Robust Discovery of Regression Models

Jennifer L. Castle [a], Jurgen A. Doornik [b], David F. Hendry [b,*]

[a] *Magdalen College and Climate Econometrics, University of Oxford, Oxford, UK*
[b] *Nuffield College and Climate Econometrics, University of Oxford, Oxford, UK*

ARTICLE INFO

ABSTRACT

Successful modeling of observational data requires jointly discovering the determinants of the underlying process and the observations from which it can be reliably estimated, given the near impossibility of pre-specifying both. To do so requires avoiding many potential problems, including substantive omitted variables; unmodeled non-stationarity and mis-specified dynamics in time series; non-linearity; and inappropriate conditioning assumptions, as well as incorrect distributional shape combined with contaminated observations from outliers and shifts. The aim is to discover robust, parsimonious representations that retain the relevant information, are well specified, encompass alternative models, and evaluate the validity of the study. An approach is proposed that provides robustness in many directions. It is demonstrated how to handle apparent outliers due to alternative distributional assumptions; and discriminate between outliers and large observations arising from non-linear responses. Two empirical applications, utilizing datasets popularized in previous applications, show substantive improvements from the proposed approach to robust model discovery.

1. Introduction

Robustness is 'a certain resilience of conclusions to deviations from assumptions of hypothetical models' (Koenker, 1982) using 'procedures that are not influenced too much by small deviations from the distributional assumptions of the model' (Ronchetti, 1985). Both are desirable, but to achieve such aims, robustness must go beyond just estimating a pre-specified model by a 'robust method' assuming omniscience in all other aspects of its formulation. Complete and correct *a priori* specifications almost never exist for models of observational data, so model discovery is unavoidable. The target must be to discover the data generating process (DGP) for the variables being modeled (or a good approximation thereto) while embedding the objective of the analysis, which is often a theory-based formulation. In a homely analogy, you carefully select a beautiful red apple from a tree, but on cutting it, you discover it is filled with bugs. Applied to empirical modeling, an equation is selected from a set of regressors, but on fitting to a subsample, you discover the estimated parameters are very different.

---

* Corresponding Author.
  E-mail addresses: jennifer.castle@magd.ox.ac.uk (J.L. Castle), jurgen.doornik@nuffield.ox.ac.uk (J.A. Doornik), david.hendry@nuffield.ox.ac.uk (D.F. Hendry).

Given a lack of omniscience, successful discovery requires robustifying models against as many contaminating influences as possible—with our potential solutions—namely:

C1 omitted relevant variables—so initially include all likely explanatory variables;
C2 erroneously included irrelevant variables—rigorous selection;
C3 mis-specified linearity—by a low-dimensional non-linear representation;
C4 outliers & incorrect distributional assumptions—by impulse-indicator saturation (IIS);
C5 location shifts and other non-constant parameters—by step-indicator saturation & multiplicative-indicator saturation (SIS and MIS);
C6 invalid conditioning—by checking invariance and exogeneity;
C7 inadequate dynamics—add sufficient lags for a sequential factorization;
C8 stochastic trends—by cointegration and differencing.

Problems C7 and C8 only relate to models for time series, while the others are also relevant in cross-section models. Our methods were initially developed for dynamic regression models (see Hendry and Doornik, 2014), but the applications below show that they are also relevant in cross-section models. This enables a comparison with Lasso estimation in the second application. We use time-series notation to denote the estimation sample by $t = 1, \ldots, T$, as N relates to the number of candidate variables and M to the number of simulation replications.

C1–C8 need to be addressed as jointly as possible to avoid confounding problems, albeit C6 is after modeling. To allow for these potential problems, we start modeling from a specification that is sufficiently general to characterize the target, yet permits evaluation of the objective of the analysis. This initial specification will have more candidate variables, N (not all of which need be relevant), than observations, T (not all of which may be accurately measured), so the second key component of discovery is a machine-learning multi-block search algorithm to select relevant influences and eliminate irrelevant variables and outlier data points (here we use *Autometrics*: see Doornik, 2009 and Section 3).

The structure of the paper is as follows. Section 2 introduces automatic model discovery as a way to mitigate the contaminating influences. The first part, Section 2.1, addresses formulation of the initial general model (problem C1 and C2), and creating functional-form transformations for non-linearities (problem C3). Next, Section 2.2 considers indicator saturation estimators for outliers, location shifts, and non-constant parameters (problems C4 & C5). Section 2.3 applies IIS facing incorrect distributional assumptions, and compares IIS with the robust method least trimmed squares (LTS). Section 2.4 describes testing invariance and exogeneity (problem C6). The important role of retaining subject-matter theory is discussed in Section 2.5. Modeling sufficient lags to capture dynamics in time series (problem C7) and handling stochastic trends (problem C8) are both well understood, so are not explicitly addressed here. Section 3 provides a detailed explanation of the algorithm for selection of relevant regressors when there are more variables than observations. We call these 'short-data models', although the algorithm can be used more generally. Section 4 concerns discriminating between non-linearities and outliers with a simulation illustration (see Stillwagon, 2016, for an empirical application). Section 5 reconsiders two empirical studies utilizing established datasets to demonstrate how robust model discovery improves over previously reported results. The first application re-analyzes a much-studied Boston housing market data set, then Section 5.2 unweaves the earlier study by Hettmansperger and Sheather (1992) that found an instability in least median of squares. Section 6 concludes, and Appendix A provides some simulation evaluations of the algorithm with comparisons to Lasso.

2. Automatic model discovery

Economies, societies and environmental systems evolve over time and are intermittently hit by natural shocks, wars, crises, policy changes, pandemics and other, often unanticipated, events. The lack of complete and correct *a priori* formulations for such non-stationary observational data makes specification search unavoidable to discover what actually matters. We first describe the initial model formulation of models that are robust to many of the directions of potential mis-specifications in C1–C8. We call the approach 'model discovery' as its application can reveal features that were not initially envisaged, while emphasizing that it is not purely data-based, but retains any subject-matter theory insights as in Hendry and Johansen (2015).

Although the initial models may look infeasibly large, to be robust to as many forms of potential mis-specification as feasible, very general specifications are needed. Robustness cannot be achieved unless all the complications are tackled as jointly as possible, otherwise mis-specification in one direction can lead to another aspect of the model proxying that, resulting in wrong inferences. An example is non-modeled outliers (problem C4) that lead a modeler to detect apparent non-linearities that are just an artifact. By modeling the outliers jointly with possible non-linearities, a modeler can discriminate between the competing explanations on data evidence rather than *ad hoc* imposed assumptions. Section 4 examines this particular example further.

2.1. Formulation of the general unrestricted model (GUM)

Given r variables $\mathbf{w}_t, t = 1, \ldots, T$, considered by an investigator to be sufficient to model a variable of interest y_t, these are partitioned into $\mathbf{w}'_t = (\mathbf{x}'_t : \mathbf{v}'_t)$, where $\mathbf{x}_t$ are r_1 theory specified variables (their parameters are the 'objective' of the study) and are not subject to selection, with the remaining $r_2 = r - r_1$ variables $\mathbf{v}_t$, which are orthogonalized with respect

to $\mathbf{x}_t$, to tackle problem C1. Three extensions to $\mathbf{w}_t$ can be automatically implemented to create the GUM, namely functional-form transformations for non-linearities, indicator variables to capture outliers and shifts, and dynamics.

Departures from linearity, C3, can be handled by including a small set of non-linear transformations of the principal components of $\mathbf{w}_t$ (see Castle and Hendry, 2011), although we do not use this approach in the empirical applications below.

Next, to tackle problems C4 and C5 of incorrect distributional assumptions, potential outliers, non-constant parameters and location shifts, create T impulse indicators, $\mathbf{1}_{\{i=t\}}$, which are zero except for unity at observation t for $t = 1, \ldots, T$ and/or step indicators $\mathbf{1}_{\{i \le t\}}$, possibly interacting with some regressors (depending on the problem under analysis), to be added to the set of candidate variables (Section 2.2 below explains the saturation estimators).

For time series, creating s lags of $(y_t, \mathbf{w}_t)$ implements a sequential factorization (see Doob, 1953) to tackle problem C7 (see Castle et al., 2011), and cointegration (problem C8) can be investigated as a reduction to a formulation without unit roots (see Ericsson and MacKinnon, 2002).

The resulting general unrestricted model (GUM) is given by:

$$y_t = \mu^R + \sum_{i=1}^{r_1} \theta_i^R x_{i,t} + \sum_{i=1}^{r_2} \theta_i v_{i,t} + \sum_{i=1}^{T} \delta_i \mathbf{1}_{\{i=t\}} + \sum_{i=2}^{T-1} \psi_i \mathbf{1}_{\{i \le t\}} + \left[\sum_{i=1}^{r} \sum_{j=1}^{s} \phi_{ij} w_{t-j} \right] + \left[\sum_{j=1}^{s} \rho_j y_{t-j} \right] + \epsilon_t, \tag{1}$$

where time-series components are in brackets, and the errors ϵ_t are independent and normally distributed with constant variance: $\epsilon_t \sim N[0, \sigma_\epsilon^2]$ to be tested after selection. The intercept and $\mathbf{x}_t$ are always retained, denoted by the superscript R on the coefficients, and therefore not subject to selection. When both impulse and step indicators are included, (1) has $N = r_2 + s(r+1) + 2(T-1)$ candidate regressors, so $N \gg T$. We assume that r_1 is small relative to the sample size.

Selecting from a GUM such as (1) can be automated in many ways: here we mainly use *Autometrics* (Doornik, 2009), augmented with a procedure to handle more variables than observations (Section 3). *Autometrics* seeks parsimonious well-specified final representations that retain theory insights and encompass rival models' results. Bontemps and Mizon (2008) provide an overview of encompassing and its relation to non-nested tests as in Cox (1962), and Doornik (2008) describes its role in *Autometrics*. In practice we may find several candidate final models, which are statistically close. A tie-breaker is used in that case.

A GUM like (1) is inevitably highly over-parametrized. Indeed, for $r = 10$, $s = 2$, $T = 100$ and $r_1 = 4$, there are 230 regressors in the GUM for only 100 observations. It might be thought that selection from such a large set as 2^{230} possible models would lead to an excessive null retention frequency. However, for Normal errors, provided the selection significance level α is appropriately controlled for different decisions, the probability of retaining irrelevant variables can be small: see Hendry and Doornik (2014) and Johansen and Nielsen (2009, 2016). As discussed below, IIS removes outliers to allow Normal critical values to be used. When selecting impulse and step indicators, all other variables are temporarily retained, so setting $\alpha \approx 1/(2T) = 0.005$ for $T = 100$, under the null of no outliers or shifts, on average one indicator will be adventitiously retained. Under the alternative, indicator coefficients greater than 2.85 times their estimated standard errors will be retained. Then keeping those selected indicators, selection over the other $r_2 + s(r+1)$ non-retained variables in (1) can be undertaken at 1%, still requiring relatively strong evidence to be included given $\mathbf{x}_t$.

