General-to-Specific Time Series Modelling

Felix Pretis

University of Oxford

Michaelmas 2017

Lecture 4: Model Evaluation and Theory of Reduction

Core References for Lecture 4:

- Hendry (2009)* – Through the Looking Glass
- Hendry (2017)* – Deciding between alternative approaches in macroeconomics
- Doornik (2009) – Autometrics
- Pretis, Reade, and Sucarrat (2016) – GETS in R
- Hendry (1995) Dynamic Econometrics (as overview)
- Hendry and Doornik (2014) (as overview)

"Even if one wants to test an economic hypothesis as to whether some effect is present, partial inference cannot be conducted alone, unless one is sure about the complete absence of all contaminating influences."

Hendry (2009)

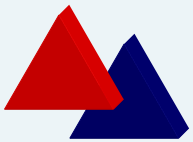
Economies high dimensional, interdependent, heterogeneous, and evolving: comprehensive specification of all events is impossible.

Data generation process (DGP):
economic mechanism plus measurement system.
Local DGP (LDGP) is for set of variables under analysis.

Models reflect LDGP – not facsimiles:
designed to satisfy selection criteria.
Aggregation over time, space, commodities, agents, endowments, essential but precludes claim to 'truth'.

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.

Experimental outputs caused by inputs:

$$\underset{\text{[output]}}{y_t} = \underset{\text{[input]}}{f(z_t)} + \underset{\text{[perturbation]}}{\nu_t} \quad (1)$$

y_t observed when z_t input; $f(\cdot)$ maps inputs to outputs.
ν_t is small, random perturbation.
'Same' outputs repeating experiments at same inputs.

In an econometric model, however:

$$\underset{\text{[observed]}}{y_t} = \underset{\text{[explanation]}}{g(z_t)} + \underset{\text{[remainder]}}{\epsilon_t} \quad (2)$$

y_t **decomposed** into two components:
$g(z_t)$ (bit explained) and ϵ_t (unexplained).
Always feasible even if y_t does not depend on $g(z_t)$.

In econometrics:

$$\epsilon_t = y_t - g(z_t) \tag{3}$$

Thus, models can be **designed** by selection of z_t.

Design criteria must be analyzed:
will lead to notion of *congruent* model.

Successive congruent models must explain earlier:
concept of **encompassing** – whereby progress achieved.

Will repeatedly use:
$P(a, b) = P(a \mid b)P(b)$
Base order on: time, theory and institutional knowledge.

All random variables in economy over $t = 1, \dots T$:
denoted $\{\mathbf{u}_t\}$ with $\mathbf{U}_1^T = (\mathbf{u}_1, \dots, \mathbf{u}_T)$
Defined on probability space $(\boldsymbol{\Omega}, \mathcal{F}, P)$.

DGP is joint data density function $D_U(\cdot)$:

$$D_U\left(\mathbf{U}_1^T \mid \mathbf{U}_0, \boldsymbol{\psi}_1^T\right) \text{ with } \boldsymbol{\psi}_1^T \in \Psi \subseteq \mathcal{R}^k, \tag{4}$$

- $\boldsymbol{\psi}_1^T \in \Psi \subseteq \mathcal{R}^k$ is $k \times 1$ parameter in space Ψ;
- $\boldsymbol{\psi}_1^T$ must not depend on $\mathcal{F}$;
- $\mathbf{U}_0$ are initial conditions.
- $\mathbf{U}_1^T$ unmanageably large: must reduce.

From **DGP** through **LDGP** & **GUM** to selected model arises by sequence of **reductions** organized into twelve stages: see Hendry (1994) and Hendry (2009).

From DGP to LDGP:

1. Aggregation
2. Data transformations
3. Data partition
4. Marginalizing
5. Sequential factorization
6. Parameters of interest

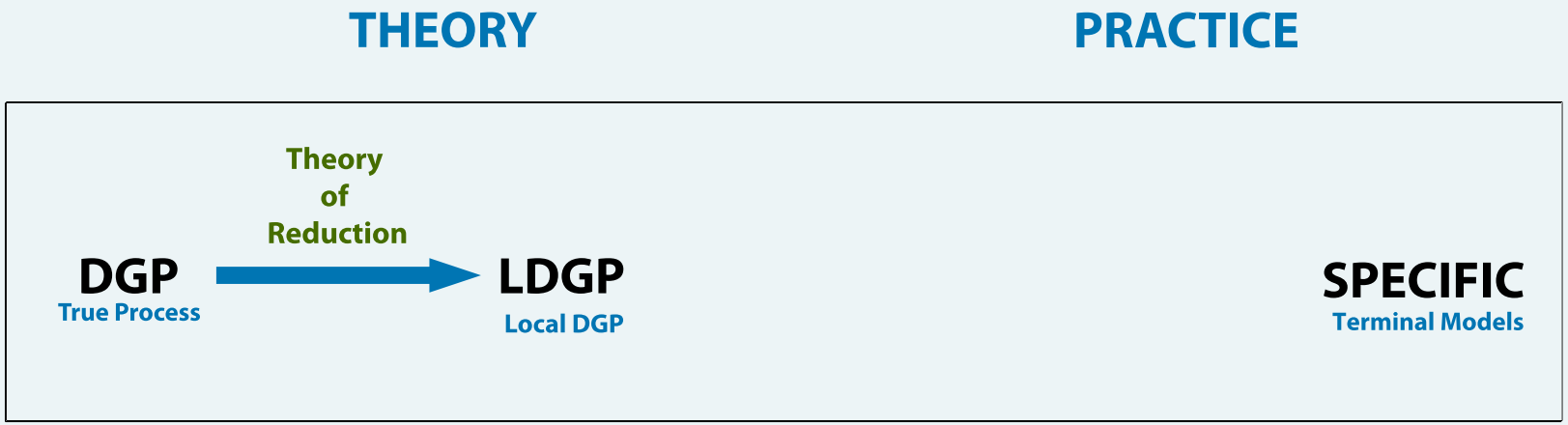
Approximating the LDGP:

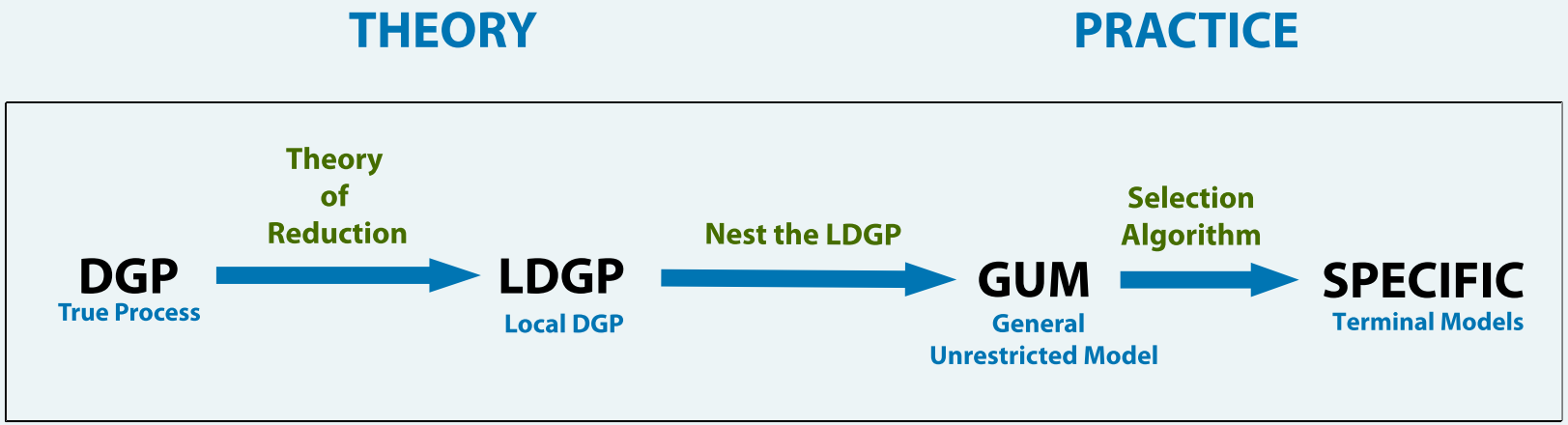
7. Lag truncation
8. Parameter constancy and invariance
9. Functional form (linearity)

Formulating a specific GUM:

10. Mapping to non-integrated data
11. Conditional factorization
12. Simultaneity

From DGP to LDGP




THEORY
PRACTICE
DGP
True Process
Theory
of
Reduction
LDGP
Local DGP
Nest the LDGP
GUM
General
Unrestricted Model
Selection
Algorithm
SPECIFIC
Terminal Models

1-1 mapping of $\mathbf{U}_1^T$ to new data set $\mathbf{W}_1^T$ where $\mathbf{U}_1^T \leftrightarrow \mathbf{W}_1^T$.

