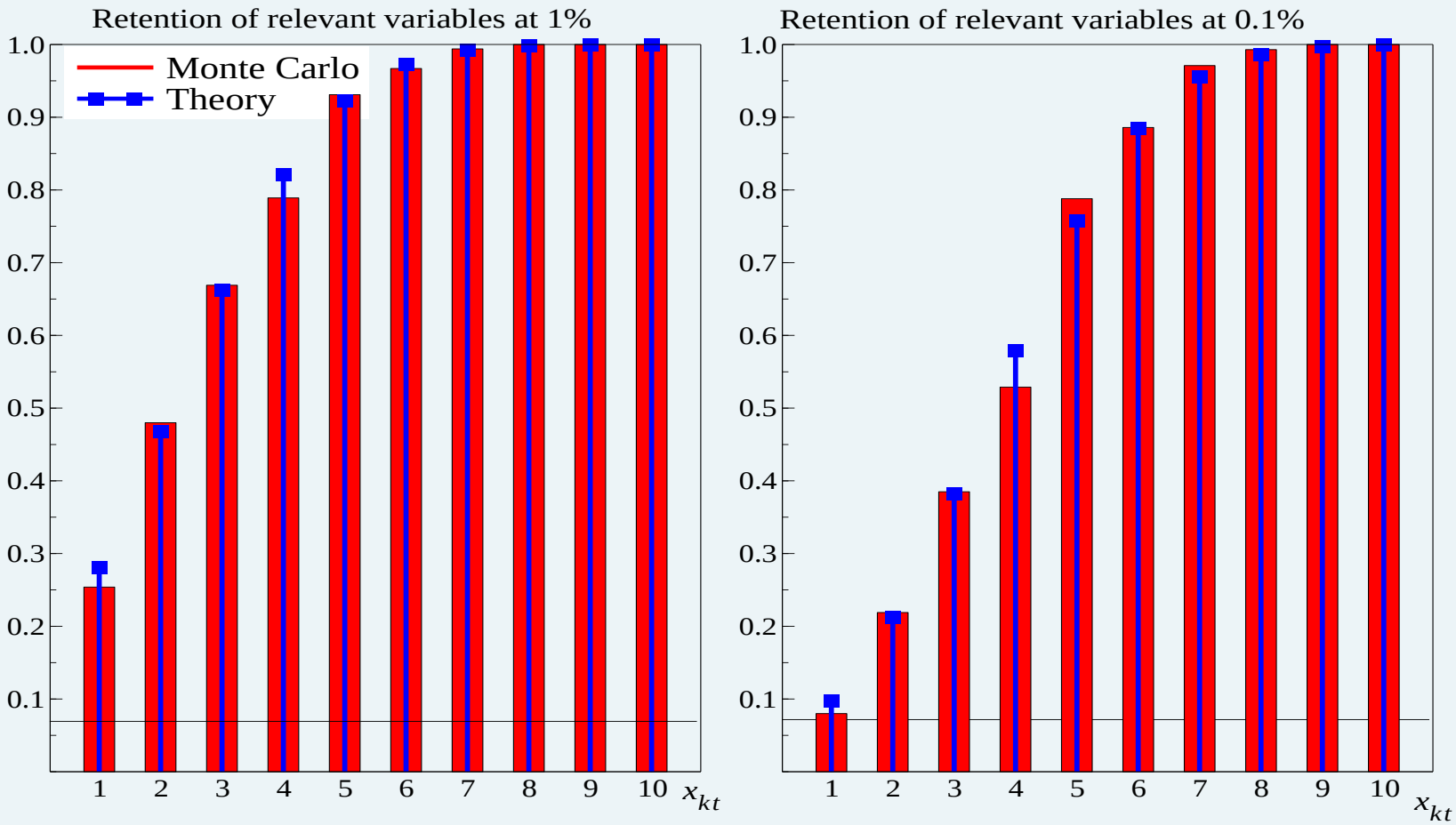
where $z_t' = (z_{1,t}, \cdots, z_{1000,t})$, $\Omega = I_{1000}$, $T = 2000$:
non-centralities of β_i are $\psi_i = 1.5 + 0.5i$ (so 2,...,6.5).

Table : Potency and gauge for 1-cut selection with $N = 1000$ variables.

α	gauge	potency	theory power
1%	1.01%	81%	81%
0.1%	0.10%	69%	68%

Gauges not significantly different from nominal sizes α:
selection is not 'oversized' even with 1000 variables
Potencies close to average theory powers of 0.811 and 0.684.
Close match between theory and evidence even when selecting just 10 relevant regressors from 1000 variables.

1-cut selection for N=1000 at 1%, 0.1%



Retention rates for relevant variables match theory, yet model reduced by about 990 variables on average. Bias corrections when $|t| \geqslant c_\alpha$ improve further.

Remarkable decrease in MSEs of retained irrelevant variables when bias correction–despite not knowing which are irrelevant and which relevant variables. For $N = 1000$ and $n = 10$ in (**??**):