2.2. Indicator saturation estimators

Indicator saturation estimators are a general class of methods seeking robust inference in the presence of unknown numbers and locations of outliers, shifts, breaks and parameter changes by designing indicators appropriate to the problem. Such methods do not require the numbers, signs, timings, magnitudes or durations of the breaks to be known in advance, and can handle shifts at any point in the sample (including the last observation). Five indicator-saturation techniques that have seen empirical applications are:

IIS   impulse-indicator saturation for outliers (Hendry et al., 2008, and Johansen and Nielsen, 2009);
SIS   step-indicator saturation for location shifts (Castle et al., 2015);
TIS   trend-indicator saturation for trend breaks (Castle et al., 2019, with an application in Walker et al., 2019);
DIS   designed-indicator saturation for specific shapes (matching volcanic eruption impacts on temperature in Pretis et al., 2016);
MIS   multiplicative saturation for changes in other parameters (Ericsson, 2012, Kitov and Tabor, 2015).

All saturation estimators lead to formulations with more variables than observations. Feasible estimators select from the candidate variables in blocks. In the simplest case of IIS, this is the same as partitioning the estimation sample to select reliable observations. But, when combined with selection over other variables, it is more convenient to treat each impulse indicator as just another variable. Selection over blocks of indicators proceeds iteratively until convergence. After this, a final model selection step can be undertaken. Section 3 provides details of the algorithm. In general, different partitionings into blocks can lead to different selected models. Efficient implementation of the estimation algorithm is useful considering the large search space.

To derive the null distribution of IIS, Hendry et al. (2008) used a 'split-half analysis' where each half of the indicators is entered in turn, recording any outliers. Similar analyses have been applied to the other saturation estimators, modified by their specific formulations, noting that SIS and TIS are MIS for the intercept and trend respectively. Both SIS and TIS are

used in the COVID-19 forecasts of Doornik et al. (2020). TIS illustrates a general property of saturation estimators: you not only discover when and how many shifts have occurred, but retained trends match those you would have found on that subsample historically. TIS, DIS and MIS for changes in regression parameters are not considered further here.

Johansen and Nielsen (2009) develop a more complete IIS theory for both stationary and unit-root autoregressions. When the error distribution is symmetric, in a stationary regression with k retained variables, a parameter of interest $\boldsymbol{\beta}$ and scaled asymptotic second moment $\boldsymbol{\Sigma}$, selecting from T impulse indicators at the critical value c_α, under the null of no outliers, leads to:

$$T^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\mathrm{D}}{\to} \mathrm{N}_k\big[\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega}_\alpha\big], \tag{2}$$

where N_k is the k-variate normal distribution. The efficiency of the IIS estimator $\widetilde{\boldsymbol{\beta}}$ with respect to OLS $\widehat{\boldsymbol{\beta}}$ measured by $\boldsymbol{\Omega}_\alpha$ depends on c_α and the distribution, but is close to $(1-\alpha)^{-1}\mathbf{I}_k$ for small α. Thus even with $N > T$, the usual $\sqrt{T}$ stationary convergence rate to a Normal distribution holds, correctly centered on $\boldsymbol{\beta}$ and with almost the usual asymptotic variance matrix $\sigma_\epsilon^2 \boldsymbol{\Sigma}^{-1}$ but weighted by the efficiency factor. Despite T extra candidates, there is a small loss of efficiency under the null for small α, against potentially large gains under alternatives of multiple outliers and shifts. Johansen and Nielsen (2016) link IIS to robust statistics by showing it is an iterated 1-step Huber-skip M-estimator given the model, so we now compare IIS to least trimmed squares (LTS) facing selection with a fat-tailed error distribution. Both IIS and LTS classify observations as outliers.

2.3. Comparing IIS and LTS facing fat-tailed distributions

Least trimmed squares (LTS) and least median of squares (LMS) are robust estimators introduced by Rousseeuw (1984). LTS(0.5) finds the half of the observations that minimize the residual sum of squares (LMS does the same for the median). LTS is often used as a starting point for more efficient robust methods, because it does not require a preliminary estimate of the scale. LMS has a lower rate of convergence, and is less used. Víšek (1999) provides asymptotic analyses, while Berenguer-Rico et al. (2019) derive settings where LTS and LMS are the maximum likelihood estimators.

Inference in empirical modeling is often based on approximate Normality, so a general selection procedure that is robust to incorrect distributional assumptions, yet remains relatively efficient under the null, is valuable. Following Johansen and Nielsen (2009), IIS can provide low-cost robustness to fat-tailed error distributions such as Student-t distributions with small degrees of freedom, as impulse indicators capture the 'outliers'. An important advantage of IIS when the error distribution is unknown is that approximately correct Normal critical values can be used to select regressors. We illustrate this property as part of discovering the underlying DGP, and compare it to applying LTS for a given model.

Consider the regression in (3) with $\mathbf{z}_t' = (z_{1,t}, \cdots, z_{12,t})$ independent and identically distributed (IID), and fixed across replications:

$$y_t = \beta_1 z_{1,t} + \cdots + \beta_{12} z_{12,t} + \epsilon_t, \ t = 1, ..., T, \tag{3}$$

$$\mathbf{z}_t \sim \mathrm{N}_{12}[\mathbf{0}, \mathbf{I}_{12}], \tag{4}$$

$$\epsilon_t \sim \mathrm{t}(3), \ \text{so} \ \sigma_\epsilon \approx 1.73. \tag{5}$$

We created 3 different DGPs from (3), all with $\beta_7 = \cdots = \beta_{12} = 0$:

DGP-A : $\beta_1 = \cdots = \beta_6 = 0$;
DGP-B : $\beta_1 = 0.1; \beta_2 = 0.2; \beta_3 = 0.3; \beta_4 = 0.4; \beta_5 = 0.6; \beta_6 = 0.8$;
DGP-C : as DGP-B, except (4) is replaced by $z_{j,t} \sim \mathrm{t}(3), \ j = 1, \ldots, 12$.

Four selection simulation experiments were undertaken for each DGP with and without IIS:

(NONE) no selection and no IIS,
(IIS) selection is only conducted for IIS so all 12 variables are retained,
(Z) selection is over the z_i variables but without IIS,
(IIS+Z) conducts selection for both variables and impulse indicators.

All decisions are made using the critical value $c_\alpha = 2.85$ (Normal significance level $\alpha = 0.005$) from the initial model:

$$y_t = \mu^R + \gamma_1 z_{1,t} + \cdots + \gamma_{12} z_{12,t} + u_t.$$

The intercept was always included, so is omitted from the evaluation. We ran $M = 5000$ replications with $T = 100$, using common random numbers between settings to reduce simulation variance.

The effectiveness of selection procedures is measured through the *gauge*, which is defined as the fraction of retained irrelevant variables (the null retention frequency: see Johansen and Nielsen, 2016), and the *potency*: the average retention frequency of DGP variables.

Table 1

MCSDs and average retention rates of z variables in DGP-B, with and without IIS, and with and without variable selection. Conditional MCSD is for selected variables; retention corresponds to *potency* for relevant variables, and *gauge* for irrelevant; $M = 5000$, $\alpha = 0.005$.

DGP-B($\cdot$)	none	iis	z	iis+z	z	iis+z	z	iis+z
	GUM MCSD (z)				Conditional MCSD (z)		Retention rate (z)	
Relevant	0.18	0.15	0.22	0.19	0.18	0.12	**0.394**	**0.454**
Irrelevant	0.18	0.15	0.06	0.04	0.30	0.36	*0.032*	*0.009*
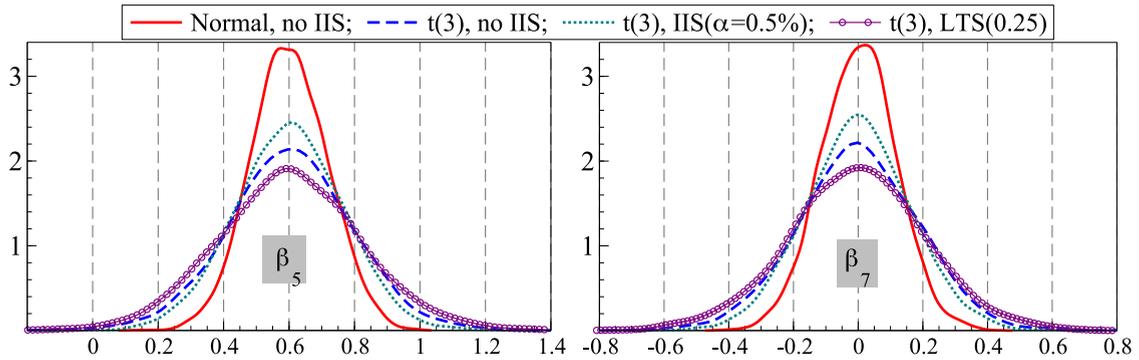
Fig. 1. Densities of estimated coefficients of $z_{5,t}$ (left) and $z_{7,t}$ (right) in DGP-B without variable selection. Normal errors (solid line) and t(3) errors (dashed line). IIS with t(3) errors (dotted line) and LTS(0.25) with t(3) errors (line with circles).

A t-test on an impulse indicator is $\widehat{\delta}_i$ divided by $\widehat{\sigma}_\epsilon$, so for $c_\alpha = 2.85$, $\widehat{\delta}_i$ would need to exceed 4.94 to be retained, where $\Pr(|\widehat{\delta}_i| \geq 4.94) \approx 0.016$. The simulation average retention rates of impulse indicators lay in a narrow range between 0.028–0.032 across the six experiments with IIS, noting that $\widehat{\sigma}_\epsilon < \sigma_\epsilon$ once the largest outliers are removed. This outcome also applied to DGP-C despite a t(3) distribution for all $z_{i,t}$, so IIS did not get confused by conditioning variables having the same distribution as the error. However, the Monte Carlo standard deviations (MCSDs) were large at around 0.018, reflecting considerable variation in the distributions of retention rates of indicators across replications, with a maximum of around 14%.

Table 1 records the MCSDs for the z variables in DGP-B for the GUM and selected model. Results are averaged separately across relevant (i.e., with $\beta_i \neq 0$) and irrelevant variables. If the DGP distribution is incorrectly assumed to be Normal when it is t(3), the gauge of 3.2% is too high when $\alpha = 0.5\%$. Using $c_\alpha = 2.85$, the average power of separate t-tests on (3), each given by $t_{\beta_i=0} = \widehat{\beta}_i/\text{SE}[\widehat{\beta}_i] \to \sqrt{T}\beta_i/\sigma_\epsilon = 5.77\beta_i$, would be 0.36, so both selection and IIS improve potency. Also with IIS, the gauge falls to 0.9%: IIS removes the fat tails, bringing the error distribution closer to Normality. In these simulations using IIS, selection is not too 'over-gauged' for irrelevant regressors and can improve potency. Conditional MCSDs are large for irrelevant variables which are only selected when their coefficients are far from zero, which happens rarely even for t(3) errors, as the gauge of 0.9% shows.