DGP of $\mathbf{U}_1^T$, and so of $\mathbf{W}_1^T$, characterized by joint density:

$$\mathsf{D}_U\left(\mathbf{U}_1^T \mid \mathbf{U}_0, \boldsymbol{\psi}_1^T\right) = \mathsf{D}_W\left(\mathbf{W}_1^T \mid \mathbf{W}_0, \boldsymbol{\phi}_1^T\right) \tag{5}$$

where $\boldsymbol{\psi}_1^T \in \boldsymbol{\Psi}$ and $\boldsymbol{\phi}_1^T \in \boldsymbol{\Phi}$.

Transformation from $\mathbf{U}$ to $\mathbf{W}$ affects parameter space
so $\boldsymbol{\Psi}$ transformed into $\boldsymbol{\Phi}$.
But $\mathbf{W}_1^T$ contain aggregates of interest.

Key to all reductions:

- impact on parameters – $\boldsymbol{\psi}_1^T$ versus $\boldsymbol{\phi}_1^T$.

Transform joint probabilities to a product:

$$\begin{aligned} \mathrm{P}(\mathrm{a}, \mathrm{b}) &= \mathrm{P}(\mathrm{a} \mid \mathrm{b})\,\mathrm{P}(\mathrm{b}) \\ &= \mathrm{P}(\mathrm{b} \mid \mathrm{a})\,\mathrm{P}(\mathrm{a}) \end{aligned}$$

Only use one ordering (often time); and use repeatedly!

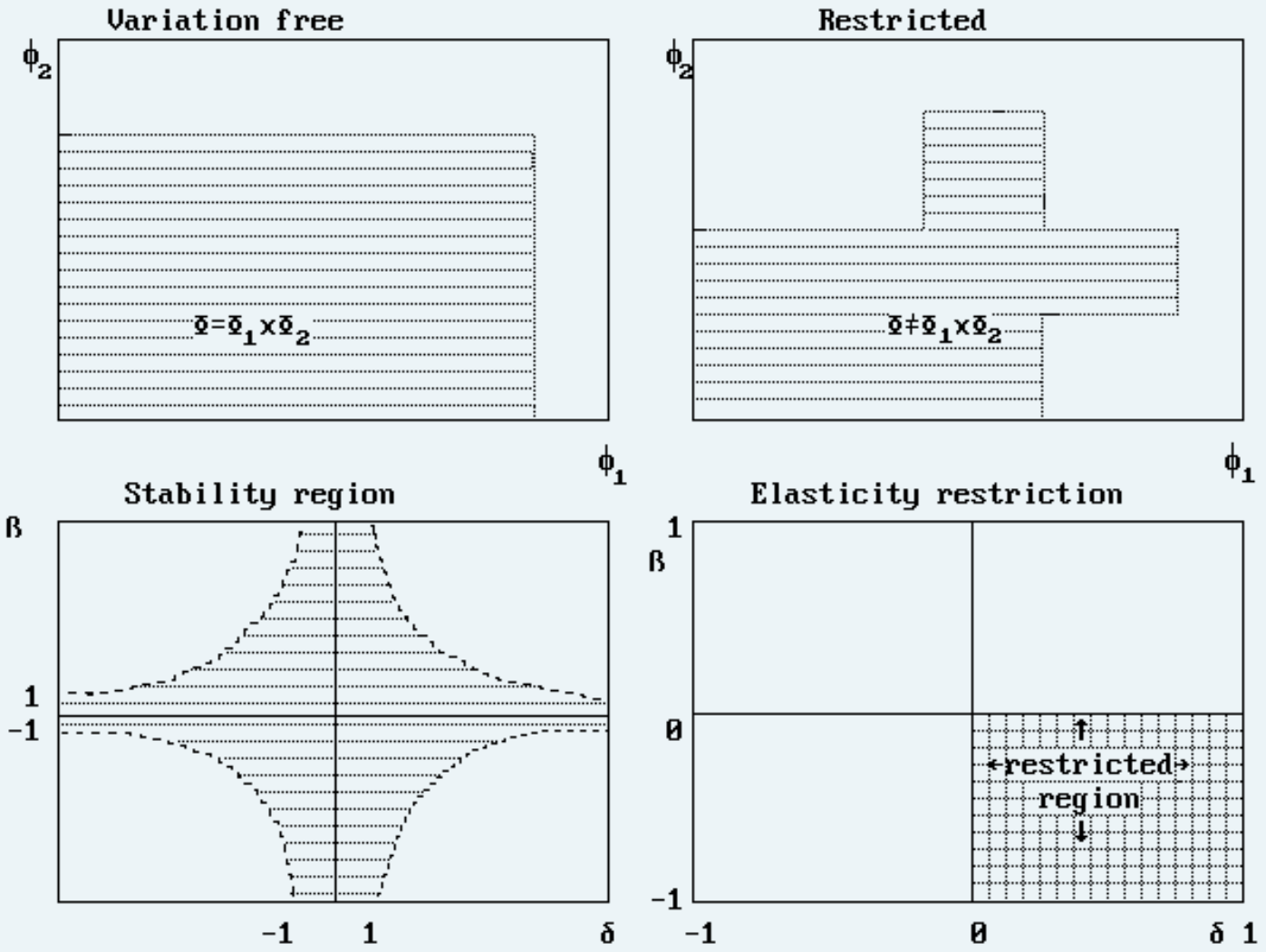
$$\begin{aligned} \mathrm{P}(\mathrm{a}, [\mathrm{b}, \mathrm{c}]) &= \mathrm{P}(\mathrm{a} \mid [\mathrm{b}, \mathrm{c}])\,\mathrm{P}(\mathrm{b}, \mathrm{c}) \\ &= \mathrm{P}(\mathrm{a} \mid [\mathrm{b}, \mathrm{c}])\,\mathrm{P}(\mathrm{b} \mid \mathrm{c})\,\mathrm{P}(\mathrm{c}) \end{aligned}$$

Allowing for parameters, we write:

$$\mathrm{P}(\mathrm{a}, \mathrm{b} \mid \boldsymbol{\psi}) = \mathrm{P}(\mathrm{a}, \mathrm{b} \mid \boldsymbol{\theta}) = \mathrm{P}(\mathrm{a} \mid \mathrm{b}, \boldsymbol{\theta}_1)\,\mathrm{P}(\mathrm{b} \mid \boldsymbol{\theta}_2)$$

where $\boldsymbol{\theta} = \mathbf{f}(\boldsymbol{\psi})$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta \subseteq \mathcal{R}^{\mathrm{k}}$.

Crucial issue: does $\Theta = \Theta_1 \times \Theta_2$?


Variation free
ϕ_2
$\Theta=\Theta_1 \times \Theta_2$
ϕ_1
Restricted
ϕ_2
$\Theta \neq \Theta_1 \times \Theta_2$
ϕ_1
Stability region
ß
1
-1
-1
1
δ
Elasticity restriction
1
ß
0
-1
restricted
region
-1
0
δ 1

Partition $\mathbf{W}_{\mathsf{T}}^{1}$ into the two sets:

$$\mathbf{W}_{\mathsf{T}}^{1} = \left(\mathbf{X}_{\mathsf{T}}^{1} : \mathbf{V}_{\mathsf{T}}^{1}\right) \tag{6}$$

where $\mathbf{X}_{\mathsf{T}}^{1}$ is $\mathsf{T} \times \mathrm{n}$.

Everything desired from the analysis must be learnt from analyzing $\mathbf{X}_{\mathsf{T}}^{1}$ **alone**:
$\mathbf{V}_{\mathsf{T}}^{1}$ must not be essential to inference.

Hidden condition that parameters of their distributions must be **variation free.**

$$\begin{aligned} \mathsf{D}_{\mathsf{W}}\left(\mathbf{W}_{\mathsf{T}}^{1} \mid \mathbf{W}_{0}, \boldsymbol{\phi}_{\mathsf{T}}^{1}\right) &= \mathsf{D}_{\mathsf{V}|\mathsf{X}}\left(\mathbf{V}_{\mathsf{T}}^{1} \mid \mathbf{X}_{\mathsf{T}}^{1}, \mathbf{W}_{0}, \boldsymbol{\varphi}_{\mathrm{a},\mathsf{T}}^{1}\right) \times \\ &\quad \mathsf{D}_{\mathsf{X}}\left(\mathbf{X}_{\mathsf{T}}^{1} \mid \mathbf{W}_{0}, \boldsymbol{\varphi}_{\mathrm{b},\mathsf{T}}^{1}\right) \end{aligned} \tag{7}$$

Eliminate $\mathbf{V}_{\mathsf{T}}^{1}$ by **discarding conditional density** [$\mathsf{D}_{\mathsf{V}|\mathsf{X}}\left(\mathbf{V}_{\mathsf{T}}^{1}|\mathbf{X}_{\mathsf{T}}^{1}, \mathbf{W}_{0}, \boldsymbol{\varphi}_{\mathrm{a},\mathsf{T}}^{1}\right)$ in (7)].