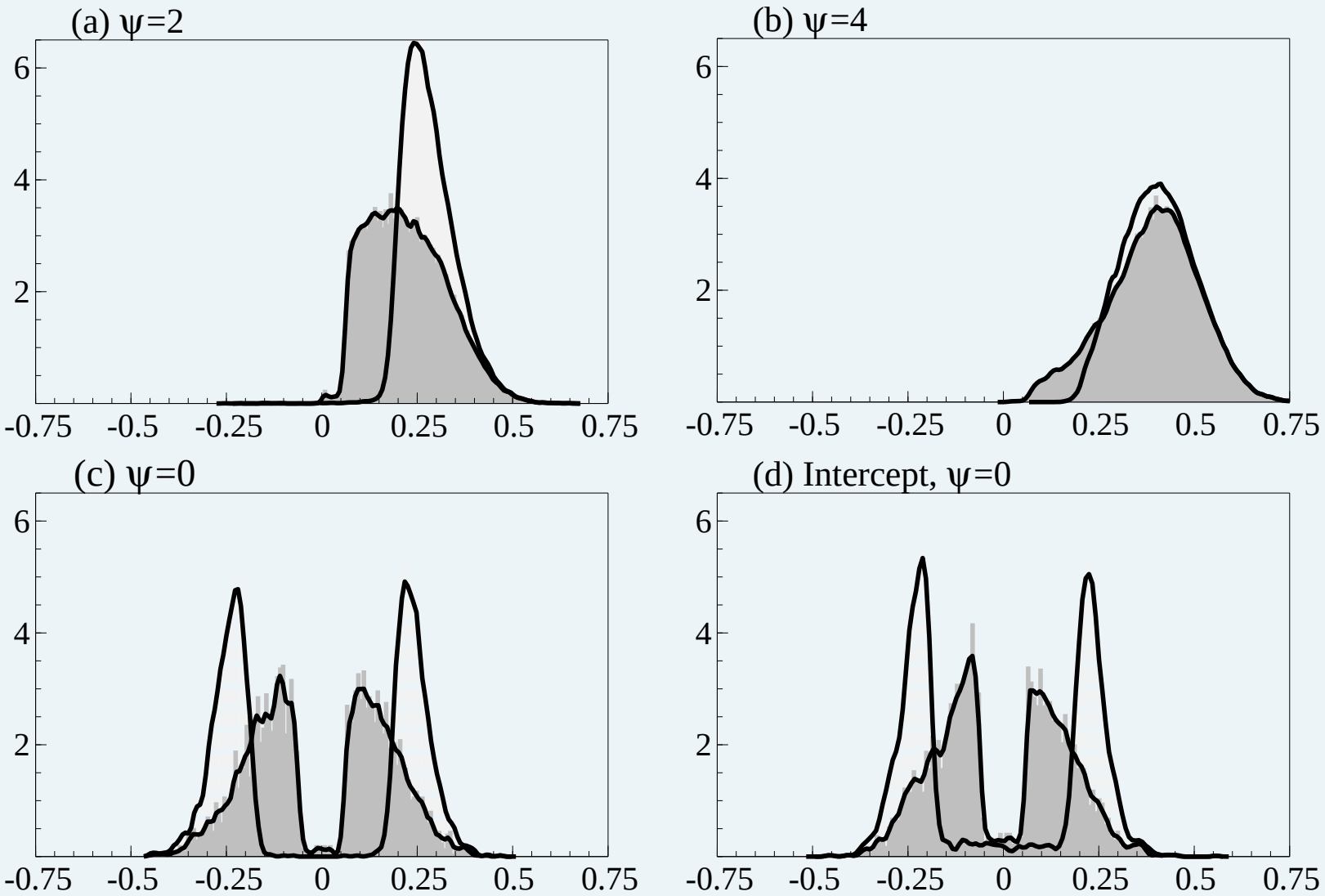
Table : Average CMSEs, times 100, for retained relevant and irrelevant variables (excluding β_0), with and without bias correction.

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
$\widetilde{\widetilde{\beta}}$ after correction	0.38	0.60	1.2	1.3

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

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

Bias correcting conditional distributions at 5%



Route map

(1) Selecting empirical models
(2) Simulating '1-cut' selection
(3) **Automatic model extensions:** *Autometrics*
(4) Detecting and modelling multiple location shifts
(5) Mis-specification testing and encompassing
(6) Empirical model discovery and theory evaluation
(7) Modelling UK real wages over the last 150 years
(8) Conclusions

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) indicator saturation (IIS/SIS) for **parameter non-constancy**.

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

$$y_t = \beta_0 + \sum_{i=1}^{s} \lambda_i y_{t-i} + \sum_{i=1}^{r} \sum_{j=0}^{s} \beta_{i,j} z_{i,t-j} + \epsilon_t \tag{5}$$

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

$$\mathbf{x}_t = \gamma + \sum_{j=1}^{s} \mathbf{\Gamma}_j \mathbf{x}_{t-j} + \epsilon_t \tag{6}$$

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

(ii) Approximate non-linearity by functions of **principal components** $\mathbf{w}_t$ of the $\mathbf{z}_t$: Castle and Hendry (2010).

Let $\mathbf{z}_t \sim \mathsf{D}_n[\boldsymbol{\mu}, \boldsymbol{\Omega}]$, where $\boldsymbol{\Omega} = \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}'$ with $\mathbf{H}'\mathbf{H} = \mathbf{I}_n$.

Then $\mathbf{w}_t^* = \mathbf{H}'\mathbf{z}_t \Rightarrow \mathbf{w}_t^* \sim \mathsf{D}_n[\mathbf{H}'\boldsymbol{\mu}, \boldsymbol{\Lambda}]$.

Empirically $\widehat{\boldsymbol{\Omega}} = \mathsf{T}^{-1}\sum_{t=1}^{\mathsf{T}}(\mathbf{z}_t - \bar{\mathbf{z}})(\mathbf{z}_t - \bar{\mathbf{z}})' = \widehat{\mathbf{H}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{H}}'$

so that $\mathbf{w}_t = \widehat{\mathbf{H}}'(\mathbf{z}_t - \bar{\mathbf{z}})$.

Implemented by squares, cubics and exponential functions:

$u_{1,i,t} = w_{i,t}^2$; $u_{2,i,t} = w_{i,t}^3$; $u_{3,i,t} = w_{i,t}e^{-|w_{i,t}|}$.

When $\boldsymbol{\Omega}$ is non-diagonal, each $w_{i,t}$ is a linear combination of every $z_{i,t}$, so $w_{i,t}^2$ involves squares and cross-products of every $z_{i,t}$ etc.

Number of potential regressors for cubic polynomials is:

$$\mathsf{M}_{\mathsf{K}} = \mathsf{K}(\mathsf{K}+1)(\mathsf{K}+5)/6.$$

Explosion in number of terms as K increases:

K	1	5	10	20	30	40
M_{K}	3	55	285	1539	5455	12300

Quickly reach huge M_{K}: **but only** $3\mathsf{K}$ **if use** $u_{k,i,t}$, $k = 1, 2, 3$.

Route map

(1) Selecting empirical models
(2) Simulating '1-cut' selection
(3) Automatic model extensions: *Autometrics*
(4) **Detecting and modelling multiple location shifts**
(5) Mis-specification testing and encompassing
(6) Empirical model discovery and theory evaluation
(7) Modelling UK real wages over the last 150 years
(8) Conclusions

'Portmanteau' approach to detect location shifts anywhere in sample **while also selecting over many candidate variables, lags etc**.

Impulse-indicator saturation

IIS creates complete set of indicator variables:
$\{1_{\{j=t\}}\} = 1$ when $j = t$, and 0 otherwise for $j = 1, \ldots, T$.
Add all T indicators to set of candidate variables when T observations.

Feasible 'split-sample' algorithm:
Hendry, Johansen, and Santos (2008).

Include first half of indicators, record significant on 1-cut:
'dummying out' first $T/2$ observations when estimating parameters.

Omit first half of indicators, include other half, record again.

Combine retained sub-sample indicators, & select significant.

αT indicators selected on average at significance level α.

Chow (1960) test is sub-sample IIS over $T - k + 1$ to T.
Salkever (1976) tests parameter constancy by impulse indicators.

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 β (not selected) and second moment Σ:

$$T^{1/2}(\widetilde{\beta} - \beta) \xrightarrow{D} N_n\left[0, \sigma_\epsilon^2 \Sigma^{-1} \Omega_\alpha\right]$$

Efficiency of IIS estimator $\widetilde{\beta}$ with respect to OLS $\widehat{\beta}$ measured by Ω_α depends on c_α and distribution, but close to $(1-\alpha)^{-1}\mathbf{I}_n$.