Fig. 1 plots the densities of two parameter estimates in the GUM for DGP-B. In the first case, the solid line, the errors are standard Normal. The second case, the dashed line, is for DGP-B(NONE), showing that neglected t(3) errors make inference less precise. Using IIS, as shown in the dotted line, moves the density towards that with Normal errors. If the error distribution were known, selection could use criteria based on that. In practice, with unknown error distributions, IIS offers some robustness, so that using critical values based on Normality is then a reasonable approximation.

The lines with circles in Fig. 1 show the densities of the estimated coefficients when using LTS(0.25), i.e. removing a quarter of the observations. While the objectives are the same, IIS and LTS have a very different impact on the estimated coefficients in this setting: here, LTS exacerbates the problem caused by t(3) errors. Although jointly selecting variables and indicators with IIS is easy, it is not with LTS which only achieves robustness to outliers.

2.4. Testing the validity of conditioning using IIS and SIS

To test parameter constancy and valid conditioning in the conditional equation of interest, IIS can be applied to models of its putative exogenous regressors, regarded as marginal processes. To illustrate the procedure, consider the bivariate Normal distribution:

$$\begin{pmatrix} y_i \\ z_i \end{pmatrix} \sim \text{N}_2 \left[\begin{pmatrix} \mu_{y,i} \\ \mu_{z,i} \end{pmatrix}, \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix} \right] = \text{N}_2[\mu_i, \boldsymbol{\Omega}],$$

where $\mathrm{E}[y_i] = \beta_0 + \beta_1 \mathrm{E}[z_i]$ is the theory model of interest so $\mu_{y,i} = \beta_0 + \beta_1 \mu_{z,i}$. Then for $\gamma = \omega_{22}^{-1} \omega_{12}$, the conditional expectation is:

$$\mathrm{E}[y_i \mid z_i] = \mu_{y,i} + \omega_{22}^{-1} \omega_{12} (z_i - \mu_{z,i}) = \beta_0 + (\beta_1 - \gamma)\mu_{z,i} + \gamma z_i. \tag{6}$$

Thus, unless $\beta_1 = \gamma$, the conditional expectation $\mathrm{E}[y_i|z_i]$ depends on $\mu_{z,i}$, and will shift whenever the mean of z_i changes. Valid conditioning and parameter constancy in the regression of y_i on z_i depend on the absence of $\mu_{z,i}$ from (6), which hypothesis can be tested by discovering all the shifts in the marginal model of z_i and testing their relevance in the conditional regression. IIS and SIS provide ways of doing so: see Hendry and Santos (2010) and Castle et al. (2017).

2.5. Retaining subject-matter theory

When the theory-model objective is retained while selecting over a much larger nesting specification, it becomes the null hypothesis to be stringently evaluated against a range of likely alternatives (see Mayo and Spanos, 2006, and Mayo, 2018). Free lunches are rare in economics but this approach will corroborate the theory when it is complete and correct, in the sense of Karl Popper (1959,1963), and deliver an improved model when the theory-model is rejected by additional variables being highly significant. Although discovering which variables led to rejection does not fully overcome objections to 'accepting the alternative' when the null is rejected (see e.g., Harding, 1976), such information points towards an improved model–which could even be consistent with the initial theory (see e.g., Hendry and Mizon, 2011).

3. Selection in short-data models

Given initial GUMs may have more candidate regressor variables, N, than observations, T, we address selection of regression models in that situation. We assume sparsity in that not all N variables matter, and we wish to find those that do, and also that the final models are sufficiently small to be estimated using standard regression methods. In practice, we start using the methods of this section as soon as N gets close to the sample size T; our rule for this is $N > 0.75T$.

Doornik and Hendry (2015) identify three basic shapes of the design matrix for 'big data', namely 'tall' (not so many variables but many observations, with $T \gg N$), 'fat' (many variables, but not so many observations, $N > T$) and 'huge' (many variables and many observations, $T > N$). This section discusses the fat design case, noting that the sample size need not be large for this to occur. For convenience, we refer to the fat big-data design as 'short data', which better reflects the setting.

There is a potential for short data in many settings, e.g., when modeling developing economies, or allowing for many effects with generous lag lengths. They also naturally arise with saturation estimators, e.g., when adding an impulse indicator ('dummy') for every observation with IIS (Section 2.2).

When selection is jointly over indicators and variables, it is more convenient to treat them in the same way, so the split-sample analysis in (e.g.) Hendry et al. (2008) is not feasible. An alternative is proposed in the next subsection where the candidate set is still partitioned in blocks, but expanding and contracting searches then alternate until convergence. Selecting blocks of impulses in a regression model with IIS amounts to dropping observations. This is not the case in more general settings when we also select over variables.

With saturation estimation the number of regressors grows with the sample size. When T is large, say $T > 512$, it could be useful to divide the problem into smaller separate sections. With each at 256, e.g., the operation count is reduced from order T^3 to a multiple of $256T^2$.

3.1. A learning algorithm for short data

3.1.1. Setup
The basic setup consists of a $T \times P$ multivariate dependent variable $\mathbf{Y} = (y_{it})$, made up of P individual variables, each of T observations. There are N potential explanatory variables, collected in the $T \times N$ matrix $\mathbf{X} = (x_{it})$. To describe the selection of columns of $\mathbf{X}$, we consider this to be our set of candidate variables $\mathcal{X} = \{x_1, ..., x_N\}$.

At our disposal is a selection method M. Assume that some part $\mathcal{X}^R$ of $\mathcal{X}$ is retained in the model, but their coefficients are freely estimated. Write $\overline{\mathcal{X}}$ for the free variables, i.e., $\mathcal{X}$ with $\mathcal{X}^R$ removed: $\overline{\mathcal{X}} = \mathcal{X}/\mathcal{X}^R$. M is used to select (the target $\mathbf{Y}$ is retained throughout):

$$\mathcal{S} = \mathrm{M}(\overline{\mathcal{X}} \mid \mathcal{X}^R).$$

M selects a subset $\mathcal{S}$ of $\mathcal{X}$ that is not already retained, so the complete final model is $\mathcal{S} \cup \mathcal{X}^R$.

It is often the case that M needs more observations than variables to be operational. Then a way around this restriction is to partition the candidate set in B smaller blocks:

$$\overline{\mathcal{X}} = \overline{\mathcal{X}}_1 \cup ... \cup \overline{\mathcal{X}}_B, \tag{7}$$

and apply model selection to each block:

$$\mathcal{S} = \cup_{i=1}^{B} \mathrm{M}(\overline{\mathcal{X}}_i \mid \mathcal{X}^R). \tag{8}$$

Now (8) enables 'short data' estimation with $N \geq T$, provided we keep $\dim(\overline{\mathcal{X}}_i \cup \mathcal{X}^R) < \eta T$, which allows for the use of a standard estimation method ($0 < \eta < 1$).

Table 2

Algorithms for the expansion step

$\mathbf{E}_1(\overline{\mathcal{X}} \mid \mathcal{C}; \; \alpha, N^B)$:
  1. Partition $\overline{\mathcal{X}} = \overline{\mathcal{X}}_1, ..., \overline{\mathcal{X}}_B$ and select: $\mathcal{S} = \cup_{i=1}^B M(\overline{\mathcal{X}}_i \mid \mathcal{C}; \; \alpha)$.
  2. Sort the elements of $\mathcal{S}$ by their significance, most significant first.
  3. Return the sorted set.   ∎
$\mathbf{E}(\overline{\mathcal{X}} \mid \mathcal{C}; \; \alpha, N^{\min}, N^B, \lambda)$:
  1. Let $\mathcal{S} = \mathbf{E}_1(\overline{\mathcal{X}} \mid \mathcal{C}; \; \alpha, N^B)$.
  2a. If $\dim \mathcal{S} < N^{\min}$ reselect $\mathcal{S} = \mathbf{E}_1(\overline{\mathcal{X}} \mid \mathcal{C}; \; f(\alpha, \lambda), N^B)$;
  2b. else if $\dim \mathcal{S} > N^B$ reselect $\mathcal{S} = \mathbf{E}_1(\mathcal{S} \mid \mathcal{C}; \; \alpha, N^B)$.
  3. Return the first N^B variables of $\mathcal{S}$.   ∎

Table 3

Learning from expansion and reduction

$\mathbf{L}(\mathcal{X} \mid \mathcal{C}; \; \widetilde{\mathcal{C}}, \alpha_e, \alpha_r, N^{\min}, N^B, N^{\max}, j^{\max}, \lambda)$:
  1. Set $\mathcal{C}^{(0)} = \mathcal{C}$, $\widetilde{\mathcal{C}}^{(0)} = \widetilde{\mathcal{C}}$, $j = 0$. If $\dim \mathcal{C}^{(j+1)} \geq N^{\max}$ terminate, else go to *expansion*:
  2. *Expansion* Set $\overline{\mathcal{X}}^{(j)} = \mathcal{X}/\mathcal{C}^{(j)}$ and $\mathcal{S}^{(j)} = \mathbf{E}(\overline{\mathcal{X}}^{(j)} \mid \mathcal{C}^{(j)}; \; \alpha_e, N^{\min}, N^B, \lambda)$; then:
  3. *Reduction* $\mathcal{C}^{(j+1)} = \mathbf{E}_1(\mathcal{S}^{(j)} \cup \mathcal{C}^{(j)}; \; \alpha_r, N^{\max})$ and set $\widetilde{\mathcal{C}}^{(j+1)} = \widetilde{\mathcal{C}}^{(j)} \cup \mathcal{C}^{(j+1)}$; then:
  4. *Termination* if $j = j^{\max}$ or $\widetilde{\mathcal{C}}^{(j+1)} = \widetilde{\mathcal{C}}^{(j)}$ or $\dim \mathcal{C}^{(j+1)} \geq N^{\max}$:
       then finish with $\mathcal{C}^{(j+1)}$, $\widetilde{\mathcal{C}}^{(j+1)}$,
       else increment j and return to *expansion*.   ∎

An algorithm for estimating short data models can be built upon (7) and (8). Some form of iteration will be needed: we may, e.g., discover the most important variable in the final block, making all previous block estimates effectively void. We describe an algorithm that learns from previous rounds of estimation, and investigate its properties through some Monte Carlo experiments.

Thus, we propose to alternate between 'expansion' and 'reduction' steps while the selected set changes. The expansion step selects from all blocks, while the reduction consolidates this into a candidate model. The main practical consideration, namely how to choose the blocks, is deferred.