Retain marginal density $\mathsf{D}_{\mathsf{X}}\left(\mathbf{X}_{\mathsf{T}}^{1}|\mathbf{W}_{0}, \boldsymbol{\varphi}_{\mathrm{b},\mathsf{T}}^{1}\right)$.

Key is **no loss of relevant information**:
a **cut** is required, so that: $\left(\boldsymbol{\varphi}_{\mathrm{a},\mathsf{T}}^{1} : \boldsymbol{\varphi}_{\mathrm{b},\mathsf{T}}^{1}\right) \in \boldsymbol{\Phi}_{\mathrm{a}} \times \boldsymbol{\Phi}_{\mathrm{b}}$.

Thus, parameters must be **variation free**.

Innovation process created by sequentially factorizing $\mathbf{X}_1^T$:

$$\begin{aligned} \mathsf{D}_X\left(\mathbf{X}_1^T \mid \mathbf{W}_0, \boldsymbol{\varphi}_{b,T}^1\right) &= \mathsf{D}_x\left(\mathbf{x}_T | \mathbf{X}_{T-1}^1, \mathbf{W}_0, \boldsymbol{\varphi}_{b,T}\right) \times \\ &\quad \mathsf{D}_X\left(\mathbf{X}_{T-1}^1 | \mathbf{W}_0, \boldsymbol{\varphi}_{b,T-1}^1\right) \\ &\quad \vdots \\ \mathbf{LDGP} &= \prod_{t=1}^{T} \mathsf{D}_x\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^1, \mathbf{W}_0, \boldsymbol{\varphi}_{b,t}\right) \end{aligned} \tag{8}$$

Creates mean innovation error process: $\boldsymbol{\epsilon}_t = \mathbf{x}_t - \mathsf{E}\left[\mathbf{x}_t | \mathbf{X}_{t-1}^1\right]$.
Can treat sequential densities 'as if' independent:
allows laws of large numbers, central limit theorems etc.
as $\mathsf{E}\left[\boldsymbol{\epsilon}_t | \mathbf{X}_{t-1}^1\right] = \mathbf{0}$. (implicitly assume $\mathsf{E}_t[\cdot] = \mathsf{E}[\cdot]$).

Conditional, sequential distribution of $\{\mathbf{x}_t\}$ must not depend on $\mathbf{V}_{t-1}^1$ (Granger non-causality):

$$\mathsf{D}_x\left(\mathbf{x}_t \mid \mathbf{V}_{t-1}^1, \mathbf{X}_{t-1}^1, \mathbf{W}_0, \boldsymbol{\phi}_{b,t}\right) = \mathsf{D}_x\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^1, \mathbf{W}_0, \boldsymbol{\phi}_{b,t}^*\right) \quad (9)$$

If so:

$$\prod_{t=1}^{T} \mathsf{D}_x\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^1, \mathbf{W}_0, \boldsymbol{\phi}_{b,t}^*\right) \quad (10)$$

from (9) is **LDGP** for $\{\mathbf{x}_t\}$ with $\boldsymbol{\phi}_{b,t}^* = \boldsymbol{\varphi}_{b,t}$
so (9) coincides with last line of (8) above.

LDGP well defined as a reduction of DGP, but may be high-dimensional, complicated, non-linear, & evolving.
Knowledge of LDGP is best that can be achieved in space of $\{\mathbf{x}_t\}$–target for model selection.

Denoted $\boldsymbol{\mu} \in \mathcal{M}$.
Should be:

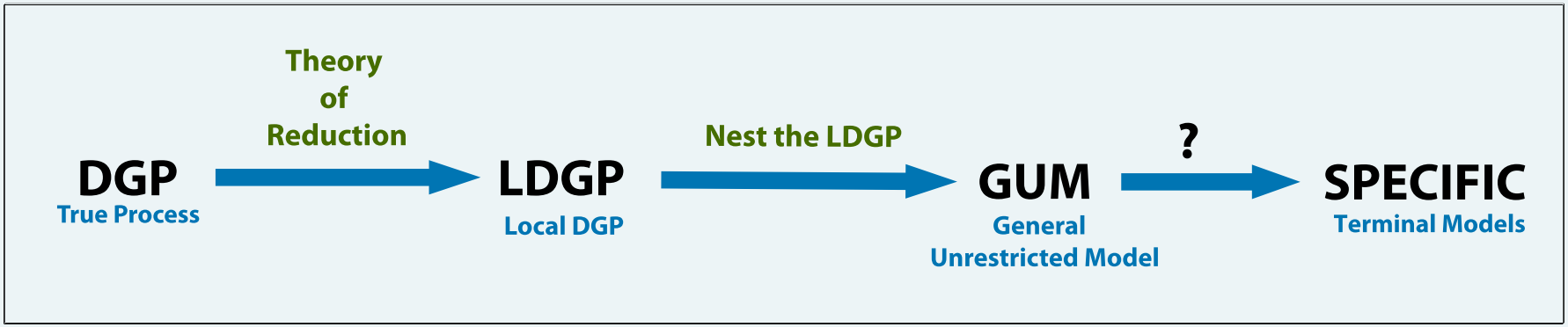
- **identifiable**,
- **constant**, and
- **invariant** to relevant class of interventions.

Requires that $\boldsymbol{\mu} = \mathbf{f}(\boldsymbol{\phi}_{\mathsf{T}}^{1})$ from (5).

Guides choice of:

- data to analyze: $\mathbf{X}_{\mathsf{T}}^{1}$;
- data transformations: $\mathbf{h}(\mathbf{x}_{\mathsf{t}})$;
- form of model: $\mathbf{g}[\mathbf{h}(\mathbf{X}_{\mathsf{T}}^{1})] = \boldsymbol{\epsilon}_{\mathsf{t}}$.

THEORY **PRACTICE**



Fix extent of history of $\mathbf{X}_{t-1}^{1}$ in (8) at s earlier periods:

$$\mathsf{D}_{\mathsf{x}}\left(\mathbf{x}_{t} \mid \mathbf{X}_{t-1}^{1}, \mathbf{W}_{0}, \boldsymbol{\varphi}_{b,t}\right) = \mathsf{D}_{\mathsf{x}}\left(\mathbf{x}_{t} \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{W}_{0}, \boldsymbol{\varphi}_{b,t}\right) \tag{11}$$

Obvious checks on validity of such a reduction are whether longer lags matter, or error remains an innovation:
key criterion is impact on $\{\boldsymbol{\varphi}_{b,t}\}$.

The parameters in question are those that characterize the distribution in (11).

$\{\varphi_{b,t}\}$ may have different elements for different times:

$$\{\varphi_{b,t}\} = (\varphi_{b,1}, \varphi_{b,2} \cdots \varphi_{b,T-1}, \varphi_{b,T}) \tag{12}$$

Constancy entails that $\{\varphi_{b,t}\}$ depends on a smaller set of parameters that are constant, at least within regimes.
Complete parameter constancy is:

$$\varphi_{b,t} = \boldsymbol{\lambda}_0 \ \in \boldsymbol{\Lambda}_0 \ \forall t. \tag{13}$$

E.g.: $y_t = \rho_t z_t + \epsilon_t$ where $\rho_t = \rho_{t-1} + \eta_t \implies \rho_t = \rho_0 + \sum_{i=1}^{t} \eta_i$

Most misunderstood, and one of least analyzed, concepts

- not improved by '**random parameters**';
- not the same as **invariance**;
- and non-constancy not entailed by **forecast failure**.

Invariance is under extensions of the information set:

1. **over time**: if extend data series, parameters stay the same;
2. **across regimes**: if change input variable, relation to output variable is unchanged and;
3. **new sources**: if add additional variables to the analysis, parameters are unaltered.

All three are necessary conditions, and are easily **testable**:

1. test for parameter constancy at end of sample;
2. change a regressor and test parameter constancy;
3. add potential candidate variables and test they are irrelevant.

Map $\mathbf{x_t}$ into $\mathbf{x_t^* = h\left(x_t\right)}$.