Must lose efficiency under null; small loss αT of 1 observation at $\alpha = 1/T$ if $T = 100$, despite T extra candidates.

Potential for major gain under alternatives of breaks and/or data contamination: but can be done jointly with all other selections.

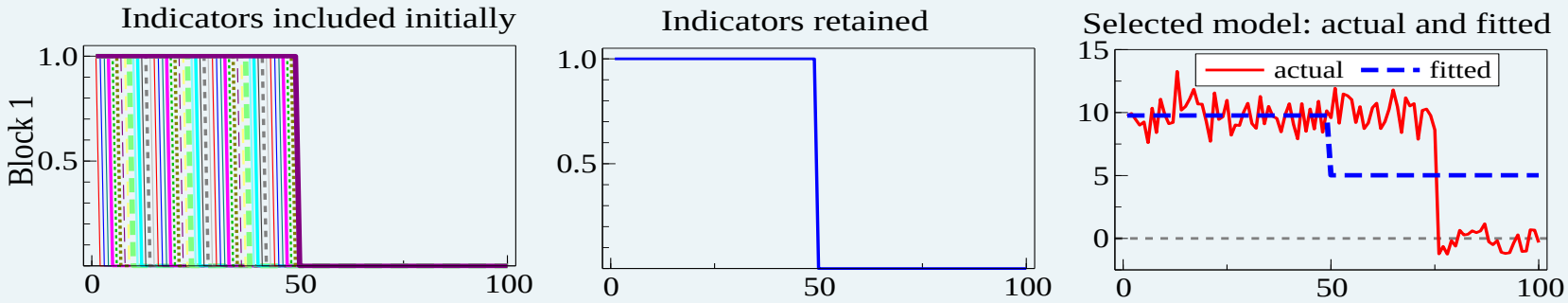
Add a complete set of step indicators $\mathcal{S}_1 = \{1_{\{t \leqslant j\}}, j = 1, \ldots, T\}$, where $1_{\{t \leqslant j\}} = 1$ for observations up to j, and zero otherwise. Step indicators cumulate impulse indicators up to each next observation.

IIS: Impulses SIS: Steps

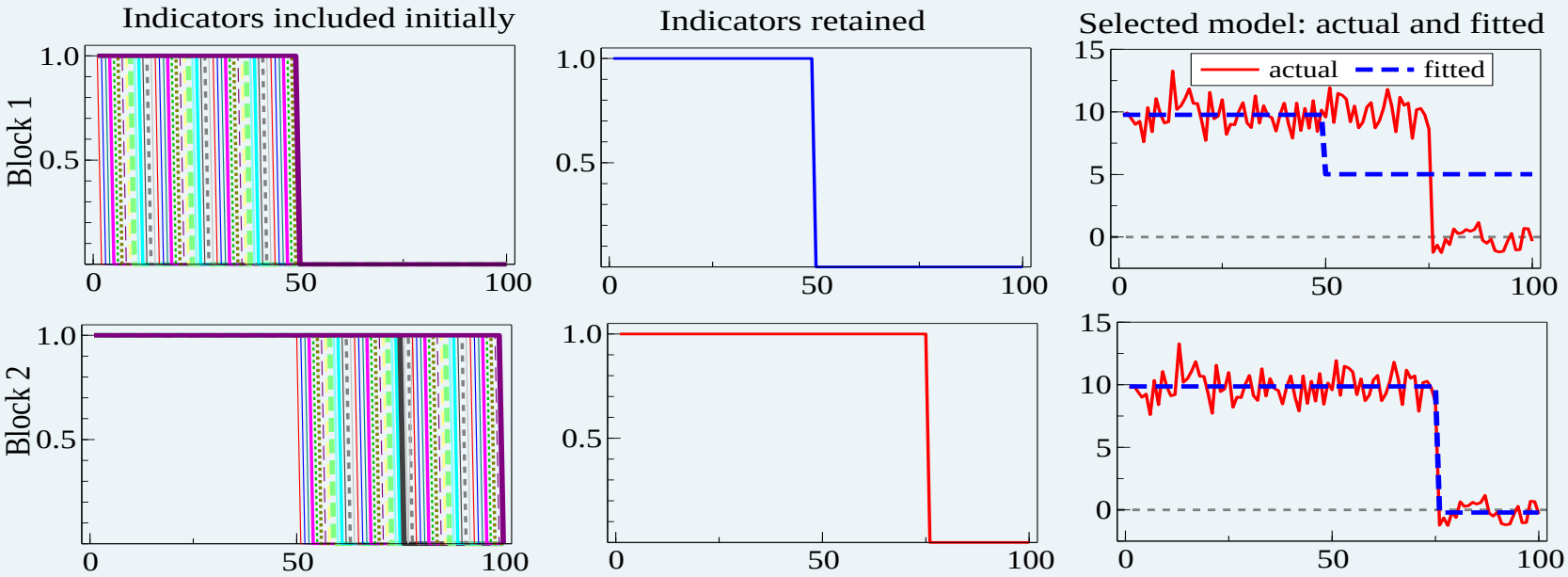
$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \ddots \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

SIS has correct null retention frequency in constant conditional models for a nominal test size of α, and a higher probability than IIS of finding location shifts.