3.1.2. Expansion step

Let α be the adopted significance level (or more generally, model selection settings for M), and N^B the target block size. The first part of the expansion step in Table 2 is a procedure $\mathbf{E}_1$ that splits the free candidates $\overline{\mathcal{X}}$ into blocks and selects given the currently retained set $\mathcal{C}$. The partitioning procedure is described later. The selection from blocks in step 1 could be run in parallel.

Procedure $\mathbf{E}$ is built around $\mathbf{E}_1$, imposing limits on the size of the selected set. If the selection is too small (line 2a), an optional 'boost' of magnitude $\lambda > 1$ is applied (but only if $N^{\min} > 0$) and selection repeated at a relaxed significance level. This offers some protection against missing a factor that may matter, possibly at the expense of raising the gauge under the null that nothing matters. [$f(\alpha, \lambda) = \lambda\alpha$ if $0.94 \leq 1 + \alpha(\lambda - 1) \leq 1.06$; $f(\alpha, \lambda) = \lambda\alpha[1 + \alpha(\lambda - 1)]^{-1}$ otherwise.] On the other hand, if the selected set is too large (line 2b), $\mathbf{E}_1$ is applied to the selection. Then, if still too large the N^B most significant variables are returned, where the significance is based on the estimates in each block.

3.1.3. Expansion and reduction

The 'learning' algorithm $\mathbf{L}$ alternates between expansion and reduction until no new variables are discovered. The expansion step is given as $\mathbf{E}$, the reduction can be done using $\mathbf{E}_1$, both defined in Table 2. As part of the learning process we wish to keep track of two sets. The first is the current model $\mathcal{C}$ after each expansion/reduction pair; the second is $\widetilde{\mathcal{C}}$, the history of variables that have been selected. Note that variables in the history need not be in the current model anymore. The learning process can now be expressed as in Table 3.

The remaining arguments for $\mathbf{L}$ mainly relate to the significance level and dimensions of the blocks. A separate significance level is specified for expansion and for reduction, α_e and α_r, respectively, but they are usually the same. $N^{\min}$ and $N^{\max}$ are lower and upper bounds of the number of selected variables, while N^B is the block size. Finally, λ is the optional boost, and the maximum number of iterations is set through $j^{\max}$.

The default block size is set to $N^B = \min(\lceil 0.2T \rceil, 128)$, although another value can be specified. $N^{\min}$ is mostly zero, and

$$N^{\max} = \lfloor \kappa(T - N^R - [P - 1]) \rfloor - \lceil 0.2T \rceil,$$

where $\kappa = 0.8$ by default, leaving space for the largest expansion set based on the default N^B. N^R denotes the number of variables that are retained throughout. When the candidate model size reaches $N^{\max}$ at any stage, the algorithm terminates prematurely. The value of λ is 8 for $\alpha_e \leq 0.01$, 4 for $\alpha_e \leq 0.02$, 2 for $\alpha_e \leq 0.03$, and 1 otherwise.

To have more control over the algorithm, and provide some speed-up, we divide the iterations in procedure $\mathbf{L}$ into four stages. Stage A is just the first iteration, starting from the empty model and history. It has the expansion boost if the initial selection is too small. The first iteration of stage B also has this boost, provided the selection is too small and it was not used in A. Otherwise $N^{\min} = 0$. Stage B is the continuation of A, up to ten iterations or convergence, whichever comes first.

Table 4

Block search algorithm with learning

A. $[\mathcal{C}^A, \widetilde{\mathcal{C}^A}] = \mathbf{L}(\mathcal{X} \mid \emptyset; \quad \emptyset, \; \alpha, \alpha, N^{\min} = \min(\frac{1}{8}N^B, 8), N^B, N^{\max}, 1, \lambda);$
B. $[\mathcal{C}^B, \widetilde{\mathcal{C}^B}] = \mathbf{L}(\mathcal{X} \mid \mathcal{C}^A; \widetilde{\mathcal{C}^A}, \alpha, \alpha, N^{\min} = \min(\frac{1}{8}N^B, 8) [\text{or } 0 \text{ see text}], N^B, N^{\max}, 10, \lambda);$
C. $[\mathcal{C}^C, \widetilde{\mathcal{C}^C}] = \mathbf{L}(\mathcal{X} \mid \mathcal{C}^B; \mathcal{C}^B, f(\alpha, 2), f(\alpha, \frac{1}{2}), 0, N^B, N^{\max}, 1, 1);$
D. $[\mathcal{C}^D, \widetilde{\mathcal{C}^D}] = \mathbf{L}(\mathcal{X} \mid \mathcal{C}^C; \widetilde{\mathcal{C}^C}, \alpha, \alpha, 0, N^B, N^{\max}, 10, 1).$
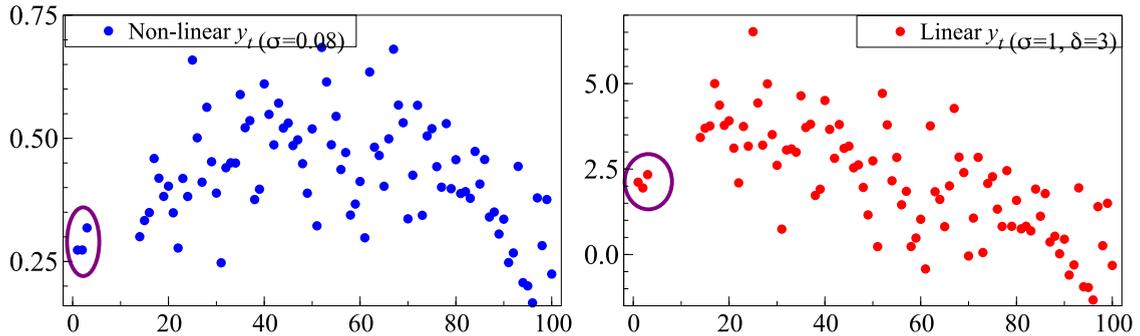
Fig. 2. Data from a quadratic DGP (panel a, left) and a linear DGP with outliers in the first three observations (panel b, right), in both cases with observations $4, \ldots, 13$ missing.

The history after stage B is reset to the current model (the third argument to $\mathbf{L}$ is $\mathcal{C}^B$ rather than $\widetilde{\mathcal{C}^B}$). Then stage C is a single iteration with a doubled significance level in expansion, $f(\alpha, 2)$, that is offset in the reduction, but only if $\alpha_e \leq 0.03$. Finally, stage D continues for up to ten iterations. This leads to our block search algorithm with learning in Table 4.

The nominal significance level $\alpha = \alpha_e = \alpha_r$ is the main control variable of the algorithm. It is chosen in advance, and kept fixed throughout; our preference is a value near $\min\{1/T, 1/(N + N^R)\}$.

Linking this procedure to an actual model selection device may require adjustments specific to that device. Appendix A presents the settings that we use with *Autometrics* and documents performance of several approaches in a range of simulation settings. We also use the learning algorithm to improve variants of the Lasso, Efron et al. (2004), in a short-data setting.

3.1.4. Limitations

The outcome of the learning algorithm will, in general, be sensitive to the composition of the blocks, in particular through the ordering of the variables and the block size. This is essentially a small-sample problem combined with limited computational power: a shortage of observations forces us to use the algorithm, while it is infeasible to estimate all possible models in most realistic settings. One option would be to experiment with some different orderings, at the cost of slower computation.

To facilitate replication, we always sort the expansion step by the database index of the variable, and within that by lag length for time series. Many other permutations are possible, but there seems to be some a priori benefit of keeping lags of a variable together, as these are often close substitutes. Random search would also be possible, but seems to have no practical advantage, except perhaps for asymptotic analysis. We experimented with several other methods of prior ordering, but found none that had a particular advantage. The likely reason is that the optimal ordering depends on knowing the final model.

4. Discriminating between non-linearities and outliers

We next investigate whether IIS can discriminate between non-linearities and outliers. We seek robustness in both directions, so if the DGP contains non-linearities, these should be modeled by selected non-linear functions rather than impulse indicators, and, if there are outliers, we wish to avoid retaining non-linear functions that proxy them.

In addition, two mis-specifications are introduced, truncation and contamination. Truncation leads to missing observations. This could result in non-linear functions appearing to be linear with outliers, or linear functions with outliers that are discrepant from the DGP in such a way as to give the impression of non-linearity. Fig. 2a shows a quadratic function with no outliers, while panel b is a linear function with observations 1,2,3, as outliers from a 3 standard deviation shift in the mean of the linear function. In both cases observations $4, \ldots, 13$ are missing, hampering discovery of the DGP.

Contamination occurs when some observations are discrepant relative to the DGP. Fig. 3 illustrates this by interacting ten consecutive observations with a step dummy. In Fig. 3a, observations $41, \ldots, 50$ of a quadratic DGP are subject to a 3 standard deviation shift in the mean of the function. In panel b such contamination is applied to a linear DGP.
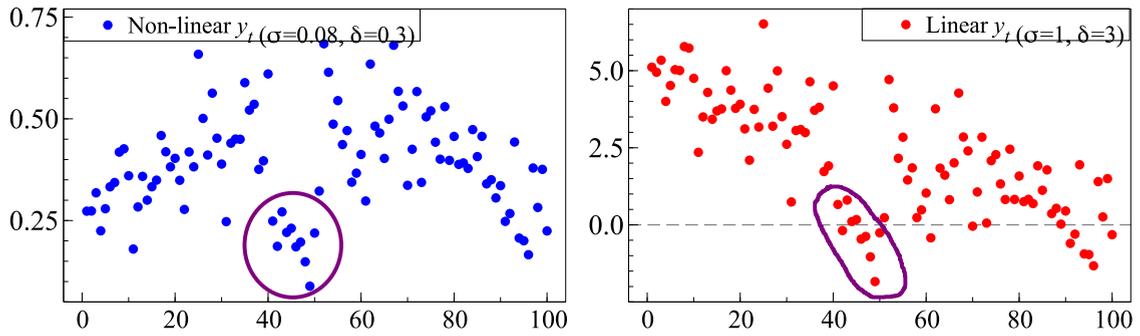
Fig. 3. Data from a quadratic DGP (panel a, left) and a linear DGP (panel b, right) with 10 contaminated observations shown in ellipses, given by a three standard deviation downward shift of observations $41, \ldots, 50$.

Table 5

Truncation experiments: observations $i = 4, \ldots 13$ missing. Contamination: observations $\mathcal{S}_1 = \{41, \ldots 50\}$ shifted by $\delta = -3$. Model selection of z_i, z_i^2 and IIS with $\alpha = 0.01$. *Rate* is the retention rate of variables and impulse indicators: **bold** denotes potency and *italic* gauge. Bias and RMSE are unconditional. $M = 5000$, $\alpha = 0.01$.