Denote resulting data by $\mathbf{X^*}$.

Assume that $\mathbf{x_t^*}$ makes $\mathsf{D}_{x^*}\left(\cdot\right)$ approximately normal and homoscedastic, $\mathsf{N}_n\left[\boldsymbol{\eta_t}, \boldsymbol{\Upsilon}\right]$, with no loss of information:

$$\mathsf{D}_x\left(\mathbf{x_t} \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{W}_0, \boldsymbol{\lambda}_0\right) = \mathsf{D}_{x^*}\left(\mathbf{x_t^*} \mid \left(\mathbf{X^*}\right)_{t-1}^{t-s}, \mathbf{W}_0, \boldsymbol{\lambda}\right) \tag{14}$$

Intimate connection between functional form and parameter constancy.

Drop * notation for simplicity, so $\mathbf{x_t}$ denotes after any relevant transformations.

For implementation, must specify precise functional form: e.g., linear model for $D_x\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{W}_0, \boldsymbol{\lambda}\right)$:

$$\mathbf{x}_t = \sum_{j=1}^{s} \mathbf{A}_j \mathbf{x}_{t-j} + \boldsymbol{\epsilon}_t \tag{15}$$

Express more generally by $\boldsymbol{\Gamma}(L)\,\mathbf{x}_t = \boldsymbol{\epsilon}_t$ where:

$$\boldsymbol{\Gamma}(L) = \sum_{j=0}^{s} \boldsymbol{\Gamma}_j L^j;$$

$\boldsymbol{\Gamma}(L)$ is a polynomial matrix of order s in lag operator L where $\boldsymbol{\Gamma}_0 = \mathbf{I}_n$.

Underpins VAR models as the GUM.

Most economic time series are non-stationary.
Key aspect is unit root (integrated data): denoted I(1).
Reduction to I(0) ensures conventional inferences valid
but many inferences valid even if I(0) reduction not enforced.

Importantly holds for most diagnostic tests
(not heteroscedasticity however).

Mapping to I(0) also helps interpret outcomes
and reduces dimensionality of parameterization.
Two possible reductions: **differencing and cointegration**:

- $\Delta \mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$ always removes unit root;
- $\boldsymbol{\beta}' \mathbf{x}_t$ sometimes does so as well.

Then both are I(0), and conventional inference applies.

Factorize $\mathbf{x}_t$ into sets of n_1 and n_2 variables, $n_1 + n_2 = n$:

$$\mathbf{x}_t' = \left(\mathbf{y}_t' : \mathbf{z}_t'\right), \tag{16}$$

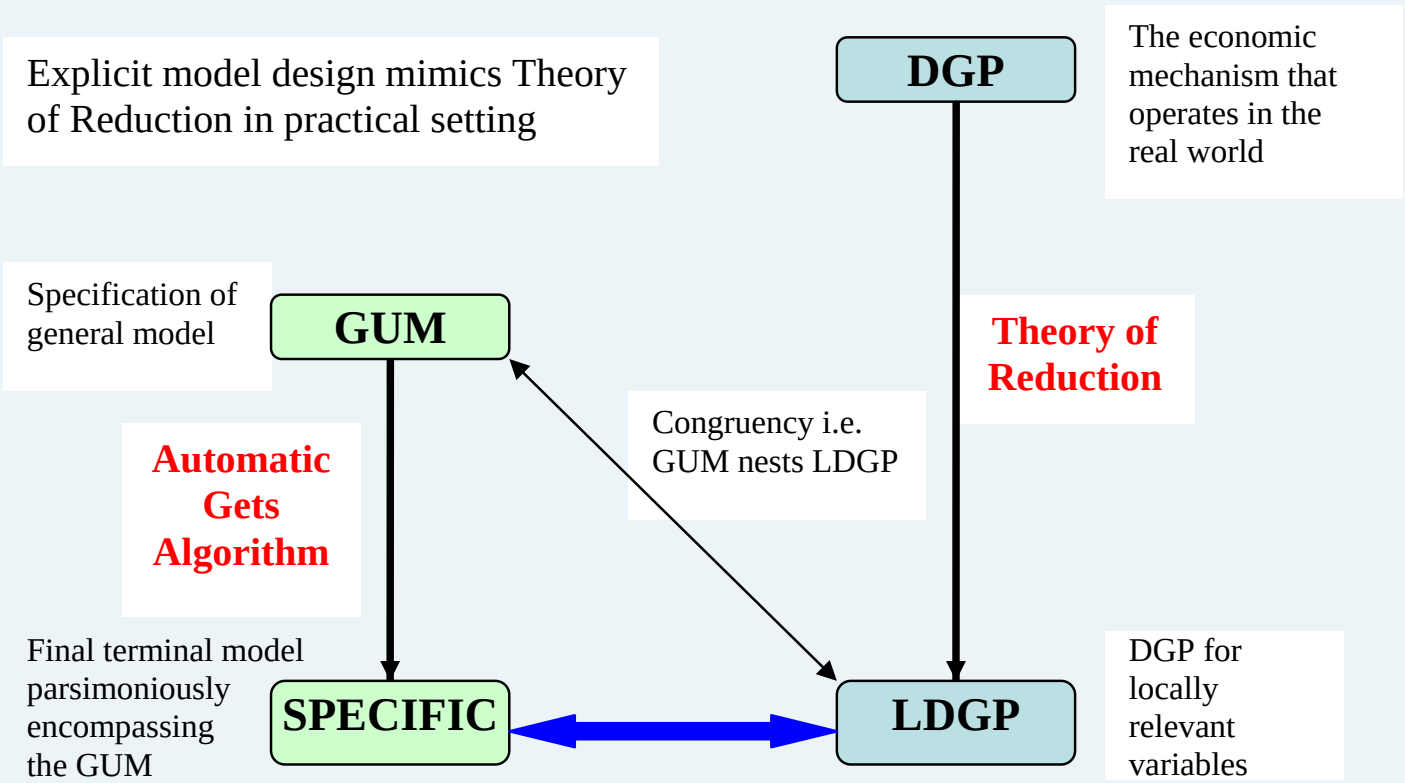
where $\mathbf{y}_t$ endogenous and $\mathbf{z}_t$ non-modelled.

$$\begin{aligned} \mathsf{D}_x\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{W}_0, \boldsymbol{\lambda}\right) &= \mathsf{D}_{y|z}\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{X}_{t-1}^{t-s}, \mathbf{W}_0, \boldsymbol{\theta}_a\right) \times \\ &\quad \mathsf{D}_z\left(\mathbf{z}_t \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{W}_0, \boldsymbol{\theta}_b\right) \end{aligned} \tag{17}$$

$\mathbf{z}_t$ **weakly exogenous** for $\boldsymbol{\mu}$ if:

1. $\boldsymbol{\mu} = \mathbf{f}\left(\boldsymbol{\theta}_a\right)$ **alone**; and
2. $\left(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b\right) \in \boldsymbol{\Theta}_a \times \boldsymbol{\Theta}_b$.

Justifies contemporaneous conditioning.


Explicit model design mimics Theory of Reduction in practical setting
DGP
The economic mechanism that operates in the real world
Specification of general model
GUM
Theory of Reduction
Automatic Gets Algorithm
Congruency i.e. GUM nests LDGP
Final terminal model parsimoniously encompassing the GUM
SPECIFIC
LDGP
DGP for locally relevant variables

Consequences of **failure of weak exogeneity** vary from:
– loss of estimation efficiency,
– through to a loss of parameter constancy.

Experimental setting where Gauss–Markov conditions seem satisfied:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim \mathsf{N}_{\mathsf{T}}\left[\mathbf{0}, \sigma_{\epsilon}^{2}\mathbf{I}\right] \tag{18}$$

when $\mathbf{Z}' = (\mathbf{z}_1 \ldots \mathbf{z}_{\mathsf{T}})$ is a $\mathsf{T} \times \mathrm{k}$ matrix, $\mathsf{rank}(\mathbf{Z}) = \mathrm{k}$, and $\boldsymbol{\epsilon}' = (\epsilon_1 \ldots \epsilon_{\mathsf{T}})$ with:

$$\mathsf{E}[\mathbf{y} \mid \mathbf{Z}] = \mathbf{Z}\boldsymbol{\beta}$$

hence $\mathsf{E}[\mathbf{Z}'\boldsymbol{\epsilon}] = \mathbf{0}$. But:

OLS need not be most efficient unbiased estimator of $\boldsymbol{\beta}$.