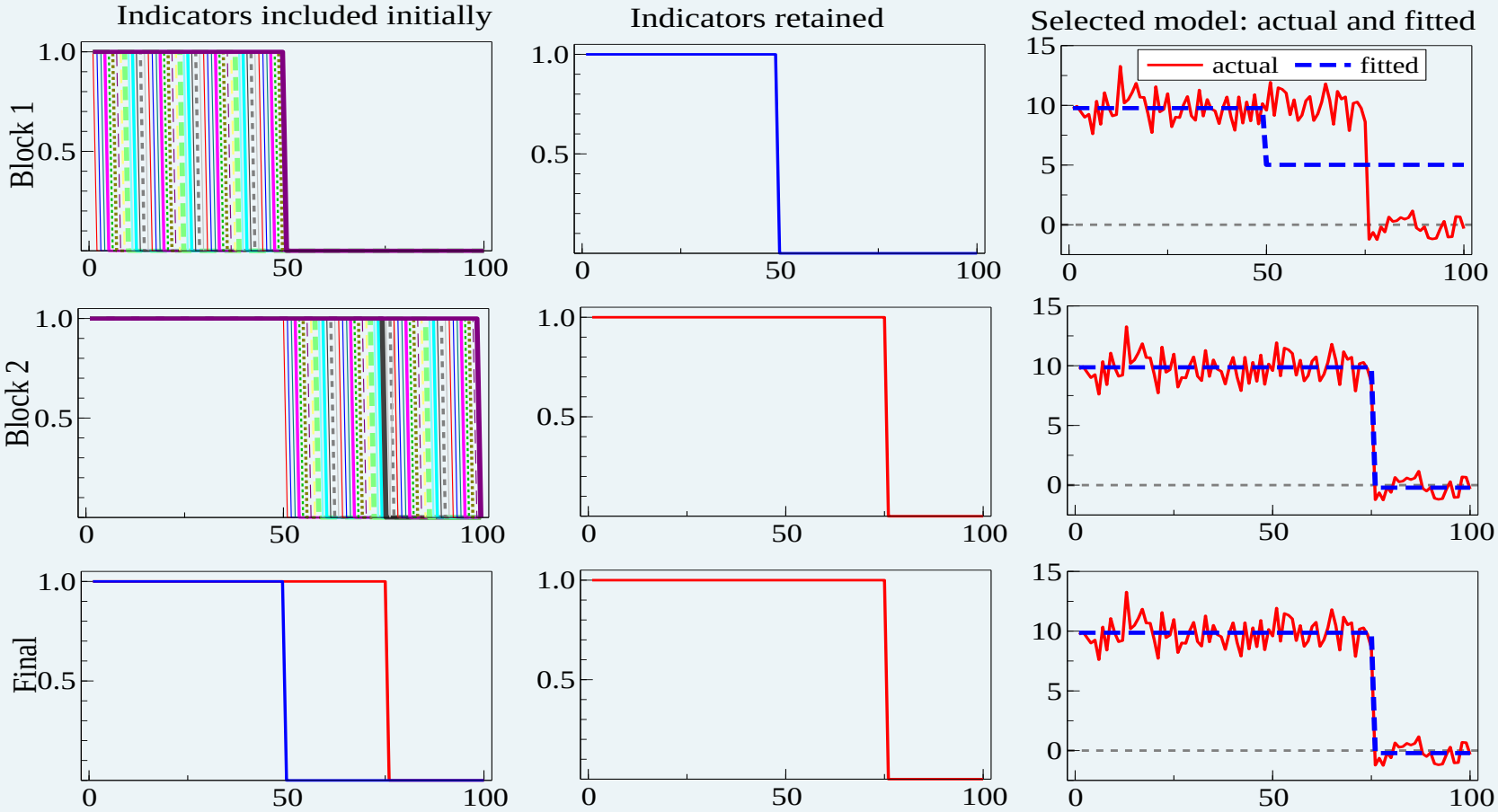
Add half indicators and select ones significant at 1%.



Drop, add other half indicators and again select at 1%.



Combine retained indicators and re-select at 1%.



Initially retains last step as mean shifts down, then finds location shift, so eliminates redundant indicator: just one step needed.

Formulation decisions of which r variables z_t;
their maximum lag lengths (s);
squares, cubics + exponentials in w_t, after orthogonalizing z_t;
location shifts (any number, anywhere) by IIS and/or SIS.
Leads to general unrestricted model (GUM):

$$\begin{aligned} y_t = & \sum_{i=1}^{r}\sum_{j=0}^{s}\beta_{i,j}z_{i,t-j} + \sum_{i=1}^{r}\sum_{j=0}^{s}\kappa_{i,j}w_{i,t-j}^2 + \sum_{i=1}^{r}\sum_{j=0}^{s}\psi_{i,j}w_{i,t-j}^3 \\ & + \sum_{i=1}^{r}\sum_{j=0}^{s}\gamma_{i,j}w_{i,t-j}e^{-|w_{i,t-j}|} + \sum_{j=1}^{s}\lambda_j y_{t-j} \\ & + \sum_{i=1}^{T}\delta_i 1_{\{i=t\}} + \sum_{i=1}^{T-1}\phi_i 1_{\{i\leqslant t\}} + \epsilon_t \end{aligned} \tag{7}$$

$K = 4r(s+1) + s + T$ potential regressors (possibly $(2T-1)$ indicators): bound to have $N > T$–exogeneity considered later.

Route map

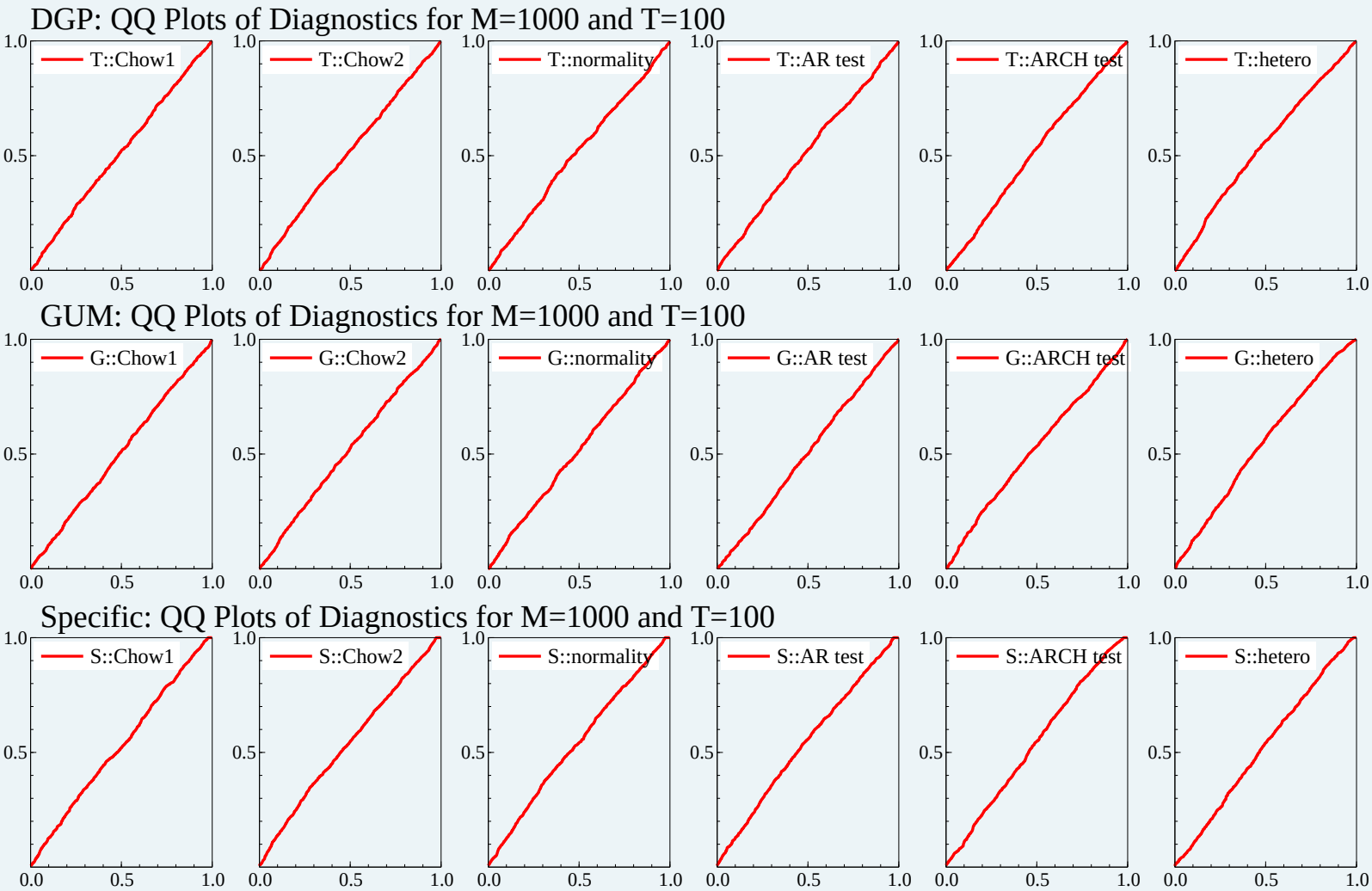
(1) Selecting empirical models
(2) Simulating '1-cut' selection
(3) Automatic model extensions: *Autometrics*
(4) Detecting and modelling multiple location shifts
(5) **Mis-specification testing and encompassing**
(6) Empirical model discovery and theory evaluation
(7) Modelling UK real wages over the last 150 years
(8) Conclusions