	Rate $\sigma = 0.08$	Bias	RMSE	*Rate* $\sigma = 0.04$	Bias	RMSE	*Rate* $\sigma = 1$	Bias	RMSE	*Rate* $\sigma = 0.5$	Bias	RMSE
	DGP-Q with truncation						DGP-L with truncation and three outliers					
z_i	**0.997**	−.0066	0.154	**1.000**	−.0017	0.073	**0.755**	1.355	2.607	**0.912**	0.537	1.543
z_i^2	**0.999**	.0061	0.139	**1.000**	.0015	0.066	*0.254*	−0.967	2.085	*0.102*	−0.365	1.268
$1_1, 1_2, 1_3$	*0.016*			*0.014*			**0.520**			**0.559**		
	DGP-Q with contamination						DGP-L with contamination					
z_i	**0.959**	−0.175	0.285	**1.000**	−0.073	0.106	**0.998**	−0.388	1.552	**1.000**	−0.191	0.762
z_i^2	**0.966**	0.180	0.285	**1.000**	0.075	0.106	*0.099*	0.460	1.539	*0.096*	0.226	0.754
$1_j, j \in \mathcal{S}_1$	**0.483**			**0.492**			**0.528**			**0.520**		
$1_j, j \notin \mathcal{S}_1$	*0.006*			*0.004*			*0.005*			*0.004*		

The quadratic and linear DGPs, using $z_i = i/T$, are given by:

$$y_i^* = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \sigma u_i, \quad u_i \sim N[0, 1], \ i = 1, \ldots, T,$$
$$\text{DGP-Q}: \quad \beta_0 = 0.25, \beta_1 = 1, \beta_2 = -1, \tag{9}$$

$$\text{DGP-L}: \quad \beta_0 = 5, \beta_1 = -5, \beta_2 = 0. \tag{10}$$

DGP-Q can be written as $y_i^* = 0.5 - (z_i - 0.5)^2 + \sigma u_i$. Two comparisons are considered:

1. DQP-Q with truncation: $y' = (y_1^*, y_2^*, y_3^*, y_{14}^*, \ldots, y_{100}^*)$ versus DGP-L with truncation and outliers:

   $$y' = (y_1^* + 3, y_2^* + 3, y_3^* + 3, y_{14}^*, \ldots, y_{100}^*).$$

   The truncated sample drops observations $i = 4, \ldots, 13$, so $T = 90$.
2. DQP-Q versus DGP-L both with contamination:

   $$y_i = y_i^* + \delta \sigma \left(\mathbf{1}_{\{41\}} + \ldots + \mathbf{1}_{\{50\}}\right).$$

The GUM consists of the intercept (always retained), z_i, z_i^2, and IIS. Table 5 reports the retention rates, biases, and root mean square errors (RMSEs) for selection using *Autometrics* at $\alpha = 0.01$ with $M = 5000$ replications. We consider two signal-to-noise ratios for each DGP with $\delta = -3$. Average retention over the first three indicators is reported, corresponding to outliers added to DGP-L, so reflect potency (bold in the table), but gauge for DGP-Q (italic). Fig. 4 shows the results for one draw, where retained indicators are labeled.

For the non-linear DGP-Q, the quadratic function is retained with near unit probability for both large and small signal-to-noise ratios, and the biases and RMSEs on the quadratic term are small: accurate estimates of the non-linearity can be obtained despite outliers, missing observations or contamination. The three irrelevant impulse indicators are retained close to the 1% target. For the linear DGP-L, the quadratic function is retained too often at 25% for $\sigma = 1$, so the indicators are only retained 50% of the time. A smaller σ helps distinguish between the two hypotheses, and z_i is retained more frequently.

The contaminated data subset is selected by about half of the indicator variables. A joint test of equal coefficients would reveal that these could be replaced by a step dummy, increasing power, or SIS could be applied initally. The irrelevant indicators are retained at a lower probability than α, so 'overfitting' is not a concern. For the linear DGP, the quadratic function is almost always excluded, and retention of the indicators matches that of the non-linear DGP, so the properties of
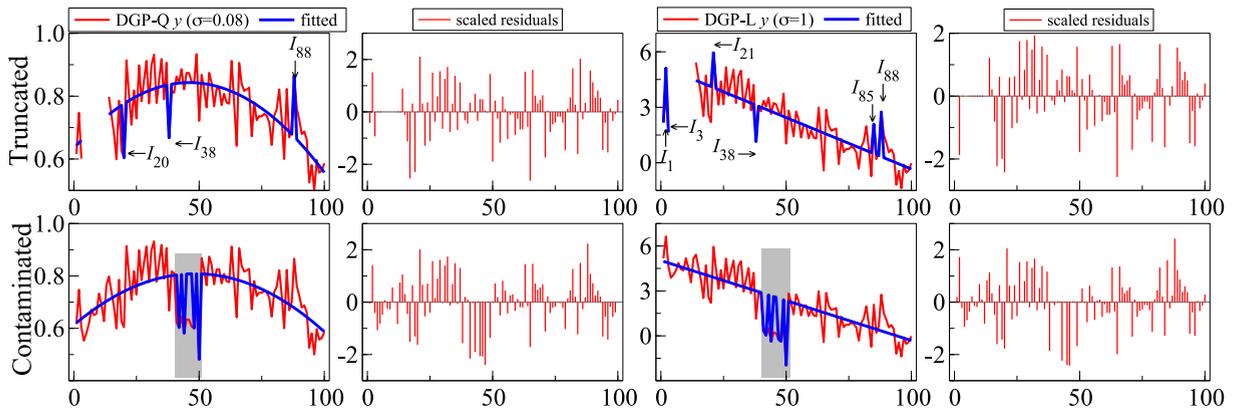
Fig. 4. Top row records one truncated experiment and bottom row a contaminated experiment. *Autometrics* model fit from IIS with the quadratic function when the DGP is non-linear with no outliers (2 left panels) or linear with outliers (2 right panels). Residuals scaled by estimated equation standard error.

IIS do not depend on the functional form of the DGP. Overall, the results show that jointly selecting impulse indicators and non-linear functions enables discrimination between these hypotheses. The costs of testing for both forms of specification are low, particularly with smaller noise.

5. Empirical applications of robust model discovery

Our strategy for robust model discovery seeks to jointly tackle all forms of mis-specification. To illustrate its application, we re-analyse two empirical datasets from previous studies that require robustness in various directions for valid inference and interpretation. The applications show how the approach can isolate sub-sample effects (Section 5.1), and can also avoid 'local instabilities' to which some robust procedures can be vulnerable (Section 5.2).

5.1. Re-analyzing the Boston Housing Market data

These data are originally from Harrison and Rubinfeld (1978), also used by Belsley et al. (1980) and most recently by Peña (2019). The data consist of 506 observations on 14 variables collected by the U.S. Census Service in 1970 on housing in the Boston metropolitan area. Belsley et al. (1980) note that observations 357–488 correspond to Boston, whereas the rest correspond to the suburbs, but a more detailed ordering is not clear. Peña (2019) finds very different regression estimates in those sub-samples. The dataset has also been extensively used as a benchmark for various robust modeling and machine learning algorithms, but no clear consensus has emerged on the underlying data generating process. Hence, it is a useful cross-section dataset for applying our methods to see if they improve on previous empirical results.

As a baseline equation, (B1) records the regression of the log of the median value of owner-occupied homes, denoted *LmedVal*, on the 13 regressors listed in Appendix B, using OxMetrics 8.20 (see Doornik and Hendry, 2018):

$$\widehat{LmedVal} = \underset{(0.20)}{4.1} - \underset{(0.001)}{0.010} Crime + \underset{(0.055)}{0.117} \textbf{Zone} + \underset{(0.002)}{0.002} \textbf{Industry} + \underset{(0.034)}{0.101} Charles$$

$$- \underset{(0.153)}{0.778} NOx + \underset{(0.017)}{0.091} Rooms + \underset{(0.0005)}{0.0002} \textbf{Age} - \underset{(0.008)}{0.049} Distance - \underset{(0.015)}{0.063} Tax \tag{B1}$$

$$+ \underset{(0.003)}{0.014} Radial - \underset{(0.005)}{0.038} PTratio + \underset{(0.011)}{0.041} BlkPop - \underset{(0.002)}{0.029} LowStat,$$

$$\widehat{\sigma} = 0.190, \text{R}^2 = 0.79, \ \text{F}_{\text{Het}}(25, 480) = 7.24^{**}, \ \chi^2_{\text{nd}}(2) = 92.5^{**}, \ \text{and} \ \text{F}_{\text{Reset}}(2, 490) = 15.4^{**}.$$

Estimated coefficient standard errors are shown in parentheses below estimated coefficients, $\widehat{\sigma}$ is the estimated residual standard deviation, R^2 is the coefficient of multiple correlation, F_{Het} is a test for residual heteroskedasticity (see White, 1980), $\chi^2_{\text{nd}}(2)$ is a test for Normality (see Doornik and Hansen, 2008), and F_{Reset} is the RESET test (see Ramsey, 1969). All 3 mis-specification tests strongly reject, and three of the regressors, shown in bold, would not be judged significant at 1%, which is a reasonable significance level for 506 observations.

The top row in Fig. 5 shows respectively the scatter plot of actual against fitted values, the residuals, scaled to unit variance, and the QQ plot, confirming a serious mismatch of model and data.

Once a specification is rejected, how to 'fix it' is unclear as the apparent problem may be due to other mis-specifications (see e.g., Mizon, 1995): going from simple to general has many possible paths, and stopping at the first unrejected equation
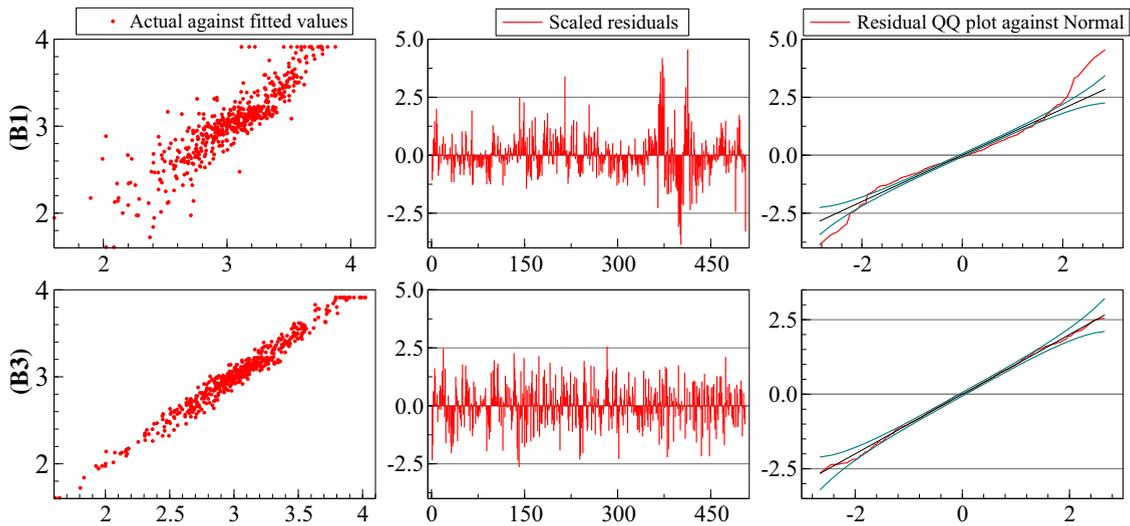
Fig. 5. Graphical results for the two models of Boston house prices. In rows: (B1) and (B3). In columns: scatter plots of actual against fitted values, residuals, and QQ plots against the Normal distribution. Residuals scaled by $\widehat{\sigma}$.

is not a good rule (see Anderson, 1962). For example, to remove the outliers, applying IIS with *Autometrics* at $\alpha = 0.001$ to (B1), retaining all its regressors, found 58 outliers with $\widehat{\sigma} = 0.11$. The scaled residuals showed no further outliers, but revealed blocks of zeroes, mainly occurring over the Boston subsample. However, the three mis-specification tests still rejected, so an investigator might try combining IIS and SIS to also capture the steps.