Explicit weak exogeneity condition required on $\mathbf{Z}$:
must preclude $\boldsymbol{\beta}$ being learned from marginal distribution.
Suppose that marginal equals:

$$\mathbf{z}_t = \boldsymbol{\beta} + \boldsymbol{\nu}_t \text{ with } \boldsymbol{\nu}_t \sim \mathsf{N}_k\left[\mathbf{0}, \boldsymbol{\Omega}_\nu\right] \tag{19}$$

allows sample mean $\bar{\mathbf{z}}$ of $\mathbf{z}$ to be an unbiased estimator:

$$\mathsf{E}[\bar{\mathbf{z}}] = \boldsymbol{\beta}, \tag{20}$$

$$\mathsf{V}[\bar{\mathbf{z}}] = \mathsf{T}^{-1}\boldsymbol{\Omega}_\nu, \tag{21}$$

so $\bar{\mathbf{z}}$ can dominate $\widehat{\boldsymbol{\beta}}$, possibly dramatically if $\boldsymbol{\Omega}_\nu$ is tiny compared to $\sigma_\epsilon^2(\mathbf{Z}'\mathbf{Z})^{-1}$.

Cointegrated systems allow testing of one aspect of weak exogeneity:

- equilibrium-correction mechanisms which cross-link equations violate long-run weak exogeneity;also shows weak exogeneity cannot necessarily be obtained merely by choosing 'parameters of interest'.
- The presence of a disequilibrium term in more than one equation is testable.

Structural breaks allow tests for **super exogeneity** and the **Lucas (1976) critique**: see e.g. Engle and Hendry (1993).

When conditional models are constant despite data moments changing, there is evidence of super exogeneity for that model's parameters: automatic test in Hendry and Santos (2010).

Sequentially factorize DGP of n-dimensional $\{\mathbf{x}_t\}$:

$$\prod_{t=1}^{T} \mathsf{D}_{\mathsf{x}}\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}, \boldsymbol{\theta}\right) = \prod_{t=1}^{T} \mathsf{D}_{\mathsf{y}|\mathsf{z}}\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{X}_{t-1}, \boldsymbol{\phi}_1\right) \mathsf{D}_{\mathsf{z}}\left(\mathbf{z}_t \mid \mathbf{X}_{t-1}, \boldsymbol{\phi}_2\right) \tag{22}$$

where $\mathbf{x}_t = \left(\mathbf{y}_t' : \mathbf{z}_t'\right)$ and $\boldsymbol{\phi} = \left(\boldsymbol{\phi}_1' : \boldsymbol{\phi}_2'\right)' = \mathbf{f}\left(\boldsymbol{\theta}\right) \in \mathcal{R}^{k}$.

If parameters of $\mathbf{y}$ and $\mathbf{z}$ processes variation free –
$\mathbf{z}_t$ weakly exogenous for parameters of interest $\boldsymbol{\psi} = \mathbf{h}\left(\boldsymbol{\phi}_1\right)$:
does not rule out that $\boldsymbol{\phi}_1$ may change if $\boldsymbol{\phi}_2$ is changed.

Super exogeneity adds parameter invariance in conditional:

$$\frac{\partial \boldsymbol{\phi}_1}{\partial \boldsymbol{\phi}_2'} = \mathbf{0}.$$

If $D_x(\cdot)$ is multivariate normal:

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} \sim IN_n \left[\begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \end{pmatrix}, \begin{pmatrix} \sigma_{11,t} & \sigma'_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{pmatrix} \right] \tag{23}$$

$\mu_{1,t}$ and $\mu_{2,t}$ are functions of $\mathbf{X_{t-1}}$.
Suppose economic theory suggests that $\mu_{1,t} = \beta \mu_{2,t}$
β is primary parameter of interest.

$$\text{Cond. Model: } E[y_t|z_t] = \mu_{1,t} + \sigma_{12,t}\sigma_{22,t}^{-1}(z_t - \mu_{2,t})$$

$$\text{Cond. Var.: } V[y_t|z_t] = \sigma_{11,t} - \sigma'_{12,t}\sigma_{22,t}^{-1}\sigma_{12,t} = \omega_t^2$$

re-write cond. model as:

$$E[y_t|z_t] = \beta \mu_{2,t} + \sigma_{12,t}\sigma_{22,t}^{-1} z_t - \sigma_{12,t}\sigma_{22,t}^{-1}\mu_{2,t}$$

when is z_t super exogenous for parameters (β, ω_t^2)?

Further re-write cond. model:

$$\begin{aligned} E[y_t|z_t] &= \beta\mu_{2,t} + \sigma_{12,t}\sigma_{22,t}^{-1}z_t - \sigma_{12,t}\sigma_{22,t}^{-1}\mu_{2,t} \\ &= (\beta - \sigma_{12,t}\sigma_{22,t}^{-1})\mu_{2,t} + \sigma_{12,t}\sigma_{22,t}^{-1}z_t \\ &= \gamma_{1,t} + \gamma_{2,t}z_t \end{aligned}$$

Econometric model:

$$y_t = \beta_0 + \beta z_t + \epsilon_t$$

For z_t to be weakly exogenous:

- $\beta = \sigma_{12}\sigma_{22}^{-1}\forall t$ then $\mu_{2,t}$ does not enter conditional model.

For z_t to be super exogenous additionally need:

- $\gamma_{2,t} = \gamma_2 \forall t$
- $\omega_t^2 = \omega^2 \forall t$
- $(\gamma_1, \gamma_2, \omega^2)$ invariant to changes in marginal $(\mu_{2,t}, \sigma_{22,t})$.

Key implication: $\gamma_{1,t}$ **depends on** $\mu_{2,t}$ **when** $\beta \neq \sigma_{12}\sigma_{22}^{-1} = \gamma_{2,t}$.

Then changes in marginal reflected in conditional.

requires $\beta = \gamma_2$, which is testable if $\mu_{2,t}$ shifts.
If γ_2 unaffected by shifts in $\mu_{2,t}$, then z_t super exogenous.

Note: **Automatic test** using indicators from marginal tested for significance in conditional.

Three attributes:

- ‘uniqueness’;
- ‘corresponds to desired entity’;
- ‘satisfies the assumed interpretation’.

Consider regression of quantity on price
and want to ‘identify as demand equation’:

1. unique function of data second moments;
2. but need not correspond to underlying demand behaviour;
3. and may be incorrectly interpreted–really supply schedule.

Uniqueness is often the sense intended in econometrics.

Simultaneous representation is a reduction.

$$\mathbf{B}\mathbf{y}_t = \mathbf{C}\mathbf{z}_t + \boldsymbol{\epsilon}_t \text{ where } \boldsymbol{\epsilon}_t \sim \mathsf{IN}_{n_1}[\mathbf{0}, \boldsymbol{\Sigma}_\epsilon] \tag{24}$$

so actual DGP is:

$$\mathbf{y}_t = \boldsymbol{\Pi}\mathbf{z}_t + \mathbf{w}_t \text{ where } \mathbf{w}_t \sim \mathsf{IN}_{n_1}[\mathbf{0}, \boldsymbol{\Omega}_w] \tag{25}$$

with:

$$\mathbf{B}\boldsymbol{\Pi} - \mathbf{C} = \mathbf{0} \tag{26}$$

(26) may entail restrictions on $\boldsymbol{\Pi}$, so is a testable reduction.

$$\text{GUM: } \mathbf{A}(\mathbf{L})\,\mathbf{h}(\mathbf{y})_t = \mathbf{B}(\mathbf{L})\,\mathbf{g}(\mathbf{z})_t + \boldsymbol{\epsilon}_t \tag{27}$$

where $\boldsymbol{\epsilon}_t \underset{app}{\sim} \mathsf{N}_{n_1}[\mathbf{0}, \boldsymbol{\Sigma}_\epsilon]$.

Notice the many design steps in reaching (27).

$\boldsymbol{\epsilon}_t$ is a **derived** and not an autonomous process:

$$\boldsymbol{\epsilon}_t = \mathbf{A}(\mathrm{L})\,\mathbf{h}(\mathbf{y})_t - \mathbf{B}(\mathrm{L})\,\mathbf{g}(\mathbf{z})_t \tag{28}$$

Designed by choices made in reduction.

Crucial insight: **model changes as reductions are altered**.

Parameters of model are **functions** of $\boldsymbol{\phi}^1_{\mathsf{T}}$ in (5):
$(\mathbf{A}(\cdot)\ \mathbf{B}(\cdot)) = \mathbf{f}(\boldsymbol{\phi}^1_{\mathsf{T}})$ alter as $\boldsymbol{\phi}^1_{\mathsf{T}}$ changes.

Test empirical model against LDGP, as represented by data properties.