Little impact of selection on test statistics.
Small change in quantiles above nominal significance level:
but increasing impact as quantile decreases.
Bound to occur:
models with significant heteroscedasticity not selected.

Not a 'distortion' of sampling properties: decision is taken for GUM.
Conditional on that, no change should occur.

Next Figure reports QQ plots of actual against reference distributions under the null for the main mis-specification tests in DGP, GUM and selected model.

Selection effects on test distributions



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

Implementing system cointegration in *Autometrics*

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

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

Encompassing plays a key role

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

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

If fails, variable retained despite insignificance on t-test, as in Doornik (2008).

***Autometrics* without encompassing loses both gauge and potency**.

***Autometrics* with encompassing is well behaved**:

gauge is close to nominal rejection frequency α.
potency is close to theory maximum of 1-off t-test.

Route map

(1) Selecting empirical models
(2) Simulating '1-cut' selection
(3) Automatic model extensions: *Autometrics*
(4) Detecting and modelling multiple location shifts
(5) Mis-specification testing and encompassing
(6) **Empirical model discovery and theory evaluation**
(7) Modelling UK real wages over the last 150 years
(8) Conclusions

Approach is not atheoretic.

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

Embed initial economic analysis $\mathbf{y} = \mathbf{f}(\mathbf{z})$ in GUM, to be retained without selection, but does not guarantee parameters will be significant.

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.

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

General approach explained in Castle, Doornik, and Hendry (2011).

Correct n valid conditioning variables, $\mathbf{z}_t$, constant parameters $\boldsymbol{\beta}$:

$$y_t = \boldsymbol{\beta}'\mathbf{z}_t + \epsilon_t \tag{8}$$

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

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{t=1}^{T} \mathbf{z}_t\mathbf{z}_t'\right)^{-1} \sum_{t=1}^{T} \mathbf{z}_t\epsilon_t \sim \mathsf{N}_n\left[\boldsymbol{\beta}, \sigma_\epsilon^2\left(\sum_{t=1}^{T} \mathbf{z}_t\mathbf{z}_t'\right)^{-1}\right] \tag{9}$$

Next, $\mathbf{z}_t$ 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:

$$y_t = \boldsymbol{\beta}'\mathbf{z}_t + \boldsymbol{\gamma}'\mathbf{w}_t + \nu_t \tag{10}$$

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

$$\mathbf{w}_t = \widehat{\boldsymbol{\Gamma}}\mathbf{z}_t + \mathbf{u}_t \tag{11}$$

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

$$y_t = \boldsymbol{\beta}'\mathbf{z}_t + \boldsymbol{\gamma}'\mathbf{w}_t + \nu_t = \boldsymbol{\beta}'\mathbf{z}_t + \boldsymbol{\gamma}'\mathbf{u}_t + \nu_t \tag{12}$$

Coefficient of $\mathbf{z}_t$ unaltered.

Consequently:

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

$$\sim \mathsf{N}_{n+k} \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \right)^{-1} \end{pmatrix} \right] \tag{13}$$

as $\sum_{t=1}^{T} \mathbf{z}_t \mathbf{u}_t' = 0$, so distribution of $\widetilde{\boldsymbol{\beta}}$ in (13) **identical** to that of $\widehat{\boldsymbol{\beta}}$ in (9), **unaffected** by model selection.

Only costs of selection are:

(a) chance retentions of some $\mathbf{u}_t$ from selection, controlled by very tight significance levels ($\alpha \leqslant min[0.001, 1/(N + T)]$; and

(b) impact on **estimated** distribution of $\widetilde{\boldsymbol{\beta}}$ through $\widetilde{\sigma}_\epsilon^2$,

offset by bias correcting.

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 $\alpha = 1/N$.

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**.

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

Have answers to every 'seminar question' before they are asked!

$T = 139$, 3 relevant and 37 irrelevant variables: all %

	HP		step-wise		Lasso: BIC		Autometrics	
	HP7	HP8	HP7	HP8	HP7	HP8	HP7	HP8
	1% nominal size							
Gauge	3.0*	0.9*	0.9	3.1	19.5	**35.1**	1.6	1.6
Potency	94.0	99.9	100.0	**53.3**	94.4	86.3	99.2	**100.0**
DGP found	24.6	78.0	71.6	**22.0**	0.1	**0.0**	68.3	**68.8**

* Only counting significant terms (tiebreaker was 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**

Approximating small relevant effects by PCs

To capture potential omissions of individually insignificant relevant effects, add $\mathbf{w}_t$, or principal components, $\mathbf{w}_{1,t}$, of unselected $\mathbf{z}_t$.

Could also reflect common trends modelled by latent factors.

Effective when factor structure of $\mathbf{z}_t$ matches relation between $\mathbf{y}_t$ and $\mathbf{z}_t$ in LDGP: then by representing individually-insignificant effects in $\mathbf{z}_t$ by $\mathbf{w}_{1,t}$, can achieve substantive reductions in RMSEs relative to just estimating the LDGP.

$\alpha = 0.01$	[A]	[B]	[C]	[D]	[E]	[F]
$n = 10; \psi_i = 1; \rho = 0.9, \sigma = 1$						
mean $\widehat{\sigma}$	1.00	1.01	1.01	1.05	1.04	1.04
mean Bias	-0.02	72.0	61.0	-0.02	-0.02	-0.02
mean RMSE	**0.32**	0.75	**0.68**	0.32	0.08	**0.06**

(A) estimating DGP; (B) selection from DGP by *Autometrics*;
(C) bias correction of (B); (D) estimation of factor model;
(E) 1-cut selection from factor model; (F) bias correction of (E).