SIS adds $T-2$ broken intercepts, where each starts with a run of ones, followed by zeros. This is helpful in a time-series settings, because all SIS terms are zero at the end, and only the intercept is extrapolated into the future. Here we have a cross-section, in which there is no natural ordering of the observations. Instead, we use the order in which the data were given. It helps the efficiency of SIS if the data is already stratified, otherwise the impulses have to capture the effect.

Applying IIS+SIS at $\alpha = 0.001$ selected 9 impulse indicators and 32 step indicators from the 1010 candidate variables, with $\widehat{\sigma} = 0.09$. Nevertheless, two mis-specification tests still rejected. To 'solve' that, an investigator might try separating Boston and the suburbs. The *isBoston* indicator is unity for the Boston subsample, and the *Bos* prefix denotes the interaction of that variable with *isBoston*. There is no interaction for *Zone*,*Industry*, *Radial*, *Tax*, and *PTratio* because they are constant within Boston. Adding these additional Boston variables to (B1), and selecting at 5% but without indicator saturation ((B2) but not reported here), still leads to all three mis-specification tests rejecting.

Alternatively, in a general to specific approach, the outcome when selecting over all variables with IIS+SIS is recorded in (B3). With the overall and Boston intercepts retained, selection (at 1%) commences from 506 impulse indicators, 504 step indicators, and 21 free regressors, so 1031 candidates in total. The initial block search for more variables than observations reduced this to 113 candidates, and after adding back all the regressors, the final selection retained 81, leading to (indicators not reported; closely similar results were found using a target size of 0.1% as most selected coefficients exceed 3 times their estimated standard errors):

$$\widehat{LmedVal} = \underset{(0.12)}{1.39} - \underset{(0.012)}{0.046Crime} + \underset{(0.025)}{0.071Zone} + \underset{(0.009)}{0.267Rooms} - \underset{(0.0002)}{0.0023Age} - \underset{(0.004)}{0.033Distance}$$

$$+ \underset{(0.003)}{0.012Radial} - \underset{(0.007)}{0.038Tax} - \underset{(0.003)}{0.018PTratio} + \underset{(0.007)}{0.069BlkPop} - \underset{(0.0012)}{0.0083LowStat} \tag{B3}$$

$$+ \underset{(0.012)}{0.043BosCrime} - \underset{(0.176)}{0.724BosNOx} - \underset{(0.016)}{0.227BosRooms} + \underset{(0.17)}{1.79isBoston},$$

$$\widehat{\sigma} = 0.074, \text{R}_d^2 = 0.88, \text{ F}_{\text{Het}}(66, 413) = 1.32, \chi^2_{\text{nd}}(2) = 3.71, \text{ and F}_{\text{Reset}}(2, 424) = 1.64.$$

The entire procedure took just over two minutes, no mis-specification tests reject, so no outliers, or hidden shifts, remain. Hence inference should be more reliable and the residual standard deviation is much smaller. The bottom row of Fig. 5 confirms the improvements in the residual distribution. *Zone* and *Age* are now significant, the latter with the opposite sign to (B1).

The overall effects for the Boston subsample can be calculated from the combination of the whole sample and subsample coefficients: the effects of crime and house size on house prices are essentially zero in the city, whereas *Age*, *Distance* and *BlkPop* are the same as the whole sample. The large positive coefficient for *isBoston* reflects higher mean house prices in the city not accounted for by other subsample regressors, whose separate effects cannot be disentangled as they are constant
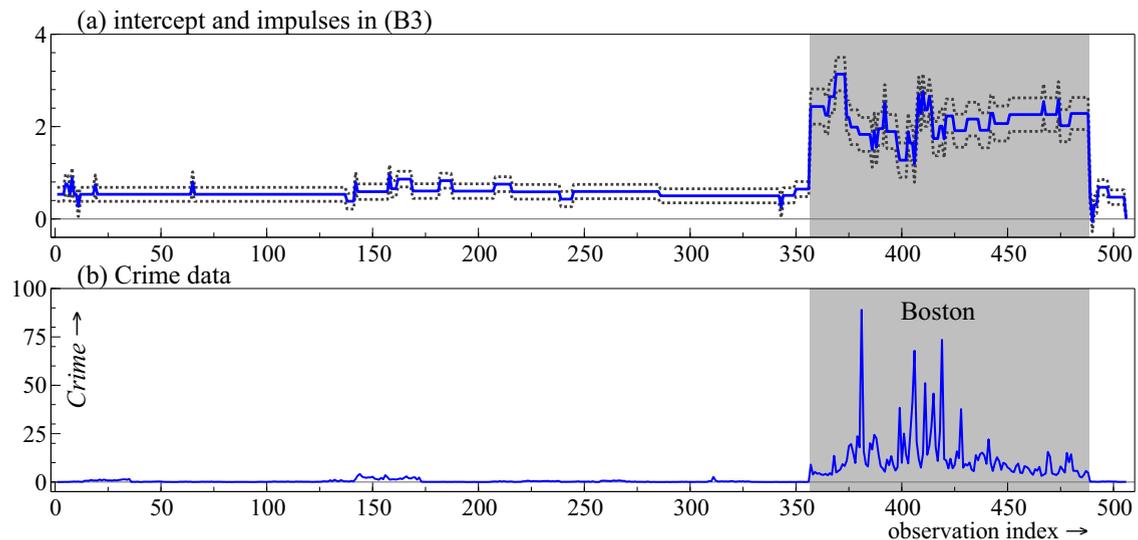
Fig. 6. (a) Contribution to the intercept for each observation from the constant term, impulse and step indicators. Dotted line is ± two standard errors. (b) Graph of *Crime* in Boston and suburbs. Boston is shaded in both graphs.

over the subsample. More than half of the indicators fall in the Boston area. The presence of so many step indicators is likely to reflect geographical clustering or other unmeasured aspects. Fig. 6(a) shows the combined magnitudes of the intercept and impulses for each observation.

The cancellation of the influences of crime between the overall effect and Boston's is a surprise as Fig. 6(b) shows it is high in the city and much lower elsewhere. To test the validity of conditioning on *BosCrime*, as there is a possibility that more valuable housing may attract that crime, we apply the method described in Section 2.4 to test for super exogeneity. First we modeled *BosCrime* by the Boston regressors and IIS+SIS at 1%, finding thirteen indicators in the marginal model that are not in (B3). Adding them to the conditional model (B3), yielded the insignificant outcome $F_{valcond}(13, 413) = 1.5$. Thus, the largest changes in *BosCrime* have the same impacts as all other changes, so *BosCrime* is super-exogenous and a valid conditioning variable.

This empirical application demonstrates how SIS can reveal important differences in subsamples of data that must be modeled to obtain a well-specified and economically interpretable equation that is robust to the different data properties of subsamples of the cross section.

5.2. Avoiding fragility of robust methods: Engine knock data

While scoring high on robustness, both LTS and LMS have a certain fragility or 'local instability', whereby a small change to one value recorded for a centrally-located observation can cause large changes in the estimates. This phenomenon can occur when two 'half-samples' correspond to different 'regimes', but nevertheless have approximately the same criterion-function value, so small changes to some observations can make the LMS or LTS solution jump from the estimates of one half sample to the other. Hettmansperger and Sheather (1992) note this issue with LMS when they accidentally made a mistake in transcribing data that seemed quite innocuous. Doornik (2016) gives an example for LTS.

We use robust model selection, rather then just applying a robust method to a 'known' model, to gain deeper insight into the empirical question. By discovering more about the underlying data, we can explain and avoid the noted fragility in the engine knock model.

Hettmansperger and Sheather (1992) took the data from Mason et al. (1989, p. 529), and aimed to predict 'engine knock' (*knock*) from a constant and four regressors called '*spark* timing', '*Air*/fuel ratio', '*intake* temperature', and '*exhaust* temperature', where italic labels are used below. Engine knock was a problem in combustion engines when lead was banned from gasoline and before new additives were found. There are 16 observations on each. However, on inputting the data, they had inadvertently entered the second observation for *Air* as 15.1 (denoted *WAir*, for wrong air) rather than the correct 14.1, and found very different estimates from those initially reported for LMS, as shown in Table 6.

Using *Air*, LTS drops observations (2, **3**, **5**, **7**, 9, **12**, 13, 15), whereas, using the miscoded *WAir*, LTS drops (**3**, 4, **5**, **7**, 11, **12**, 14, 16), where bold denotes deletions in common. In the first case, LTS drops the mismeasured observation 2, but with wrong air is kept. LMS and LTS are close within each measurement of '*Air*', but both differ considerably between measures, so lie on different planes. This difference will persist if the estimates are used as a starting point for reselection of observations.

Table 6

LMS estimates from Hettmansperger and Sheather. Our LTS(0.5) estimates with standard errors in parentheses.

	Correct air			Wrong air		
	LMS	LTS(0.5)		LMS	LTS(0.5)	
Air	2.9	3.1	(0.13)	1.2	1.1	(0.23)
intake	0.56	0.43	(0.05)	1.5	1.6	(0.05)
spark	0.21	0.06	(0.11)	4.6	3.9	(0.36)
exhaust	-0.01	-0.005	(0.002)	0.07	0.05	(0.01)
Constant	30.1	30.9	(3.3)	-86.5	-68.7	(9.2)
$\widehat{\sigma}$		0.12			0.21	
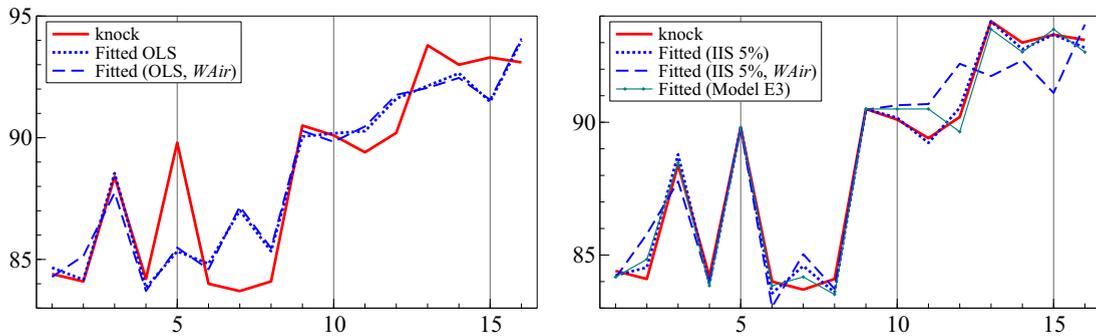
Fig. 7. Actual and fitted values from OLS (left), and IIS (right) (E1) and (E2) for the two measures of '*Air*'.