Measures of **no information loss** if correct

1. **Aggregation**: no loss of information on marginalizing wrt disaggregates if retain **sufficient statistics** for $\boldsymbol{\mu}$;
2. **Transformations**: no associated reduction, but introduce $\boldsymbol{\mu}$, and need for parameters to be **invariant** and **identifiable**;
3. **Data partition**: which variables to include/omit – **fundamental** to success of empirical modelling;
4. **Marginalizing** wrt $\mathbf{v}_t$ without loss, if $\mathbf{X}_1^T$ sufficient for $\boldsymbol{\mu}$
   marginalizing wrt $\mathbf{V}_{t-1}^1$ without loss,
   if **Granger non-causality** for $\mathbf{x}_t$ and a **cut**;
5. **Sequential factorization**:
   no loss if $\boldsymbol{\epsilon}_t$ an innovation relative to $\mathbf{X}_{t-1}^1$

6. **LDGP**: reduction of DGP for relevant variables, nested within it; properties explained by reduction;
7. **Lag truncation**: no loss if ϵ_t remains an innovation;
8. **Parameter constancy and invariance**:
   constancy across interventions on marginal process;
9. **Functional form**: no reduction when two densities in (14) are equal
10. **Integrated data**:
    reduce to I(0) by **cointegration** and **differencing**
    Conventional inference and more parsimonious;
11. **Conditional factorization**: eliminate marginal process
    No loss of information if $\mathbf{z}_t$ **weakly exogenous** for μ;
12. **Simultaneity**: parsimoniously capture **joint dependence**.

13. **Knowledge of DGP entails knowledge of all reductions thereof**
    If one model entails knowledge of others, then that model is said to **encompass** them

Unlike symptomatology approach testing for:
autocorrelation, heteroscedasticity, omitted variables, non-constancy etc. then 'correcting' them.

Invalid approach as no unique alternative to any null;
successive outcomes can contradict;

no obvious termination point–
stopping after first 'non-rejection' is disastrous, and
no account of selection process.

Commence from ‘good’ approximation to LDGP,
which embeds available economic theory,
and institutional knowledge,
checked for congruence by mis-specification tests.

Mimic reduction theory in practical research, to minimize losses due to reductions imposed

Based on notions of **congruence** and **encompassing**:
– former close to ‘well specified’,
– latter entails explaining the results of other models.

We will address how to select model of LDGP from initial general unrestricted specification

Partition own data $\mathbf{X}_1^T$ into three information sets:

1. **past data**;
2. **present data**;
3. **future data**;

$$\mathbf{X}_1^T = \left(\mathbf{X}_{1}^{t-1} : \mathbf{x}_t : \mathbf{X}_{t+1}^{T}\right) \tag{29}$$

4. **Theory information**: source of μ, and creative stimulus;
5. **Measurement information**: price index theory, identities, data accuracy; and:
6. **Rival models** (encompass congruent models)

Leads to six main model evaluation criteria.

1. homoscedastic innovation ϵ_t;
2. weakly exogenous $\mathbf{z}_t$ for μ;
3. constant, invariant μ;
4. theory consistent, identifiable structures;
5. data-admissible formulations on accurate observations;
6. encompass rival models.

Exhaustive nulls to test; but many alternatives.

Models which satisfy [1] & [2] are **well specified** on available information.
Models which satisfy [1] & [2] & [3] are (empirically) **congruent**.

Admissible, theory-consistent, encompassing, congruent model satisfies all six criteria.

- Define a starting model: general unrestricted model (GUM)
  - Designed to be congruent (diagnostic testing) and relevant,
  - Tests of reductions with approximately correct distribution,
  - Reduction can maintain congruence (or lack thereof),
  - Reduction up to a predefined significance level (backtesting w.r.t. GUM: acceptable information loss).

- Define a starting model: general unrestricted model (GUM)
  - Designed to be congruent (diagnostic testing) and relevant,
  - Tests of reductions with approximately correct distribution,
  - Reduction can maintain congruence (or lack thereof),
  - Reduction up to a predefined significance level (backtesting w.r.t. GUM: acceptable information loss).

Model selection is an iterative search procedure, need to follow several paths:

- multiple path search, or
- tree search.

- t-tests (single variable removal).
- F-tests (tests of variables removed from the GUM, encompassing aka backtesting).
- F-tests (pruning to faster search).
- diagnostic tests
  - ARCH (Engle 1982)
  - Serial correlation (Godfrey 1978, Harvey 1981)
  - Heteroscedasticity (White 1980)
  - Normality (Jarque and Bera 1980; Doornik and Hansen 1994, 2008)
  - Chow (Chow 1960 in-sample stability test)*
- information criterion (tiebreaker)
- stability tests (out of sample, optionally)

- Model selection is an iterative search procedure
  - manual search can follow a few paths: slow and tedious,
  - computer automated search can follow all paths,
    Well, not all. There are 2^k models, so need a strategy.
    $k = 100$ at 10^9/sec: $10^6\times$ age of universe.

- Model selection is an iterative search procedure
  - manual search can follow a few paths: slow and tedious,
  - computer automated search can follow all paths,
    Well, not all. There are 2^k models, so need a strategy.
    $k = 100$ at 10^9/sec: $10^6\times$ age of universe.

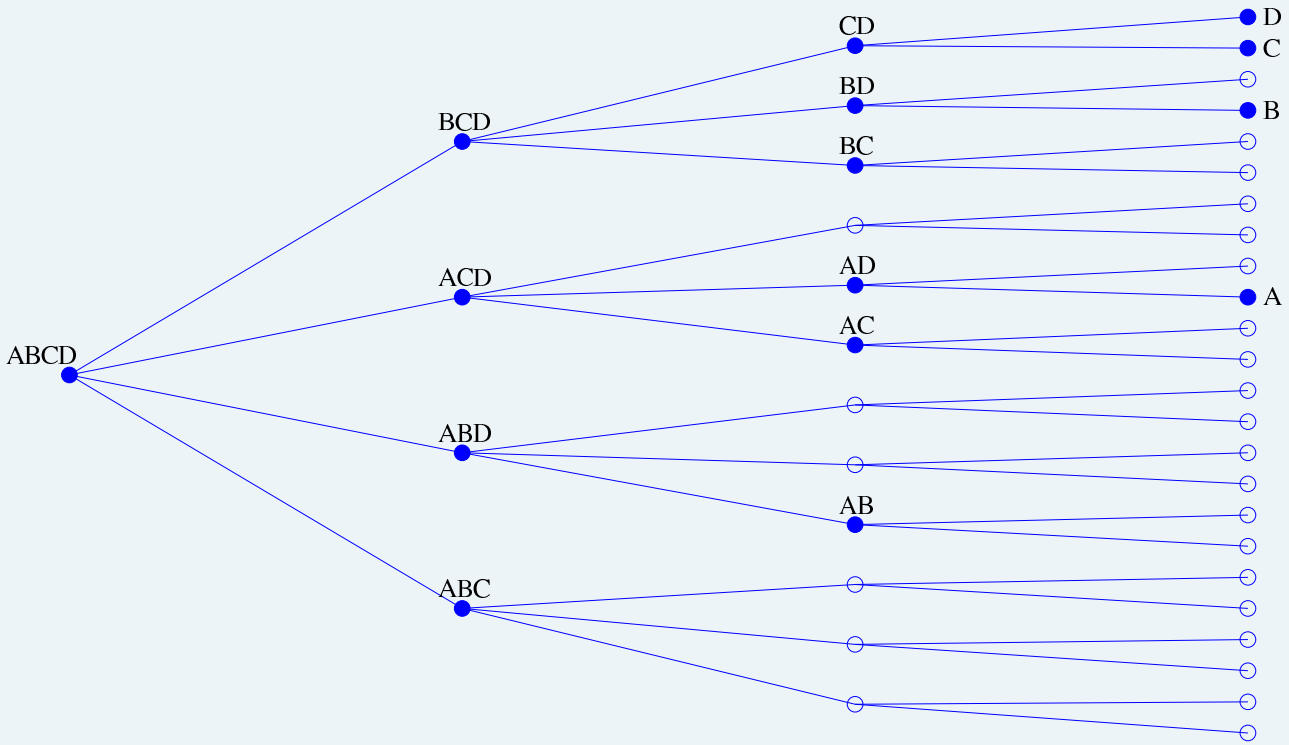
- General-to-specific model selection (Gets, 'Hendry' or 'LSE' methodology) largely driven by David Hendry (DHSY, PcGive, Alchemy, Dynamic Econometrics, ...)
  Lively debate.
- Automated Gets initiated by Hoover and Perez (1999), Hendry and Krolzig (2005) (PcGets: 2nd generation, theoretical properties, bias correction).
  Study model selection through simulation – improves debate.
- Autometrics & Getsm improve on PcGets, extended beyond standard regression models.