Route map

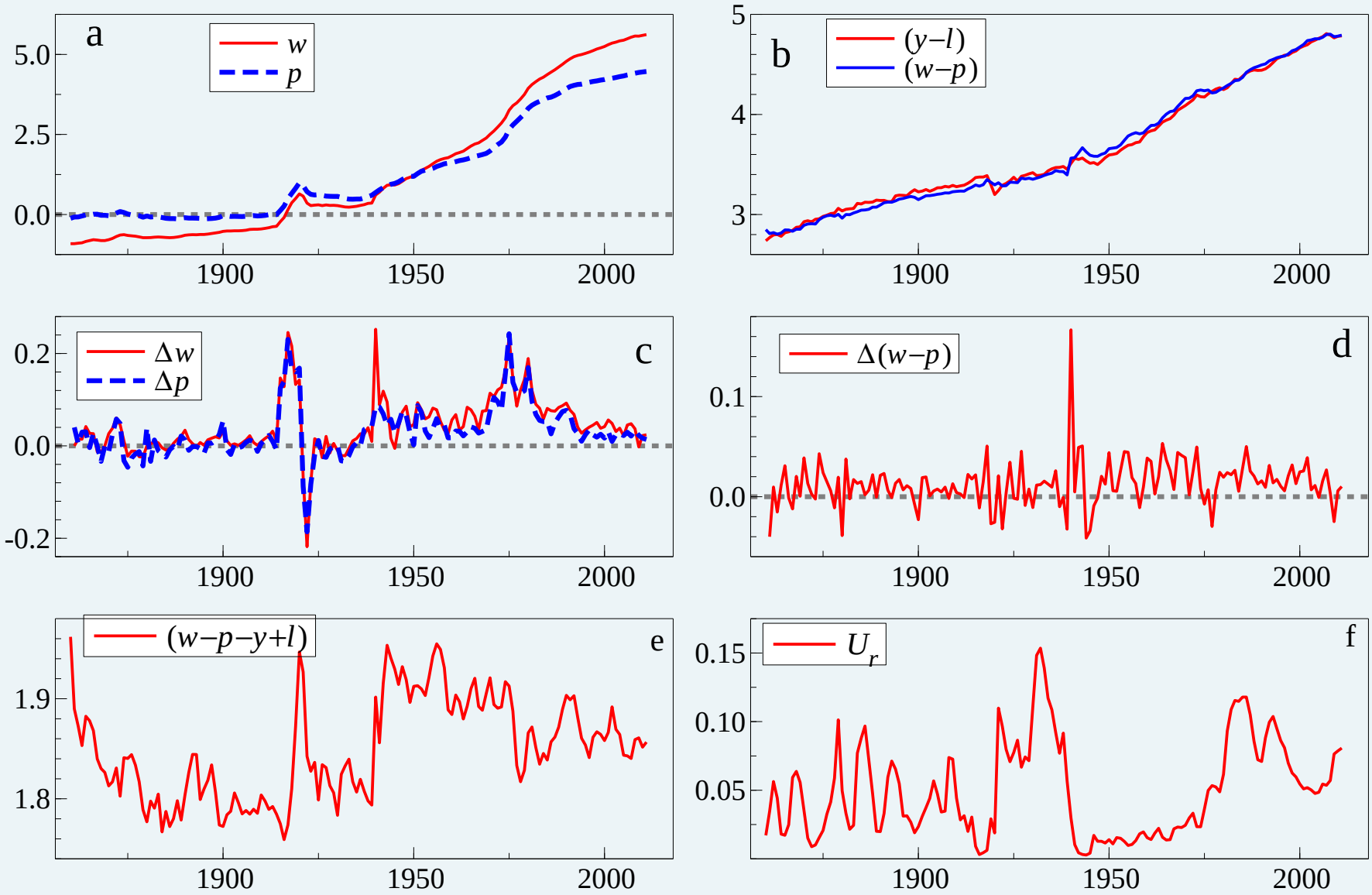
(1) Selecting empirical models
(2) Simulating '1-cut' selection
(3) Automatic model extensions: *Autometrics*
(4) Detecting and modelling multiple location shifts
(5) Mis-specification testing and encompassing
(6) Empirical model discovery and theory evaluation
(7) **Modelling UK real wages over the last 150 years**
(8) Conclusions

Example of empirical model discovery in action.

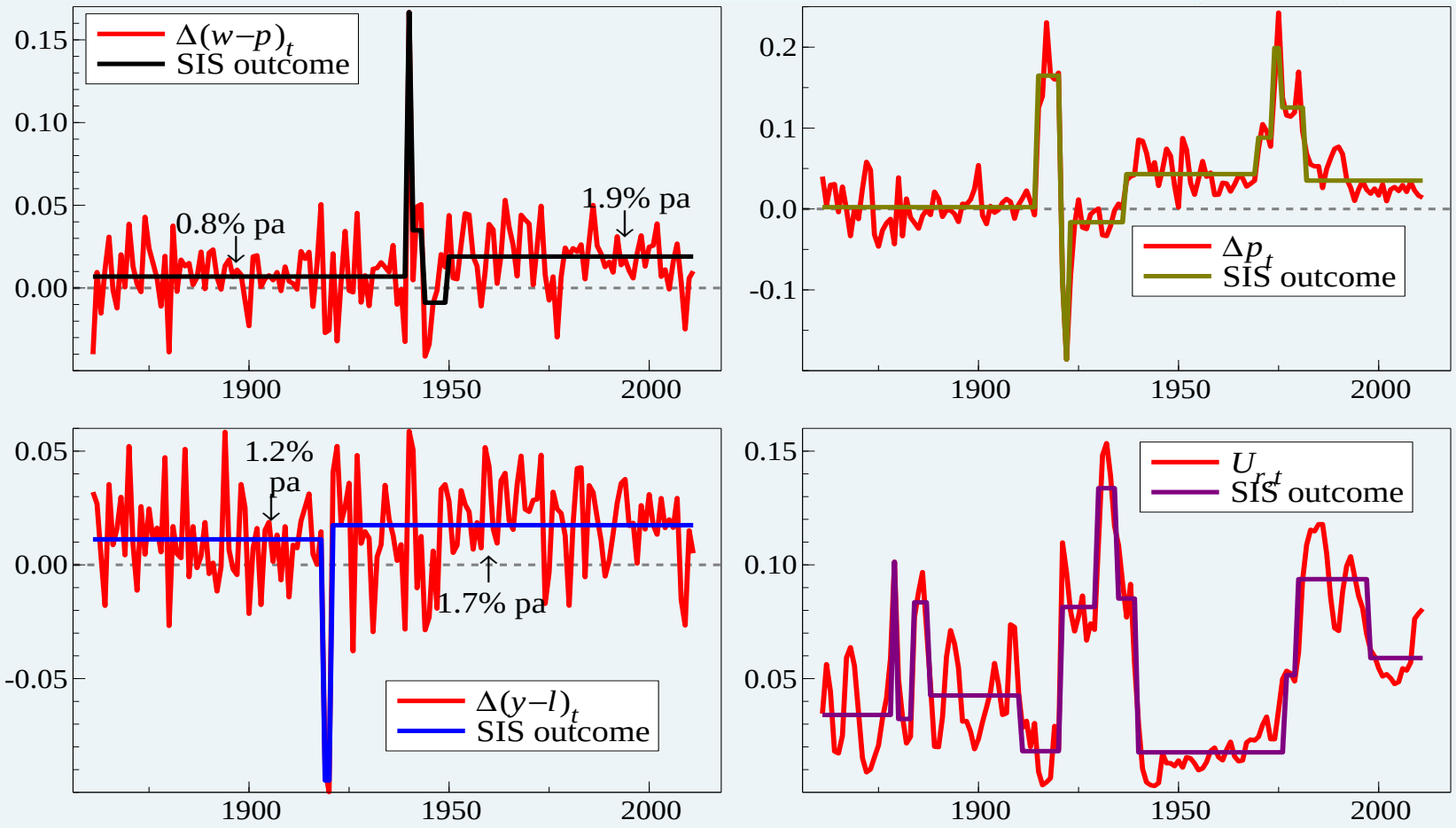
1. Examine roles of many regressors, dynamics, non-linearities, and shifts for integrated data (nominal wages rose by 68,000%).
2. Important wage-price spiral interactions.
3. Non-linear unemployment reaction.
4. Location shifts and outliers tackled by SIS.
5. Test exogeneity of all contemporaneous regressors.
6. Extended data set to forecast real wages over 'Great Recession'.