We start by applying IIS selection at 5% to the initial model with all variables, first using correct air, then wrong air. The constant is retained, so all estimated models have an intercept. For correct air, IIS finds 4 outliers, retaining *Air* and *intake*:

$$\widehat{knock}_i = \underset{(0.3)}{3.2}Air_i + \underset{(0.09)}{0.35}intake_i + \underset{(0.5)}{6.5}\mathbf{1}_{\{5\}} + \underset{(0.5)}{1.6}\mathbf{1}_{\{9\}} + \underset{(0.5)}{3.0}\mathbf{1}_{\{13\}} + \underset{(0.5)}{3.4}\mathbf{1}_{\{15\}} + \underset{(2.4)}{28.4},$$

(E1)

$\widehat{\sigma} = 0.5$, $R_d^2 = 0.99$, $F_{Het}(4, 7) = 0.6$, $\chi_{nd}^2(2) = 2.8$, $F_{Reset}(2, 7) = 5.0^*$, and $F_{nl}(2, 7) = 0.3$.

The indicator $\mathbf{1}_{\{5\}}$ reveals that observation 5 is selected as an outlier. See Section 5.1 for details of regression output and tests, with F_{nl} a test for omitted non-linearity (see Castle and Hendry, 2010). None of the mis-specification tests is significant at 5%, except for RESET which has a p-value of 4.5%. For 20 candidates, we expect to retain one by chance, which could be $\mathbf{1}_{\{9\}}$, as that disappears when running IIS at 2.5%.

Using the incorrect measure *WAir* yields:

$$\widehat{knock}_i = \underset{(0.6)}{2.1}WAir_i + \underset{(0.2)}{0.9}intake_i + \underset{(1.5)}{6.3}\mathbf{1}_{\{5\}} + \underset{(6.8)}{27.3}.$$

(E2)

$\widehat{\sigma} = 1.4$, $R_d^2 = 0.90$, $F_{Het}(4, 10) = 1.4$, $\chi_{nd}^2(2) = 0.2$, $F_{Reset}(2, 10) = 1.6$, and $F_{nl}(6, 6) = 1.5$.

The two regression estimates are now similar, and both detect that observation 5 is an outlier.

Fig. 7 records actual and fitted values by OLS and IIS for the two measures of '*Air*' showing their closeness for OLS. In the IIS case, the fitted values are close in the first half of the sample, but different in the second half. It is obvious visually that observation 5 is an outlier in OLS. In all cases, the fit for observation 9 is (almost) exact, but (E1) achieves that through $\mathbf{1}_{\{9\}}$.

5.2.1. Unweaving the findings

What actually caused the instability in LMS and LTS, and has IIS resolved it? The upper 3D plot in Fig. 8(a) of *knock* against *spark* & *intake* shows that the data split into two 'regimes': the first 8 observations on *knock* are less than 90, the last 8 greater (inside the ellipse), and are associated with wide and narrow spreads respectively. However, that split does not coincide with the observations selected by either LMS or LTS. The lower 3D graph in Fig. 8(b) of *knock* against *exhaust* and *Air*, with the incorrect second observation for *WAir* also shown, is surprisingly revealing given the irrelevance of *exhaust*. Observation 5 is a marked outlier, and the misrecorded observation 2 on *Air* is also now seen to be an outlier within the first 'regime' despite the apparent small magnitude of the mismeasurement.

The two 'regimes' visible in the upper graph (Fig. 8a) can entail a sudden switch between which subset is selected when one observation is moved between them. However, facing such a knife-edge result, it is somewhat arbitrary to decide that the selected set must be the 'good set', and the rest the 'bad set', even when the former has a smaller variance. Here, we consider both sets by creating a step indicator $S_{\{i \geq 9\}}$ equal to unity for $i \geq 9$ and zero otherwise, corresponding to the observations in the ellipse in Fig. 8a. $S_{\{i \geq 9\}}$ is interacted with the four regressors other than *exhaust*. The general model then
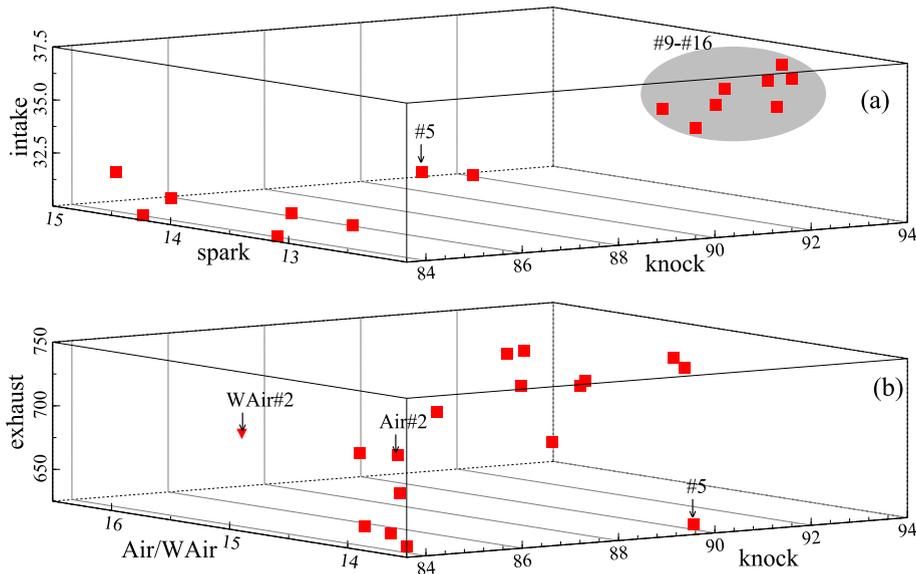
Fig. 8. 3D plot of the data (a) *knock* against *spark* & *intake*; (b) *knock* against *exhaust* & *Air, WAir*.

Table 7
LTS(0.5) estimates without *exhaust*.

	Correct air		Wrong air	
air	2.7	(0.30)	3.2	(0.31)
intake	0.61	(0.11)	0.38	(0.13)
spark	0.49	(0.27)	0.22	(0.25)
Constant	21.1	(6.5)	25.4	(5.3)
$\widehat{\sigma}$	0.34		0.35	

comprises the retained intercept, four regressors, the four interactions and sixteen impulse indicators from IIS. Selecting at 2.5% includes *Air* and $AirS_{\{i \geq 9\}}$ with coefficients of almost equal magnitude but opposite sign.

Imposing this simplification delivers:

$$\widehat{knock}_i = \underset{(0.3)}{3.3}\,Air_i S_{\{i \leq 8\}} + \underset{(0.4)}{4.3}\,spark_i S_{\{i \geq 9\}} + \underset{(0.6)}{6.6}\,\mathbf{1}_{\{5\}} + \underset{(4.5)}{38.2}, \tag{E3}$$

$$\widehat{\sigma} = 0.54, \ \ R_d^2 = 0.98, \ \ F_{\text{Het}}(4, 10) = 0.7, \ \ \chi^2_{\text{nd}}(2) = 2.9, \ \ F_{\text{Reset}}(2, 10) = 1.6, \ \text{and} \ F_{\text{nl}}(6, 6) = 2.4.$$

This equation describes the whole data set, but where *Air* only matters in the first half and *spark* only in the second. Apart from the outlier $\mathbf{1}_{\{5\}}$, the intercept is constant across both halves. There is a serious non-constancy in the impacts of *spark* and *Air* on *knock* matching the two regimes visible in Fig. 8(a). As other factors are known to influence engine knock (carbon deposits in cylinders, cleanliness of spark plugs, etc.), missing information may account for this finding.

Returning to the robust estimators LMS and LTS now also applied without *exhaust* reveals that the difference between the two measures of *Air* is no longer very large, as Table 7 records. Consequently, it seems that LTS can be affected by including empirically irrelevant variables in the model (problem C2), suggesting that there are real benefits from variable selection jointly with tackling outliers. For estimates with *Air*, LTS(0.5) now dropped (4, **5**, **7**, **9**, **12**, **13**, 14, **15**) and with *WAir*, dropped (2, 3, **5**, **7**, **9**, **12**, **13**, **15**) so six of the 8 dropped observations are in common (in bold), drawn more from the second half. Moreover, (E3) suggests there is no constant-parameter relation to be found. Indeed, from Table 6, the LTS estimates using *Air* had a coefficient for that variable close to that of $Air_i S_{\{i \leq 8\}}$, but an insignificant coefficient for *spark*, whereas with *WAir*, its own coefficient was small but that for *spark* was close to that of $spark_i S_{\{i \geq 9\}}$. Thus, the mismeasurement happened to precipitate a switch between the two regimes. Consequently, a method like *Autometrics* with IIS may be preferred because it selects over both observations and regressors and so provides some protection against knife-edge and non-constancy situations.

5.2.2. Valid conditioning

We next test for valid conditioning using the method described in Section 2.4. Modeling *spark* by *Air*, *exhaust*, *intake* and a retained constant (but not the dependent variable *knock* in the models above) using IIS at 5% yields significant values for *exhaust*, $\mathbf{1}_{\{4\}}$ and $\mathbf{1}_{\{14\}}$, so these two impulse indicators that are absent from (E3) can be used to test the validity of conditioning. Moreover, replacing *Air* by *WAir* in these regressions for *spark* yields the same two impulse indicators.