- Hoover-Perez (1999):
  1. General unrestricted model
  2. Multiple path search
  3. Encompassing test
  4. Diagnostic testing
  5. Tiebreaker
- Hendry and Krolzig (1999), PcGets (2001):
  1. Add presearch
  2. Extend multiple-path search
  3. Add iteration
  4. No out-of-sample testing (Lunch and Vital-Ahuja, 1998)
  5. Change treatment for Invalid GUM
- Autometrics (2009), Getsm (2017):
  1. Reduce role of presearch
  2. Change search path algorithm: tree search
  3. Extend scope: separation of model and algorithm
  4. Increase efficiency

Autometrics & Getsm implement underlying principle of general-to-specific model selection (*'Hendry methodology'*).

Autometrics & Getsm

- likelihood-based
- searches the whole model space:
  - tree search ensures that no model is estimated twice
  - irrelevant paths can be cut-off efficiently
  - F-tests are used to speed-up search
- implements backtracking on diagnostics: only test from terminal candidates, then backtrack if necessary
- backtesting w.r.t. GUM 0 (the initial GUM after presearches) removes need for encompassing of candidate models
- relevant terminal candidates remembered in iterated search
- implements block search for $N \geqslant T$



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

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.

DGP –

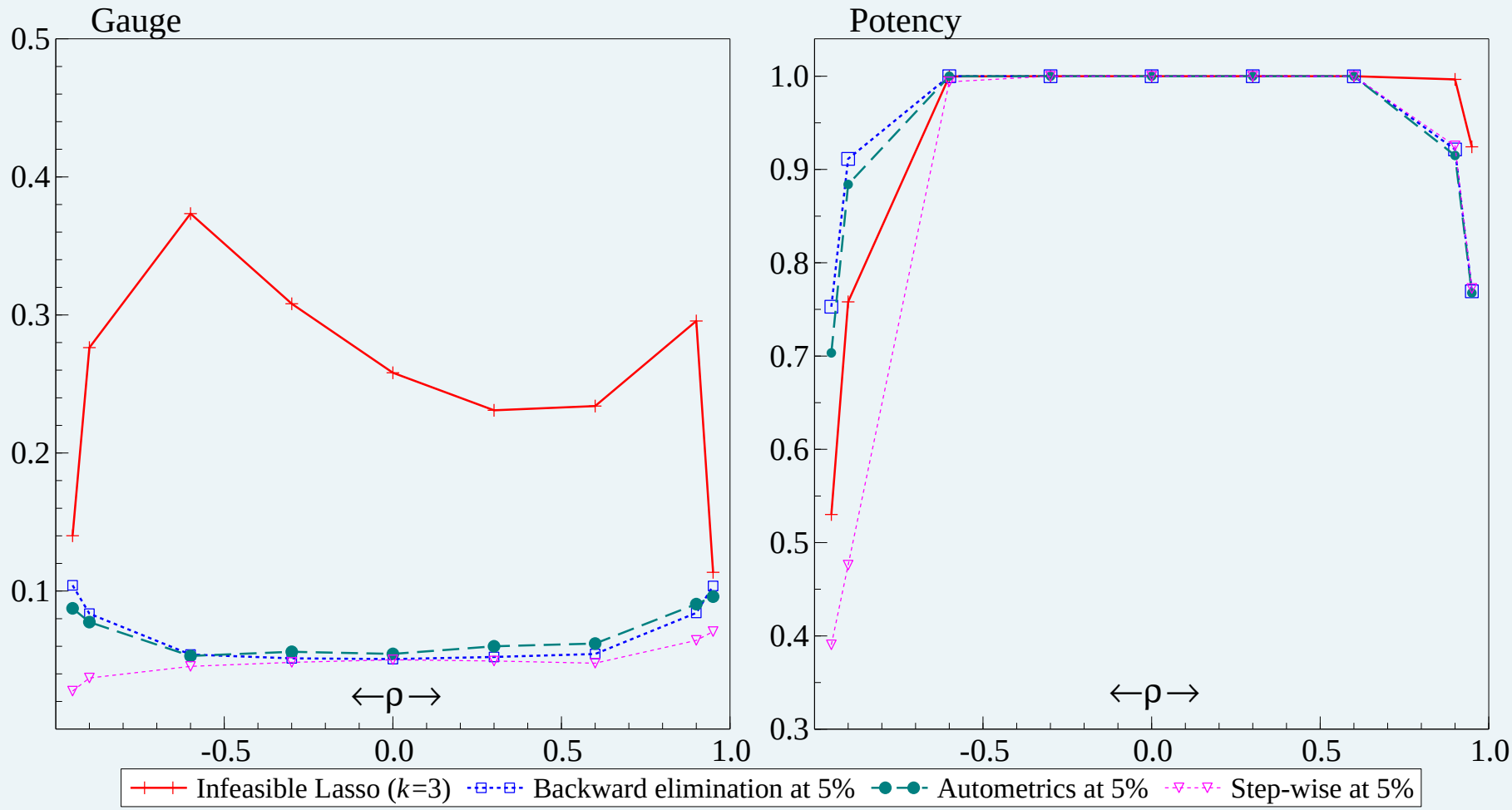
$$y_t = 0.5y_{t-1} + 0.8x_{1,t} + 0.8x_{2,t} + u_t, \quad u_t \sim \mathsf{IN}[0, 1],$$

$$\mathbf{x}_t = \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathsf{IN}\left[\mathbf{0}, \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}\right],$$

Model –

$$\begin{aligned} \mathcal{X}_{\text{fixed}} &= \{1\} \\ \mathcal{X}_{\text{free}} &= \{y_{t-1}, x_{1t}, x_{2t}, x_{3t}, x_{4t}\} \end{aligned}$$

Four selection methods



gauge: fraction of irrelevant variables (x_{3t}, x_{4t}) in the final model

potency: fraction of relevant variables (y_{t-1}, x_{1t}, x_{2t}) in the final model.

Diagnostic checking: subject every estimated model to a battery of diagnostic tests (normality, residual correlation, residual ARCH, in-sample Chow, out-of-sample Chow,...). If a test fails the reduction is rejected and the next model in line is considered.

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

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \text{ for } \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2] \tag{30}$$

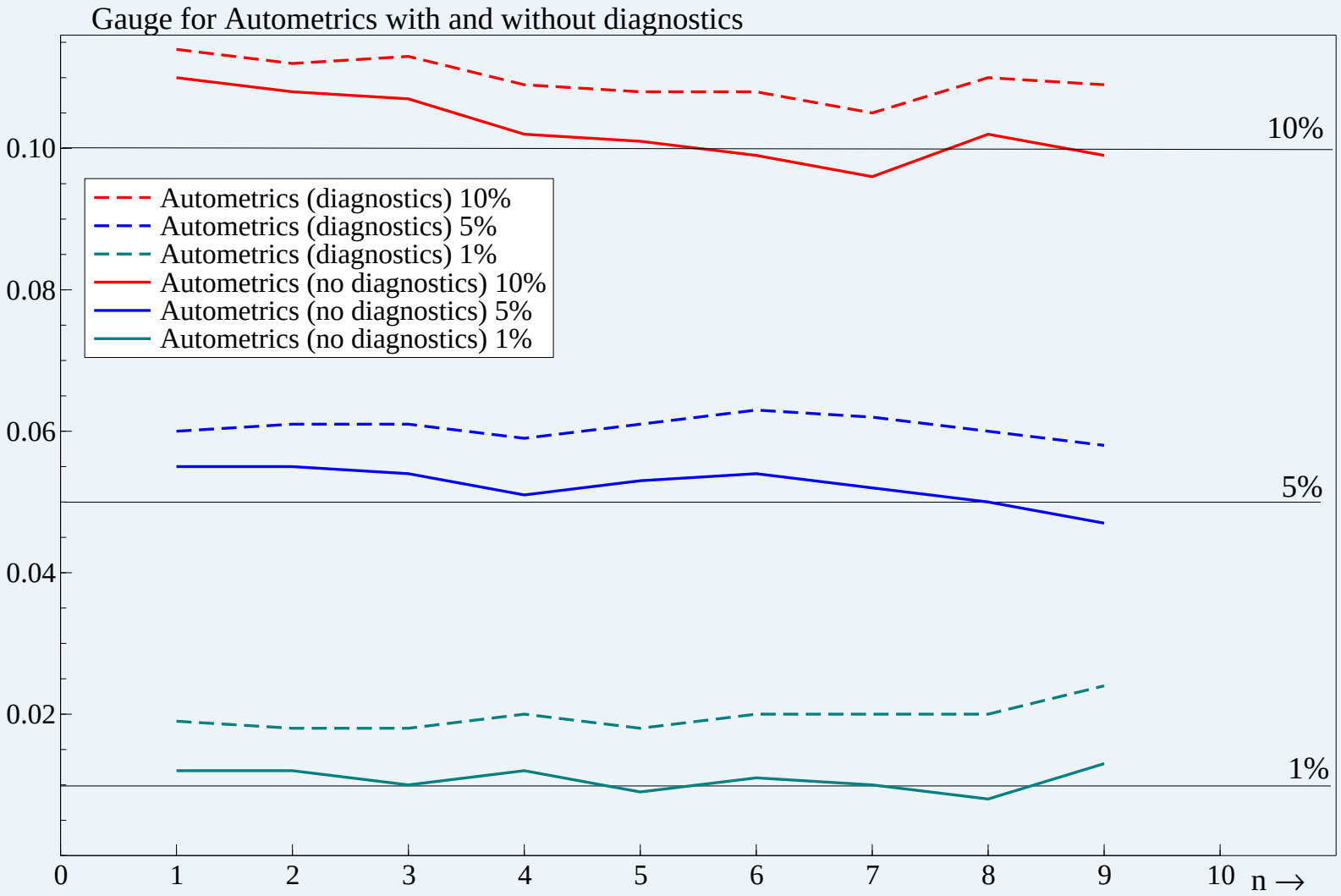
$T = 100$, $n = 1, \ldots, 10 = N$; $\beta_k = 0$ for $k > n$; $R^2 = 0.9$.