All aspects must be modelled jointly for a coherent economic explanation, including all substantively relevant variables, their dynamics, shifts, and non-linearities.

Wages, prices, productivity and unemployment



SIS and regime shifts



SIS reveals location shifts unconditionally:
two major shifts in $\Delta(w-p)_t$ and $\Delta(y-l)_t$, but different magnitudes at different times; **huge outliers do not align**.
Location shifts in $U_{r,t}$ and Δp_t also do not match.

Non-linearity test significant at $p = 0.006$ with $F(36, 91) = 1.95$.

Two elements stood out:

non-linear real-wage reaction to inflation represented by:

$$f_t \Delta p_t = \frac{-1}{1 + 1000(\Delta p_t)^2} \Delta p_t.$$

$(u_{r,t} - 0.05)^2$ was an important additional non-linearity.

Selection at $\alpha = 0.001$ for the step indicators, retaining all economic variables (see Hendry and Johansen, 2014),
then selected over those at $\alpha = 0.01$.

No diagnostsic tests significant with $\widehat{\sigma} = 1.04\%$ and RMSFE=**1.05%** over 2005–2011.

Final selection

$$
\begin{aligned}
\Delta(w-p)_t = {} & \underset{(0.003)}{0.021} + \underset{(0.042)}{0.35}\,\Delta(y-l)_t + \underset{(0.034)}{0.12}\,\Delta_2(y-l)_{t-1} - \underset{(0.029)}{0.13}\,\Delta^2 p_{t-1} \\
& - \underset{(0.028)}{0.18}\,(w-p-y+l-\widehat{\mu})_{t-2} - \underset{(0.034)}{0.18}\,(U_{r,t}-0.05) \\
& + \underset{(0.68)}{2.7}\,(U_{r,t}-0.05)^2 - \underset{(0.045)}{0.13}\,\Delta_2 U_{r,t} + \underset{(0.12)}{0.71}\,(f_t \Delta p_t) - \underset{(0.011)}{0.15}\,S_{1939} \\
& + \underset{(0.015)}{0.18}\,S_{1940} - \underset{(0.011)}{0.06}\,S_{1941} - \underset{(0.008)}{0.024}(S_{2011}-S_{1946})\Delta u_{r,t} \\
& - \underset{(0.011)}{0.036}1_{1916} + \underset{(0.006)}{0.027}\,(1_{1942}+1_{1943}-1_{1944}-1_{1945}) - \underset{(0.011)}{0.044}1_{1977}
\end{aligned} \tag{14}
$$

$$
R^2 = 0.82;\quad \widehat{\sigma} = 1.04\%;\quad SIC = -5.85;\quad T = 1864-2004;
$$

$$
\chi^2_{nd}(2) = 2.26;\quad F_{ar}(2,123) = 0.39;\quad F_{arch}(1,139) = 0.49;
$$

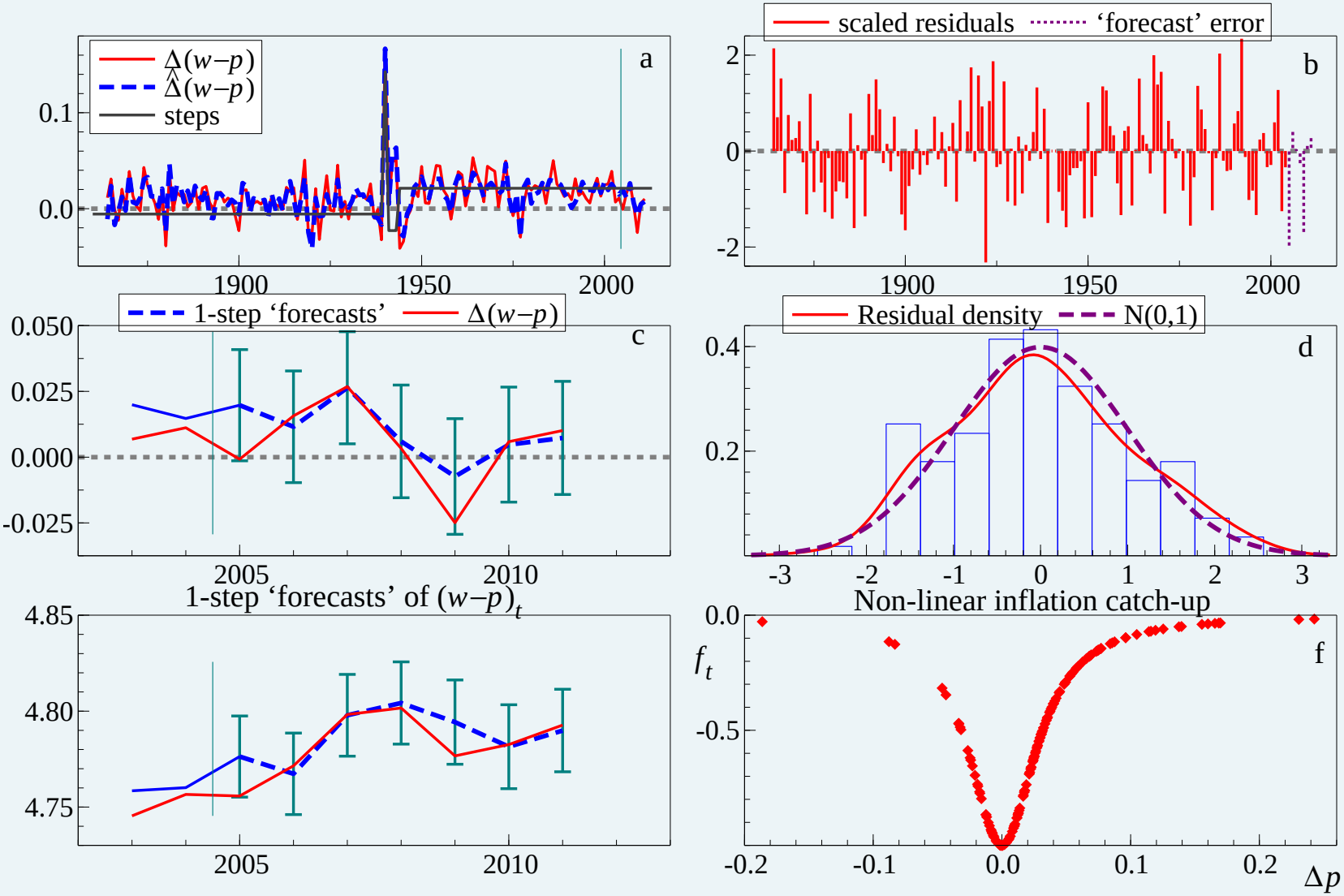
$$
F_{het}(20,116) = 0.82;\quad F_{reset}(2,124) = 2.28;\quad F_{chow}(7,125) = 0.95.
$$