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

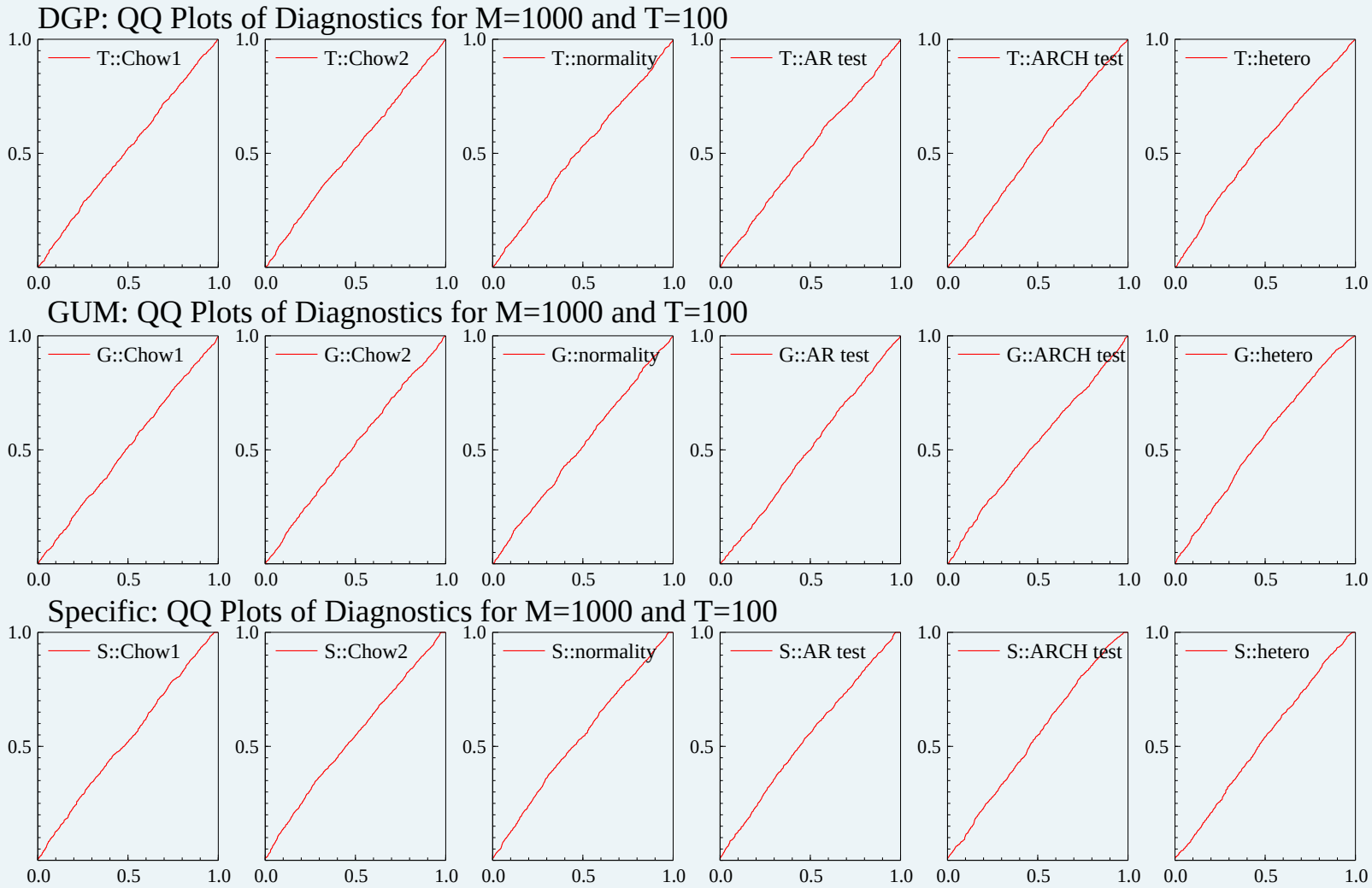
Gauge is larger than α with diagnostics **on**, when checking to ensure a congruent reduction.

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

Gauges with diagnostic tests off & on



Selection effects on test distributions



Motivated by Indicator Saturation Block-Partitioning.

More Variables Than Observations: Block Partitioning

Split variables into blocks $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$

$\mathcal{A} \cup \mathcal{B}$ select $\rightarrow \mathcal{G}_1$

$\mathcal{A} \cup \mathcal{C}$ select $\rightarrow \mathcal{G}_2$

$\mathcal{B} \cup \mathcal{C}$ select $\rightarrow \mathcal{G}_3$

$\Rightarrow \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$ select $\rightarrow$ Final Model

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.

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

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

1) The GUM has 3 DGP variables plus 37 irrelevant.

2) Then consider 141 irrelevant, larger than $T = 139$.

$N < T$: $T = 139$, **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)

$N > T$: $T = 139$, **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**

Have now established that there is little loss from using the path-search algorithms

- Gauge is close to selected α for both.
- Potency is near theory value for a 1-off test.
- Goodness-of-fit is not directly used to select models & no attempt is made to 'prove' that a given set of variables matters, and 'repeated testing' is not a concern, but choice of c_α affects R^2 and n through retention by $|t_{(n)}| \geqslant c_\alpha$.
- Likelihood estimation in general is feasible (Doornik (2009)).

Commence with general formulation – **general unrestricted model**:

$$y_t = \beta' z_t + \gamma' w_t + \sum_{j=1}^{T} \delta_{IIS,j} 1_{\{j=t\}} + \sum_{j=1}^{T-1} \delta_{SIS,j} 1_{\{j \leqslant t\}} + \nu_t \quad t = 1, \ldots, T$$

- Embed theory z_t
- Expand model w_t (almost costless if theory correct)
- Indicators δ_t (almost costless under null)
- Ensuring valid conditioning – exogeneity
  - Testable (Weak E., Strong E., Super E.)
- Theory Motivation: Reduction from DGP to LDGP to GUM to Specific

Doornik, J. A. (2009).
Autometrics.
pp. 88–121. Oxford: Oxford University Press.

Engle, R. F. and D. F. Hendry (1993).
Testing super exogeneity and invariance in regression models.
Journal of Econometrics 56, 119–139.
Reprinted in Ericsson, N. R. and Irons, J. S. (eds.) *Testing Exogeneity*, Oxford: Oxford University Press, 1994.

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

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

Hendry, D. F. (2017).
Deciding between alternative approaches in macroeconomics.
International Journal of Forecasting. Forthcoming.

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

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: 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.**

Lucas, R. E. (1976).
Econometric policy evaluation: A critique.
In K. Brunner and A. Meltzer (Eds.), *The Phillips Curve and Labor Markets*, **Volume 1 of** *Carnegie-Rochester Conferences on Public Policy*, **pp. 19–46. Amsterdam: North-Holland Publishing Company.**

Pretis, F., J. Reade, and G. Sucarrat (2016).
General-to-specific (gets) modelling and indicator saturation with the r package gets.
Oxford Department of Economics Discussion Paper 794.