$u_{r,t} = \log(U_{r,t})$ and $\widehat{\mu}$ is the sample mean of $(w-p-y+l)$.
(e.g.) S_{1939} is step indicator: 1 till 1939 and 0 after, etc.

Interpretation

- Short-run impact of changes in productivity is ≈ 0.6
- Strong equilibrium correction of -0.18 from $(w - p - y + l - \widehat{\mu})$
- Coefficient of $f_t \Delta p_t$ highly significant, but < 1
- Non-linearity in unemployment is $-0.42 U_{r,t}(1 - 6.1 U_{r,t})$: negative till unemployment rate exceeds $\approx 15\%$, then **positive**–only consistent with **involuntary unemployment**
- Step indicators needed to explain higher growth rate of real wages post war (1.9% p.a., versus 0.8% p.a. pre-1945), even though $\Delta(y - l)$ is included and has similar behaviour: spike in 1940 was a permanent location shift
- Interactions of variables with step shifts matter as well
- Both steps and impulses mainly for wars
- (14) encompasses previous models
- All mis-specification tests insignificant & constant over 2005–2011
- Super exogeneity of $(y - l)_t$, Δp_t & $U_{r,t}$ in (14) accepted.

Selected model graphical statistics



- SIS used to test exogeneity of the conditioning variables, extending Hendry and Santos (2010).
- Under null of super exogeneity, parameters in conditional model should be invariant to shifts in marginal models: so indicators in latter should not enter former.

VAR in $w - p$, $y - l$, Δp and U_r with SIS at $\alpha = 0.005$; retained indicators in the 3 marginal models tested for significance in (14).

Super exogeneity tests

Variable	null distribution	SIS test on (14)
$(y - l)_t$	$F(2, 123)$	0.77
Δp_t	$F(7, 118)$	1.87
$U_{r,t}$	$F(14, 111)$	1.37
Joint	$F(20, 105)$	1.41

No evidence against super exogeneity of $(y - l)_t$, Δp_t & $U_{r,t}$ in (14).

Route map

(1) Selecting empirical models
(2) Simulating '1-cut' selection
(3) Automatic model extensions: *Autometrics*
(4) Detecting and modelling multiple location shifts
(5) Mis-specification testing and encompassing
(6) Empirical model discovery and theory evaluation
(7) Modelling UK real wages over the last 150 years
(8) **Conclusions**

Little difficulty in eliminating almost all irrelevant variables from the GUM (a small cost of search): HP8 when $\mathsf{N} = 145 > \mathsf{T} = 139$.
Avoids huge costs from under-specified models.

When the LDGP 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 $\mathsf{N} > \mathsf{T}$ variables.

Model selection by *Autometrics* with tight significance levels and bias correction is a successful approach which allows many variables, lags, non-linearities and multiple shifts to be tackled jointly while retaining theory insights.

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

References I

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

Castle, J. L., J. A. Doornik, and D. F. Hendry (2011). Evaluating automatic model selection. *J. Time Series Econometrics 3 (1)*.

——- (2012). Model selection when there are multiple breaks. *J. Econometrics 169*, 239–246.

Castle, J. L. and D. F. Hendry (2010). A low-dimension portmanteau test for non-linearity. *J. Econometrics 158*, 231–245.

——- (2013). Semi-automatic non-linear model selection. In N. Haldrup et al. (Eds.), *Essays in Nonlinear Time Series Econometrics*. OUP.

Castle, J. L. and N. Shephard (Eds.) (2009). *The Methodology and Practice of Econometrics*. Oxford University Press.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica 28*, 591–605.

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

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

Doornik, J. A. and D. F. Hendry (2013). *Empirical Econometric Modelling using PcGive: Volume I* (7th ed.). Timberlake.

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

Hendry, D. F. and J. A. Doornik (2014). *Empirical Model Discovery and Theory Evaluation*. MIT Press.

Hendry, D. F. and S. Johansen (2014). Model discovery and Trygve Haavelmo's legacy. *Econometric Theory*, forthcoming.

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

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

Hendry, D. F. and C. Santos (2010). An automatic test of super exogeneity. In M. W. Watson, T. Bollerslev, and J. Russell (Eds.), *Volatility and Time Series Econometrics*, pp. 164–193. Oxford University Press.

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

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

——- (2013). Outlier detection in regression using an iterated one-step approximation to the Huber-skip estimator. *Econometrics 1*, 53–70.

Salkever, D. S. (1976). The use of dummy variables to compute predictions, prediction errors and confidence intervals. *Journal of Econometrics 4*, 393–397.

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

Sims, C. A., J. H. Stock, and M. W. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica 58*, 113–144.

Wooldridge, J. M. (1999). Asymptotic properties of some specification tests in linear models with integrated processes. In R. F. Engle and H. White (Eds.), *Cointegration, Causality and Forecasting*, pp. 366–384. Oxford University Press.