Adding $\mathbf{1}_{\{4\}}$ and $\mathbf{1}_{\{14\}}$ to (E3) yields $F_{valcond}(2, 10) = 0.46$, which does not reject either the validity of conditioning or the exclusion of data on *spark* for the first half of the sample. However, replacing *Air* by *WAir* in (E3) leads to rejection on F_{Het} and F_{Reset}. Retaining all the regressors without selection when redoing IIS at 1% yields:

$$\widehat{knock}_i = \underset{(0.23)}{3.2} \, WAir_i S_{\{i \leq 8\}} + \underset{(0.26)}{4.1} \, spark_i S_{\{i \geq 9\}} + \underset{(0.42)}{6.4} \, \mathbf{1}_{\{5\}} + \underset{(3.2)}{40.2} - \underset{(0.47)}{4.0} \, \mathbf{1}_{\{2\}} - \underset{(0.41)}{1.3} \, \mathbf{1}_{\{11\}}, \tag{E4}$$

$$\widehat{\sigma} = 0.37, \ R_d^2 = 0.99, \ F_{Het}(4, 8) = 0.37, \ \chi_{nd}^2(2) = 2.6, \ \text{and} \ F_{Reset}(2, 8) = 0.7.$$

As can be seen, the coefficients in common between (E3) and (E4) are closely similar, and $\mathbf{1}_{\{2\}}$ reveals the measurement error! Dropping $\mathbf{1}_{\{11\}}$ as being adventitiously significant makes the match even closer, including for $\widehat{\sigma} = 0.50$. Adding $\mathbf{1}_{\{4\}}$ and $\mathbf{1}_{\{14\}}$ to (E4) delivers $F_{valcond}(2, 8) = 0.33$, and also does not reject super exogeneity once $\mathbf{1}_{\{11\}}$ is omitted.

In this setting where LMS and LTS delivered very different estimates when a 'small' mismeasurement of one observation on one variable occurred, model selection using IIS detected the important outlier, the removal of which helped stabilize the results. It led us to notice that the mismeasurement also created a potential outlier, and that coefficients were not constant over the sample. To discover key features of the underlying process, there are advantages in selecting empirically significant regressors jointly with removing outliers and tackling potential non-constancies. Indeed, the LTS results were more similar between the correct and mismeasured variable once an apparently irrelevant regressor was eliminated.

5.2.3. Lasso estimation

Because the engine knock data is cross section with possible outliers, we could also consider using the adaptive Lasso (adaLasso) of Zhou (2006) for model selection. Here we select using the Bayesian information criterion, and always estimate the final model by OLS, so the Lasso is just a selection device.

To start, adaLasso selects *Air* and *intake* from the correct set of variables, but just *intake* when using *WAir*. So the coding error gives different models.

To allow for outliers, and using correct *Air*, we could saturate with all possible impulses, just like IIS. In that case the procedure does not know when to stop, selecting *Air* and *intake* together with impulses for observations 1,2,4,5,7,9,13,14,15,16. Appendix A confirms this problem in simulation experiments. Reducing this set with *Autometrics* at 2.5% finds the following model:

$$\widehat{knock}_i = \underset{(0.28)}{3.2} \, Air_i + \underset{(0.09)}{0.35} intake_i + \underset{(0.52)}{6.5} \, \mathbf{1}_{\{5\}} - \underset{(0.54)}{1.6} \, \mathbf{1}_{\{9\}} - \underset{(0.51)}{3.0} \, \mathbf{1}_{\{13\}} - \underset{(0.52)}{3.4} \, \mathbf{1}_{\{15\}} + \underset{(2.4)}{28.4}, \tag{E5}$$

$$\widehat{\sigma} = 0.47, \ R_d^2 = 0.99, \ F_{Het}(4, 8) = 0.63, \ \chi_{nd}^2(2) = 2.8, \ \text{and} \ F_{Reset}(2, 8) = 5.9^*.$$

This model is an alternative candidate to (E3). An encompassing test cannot distinguish between them, but (E3) is a more concise description of the data.

Appendix A suggests that the *autoLasso* provides a better approach than adaLasso estimation of models that have more variables than observations. The first step applies the block learning algorithm (as used with *Autometrics* in Section 3) with adaLasso as the selection device. The final stage is *Autometrics*, or another adaLasso step for the final model selection. Assuming that inspection also led to the discovery of the two regimes, we can apply *autoLasso* to the same general model that yielded (E3). The block learning yields *Air*, $AirS_{\{i \geq 9\}}$, $spark_i S_{\{i \geq 9\}}$ and impulses for (2,5,10,11). Then selection at 1% and combining the air variables gives (E3). In this particular case, two different approaches lead to the same result, provided the crucial discovery of the two different regimes is made.

6. Conclusion

There are various concepts of 'robustness' within econometrics and statistics. We propose a general notion of robustness to sustain model discovery which requires that model selection methods have acceptable performance facing possible outliers and shifts, leading to an incorrect distributional shape and non-constancy, omitted variables, and non-linearity, as well as mis-specified dynamics and non-stationarity in time series, plus checking the validity of exogeneity assumptions. This extends the notion of robustness from an approach to delivering good statistical properties under just one form of potential mis-specification for a pre-specified model to a more general sense of robust model discovery. As a consequence, to tackle all these forms of potential non-robustness jointly, general initial formulations combined with efficient selection methods are needed, while at the same time retaining relevant subject matter insights.

The paper outlines our approach to robust model discovery for regression equations, and the role of indicator saturation estimators therein as designed to match the likely problem. Two empirical examples demonstrate how the approach delivers robust selection and, hence, viable inference. Robustness can only be achieved if all modeling decisions are implemented jointly. The definition of an outlier requires a congruent, well-specified model. If a discrepant observation in a regression context is due to mis-specification of the regression, then the interpretation of the outliers is different to if there are contaminated observations in the DGP, which can only be detected if the model is well-specified. Distinguishing the model from the DGP allows for robust inference on the selected model when it is well-specified. Automatic model selection which also tests for congruence and encompassing in the reduction procedure will satisfy this requirement given a congruent initial specification, which a large GUM should help ensure.

Declaration of Competing interest

Acknowledgements

Appendix A. Simulation evaluations

A1. Experiments with independent regressors

We use Monte Carlo experiments to evaluate how well the algorithm described in Section 3 is able to discover the data generation process (DGP), first when the regressors are independent indicators:

$$\text{DGP:L } y_t^L = \mu + \gamma \left(\mathbf{1}_{\{\tau T+1\}} + ... + \mathbf{1}_{\{T\}} \right) + u_t, \tag{A.1}$$

$$\text{DGP:S } y_t^S = \mu + \gamma \left(\mathbf{1}_{\{1\}} + \mathbf{1}_{\{1+S\}} + \mathbf{1}_{\{1+2S\}} + ... \right) + u_t, \tag{A.2}$$

$$\text{DGP:Z } y_t^Z = \mu + \beta T^{-1/2} (z_{1t} + ... + z_{12,t}) + u_t, \tag{A.3}$$

or where $u_t, z_t \sim \text{N}[0, 1]$ and independent. As before, $\mathbf{1}_{\{T\}}$ has value one at observation $t = T$ and zero otherwise. Setting $T = 100$ and $\tau = 0.8$ in DGP:L means that twenty percent of the sample is in the break period. Defining $S = \lfloor (1 - \tau)^{-1} \rfloor$ in (A.2) with $T = 100$ and $\tau = 0.8$ shifts the mean in DGP:S by γ at observations $t = 1, 6, 11, ..., 96$. This is 20% of the observations, just as for DGP:L with $\tau = 0.8$. When $\gamma = 0$, the experiments are under the null of no break.

DGP:L corresponds to the type of structural breaks that we may observe in time-series data or from combining different subsamples. DGP:S is less realistic, looking more like neglected seasonality, but is harder for some approaches, because each subsample looks alike.

The initial model for DGP:L and DGP:S consists of the T dummies and a retained intercept, denoted MOD:L. The model for DGP:Z includes all z's and y up to lag m, with the intercept always included:

$$\text{MOD:L } y_t = \psi^R + \psi_1 \mathbf{1}_{\{1\}} + ... + \psi_T \mathbf{1}_{\{T\}} + \varepsilon_t, \tag{A.4}$$

$$\text{MOD:Z } y_t = \psi^R + \sum_{i=1}^{m} \psi_m y_{t-m} + \sum_{i=1}^{12} \sum_{i=0}^{m} \psi_{im} z_{i,t-m} + \varepsilon_t. \tag{A.5}$$

The block search algorithm with learning given in Table 4 is implemented in *Autometrics*. Table A.1 documents the settings for three versions, standard, reduced, and reference. The reference version follows Section 3.1, while the standard version is the *Autometrics* default. Because the latter has $\lambda = 8$, it is more biased away from the empty model, giving it a higher gauge than the other versions through retention of some insignificant variables. In most empirical settings the difference is small, but in a Monte Carlo where the null is empty (or small) it can make some difference.

The following approaches that are feasible in this setting are included in the initial comparison:

1. *Stepwise regression* at significance level p_a;
2. *Lasso* with optimal model selected by SC (Schwarz criterion, the same as Bayesian information criterion, BIC), subject to an upper limit of $T/2$ nonzero coefficients;
3. *IIS algorithm* of Johansen and Nielsen (2009) at significance level p_a;
4. *Backward elimination* applied to blocks, followed by *Autometrics* at significance level p_a;
5. *Proposed learning algorithm*: represented by the 'reference' block search in *Autometrics* at p_a.

Table A.2 records the gauge and potency (non-null retention frequency) for selected values of γ, i.e., the magnitude of the break in standard deviations. Stepwise regression selects the impulses in the break at a high rate when the significance is set to 5%, but at the expense of also including many irrelevant dummies; at lower significance there is no potency. The Lasso does not use a significance level, and termination is based on BIC. The potency is low, except for larger γ, but then the gauge shoots up as well.

IIS and backward elimination have similar results, with the former having best control of the gauge when there is no break, courtesy of its bias correction $\mathbf{\Omega}_\alpha$ in (2). The proposed learning algorithm, as implemented in *Autometrics*, is close

Table A.1

Settings for *Autometrics* versions of learning algorithm (Table 4).

	standard	reduced	reference
λ	$\lambda = 8$	as in Section 3.1	as in Section 3.1
Stage C history	$\widetilde{\mathcal{C}}^B$	$\widetilde{\mathcal{C}}^B$	$\mathcal{C}^B$
Stage C $f(\alpha, 2), f(\alpha, \frac{1}{2})$	always	always	only if $\alpha \leq 0.03$
Diagnostic testing in	B,C,D	B,C,D	C,D
Expansion backtesting	A–D	A–D	A,B
Reduction backtesting	A,B	A,B	A,B
Afterwards	*Autometrics*	*Autometrics* no backtesting	*Autometrics*

Table A.2

DGP:L and DGP:S have a break in mean of magnitude γ in 20% of observations. The estimated model is MOD:L, consisting of a constant and T dummies. $T = 100$ observations, $M = 1000$ replications, $p_a = 0.05, 0.01$.

	DGP:L			DGP:L			DGP:S	
	$\gamma = 0$	$\gamma = 2$	$\gamma = 4$	$\gamma = 0$	$\gamma = 3$	$\gamma = 4$	$\gamma = 4$	$\gamma = 5$
Lasso (BIC, $N_{max} = 50$)								
Gauge %				1.2	0.2	2.3	2.0	14.2
Potency %				—	6.1	16.1	14.9	77.9
	5% target			**1% target**			**1% target**	
Stepwise regression								
Gauge %	15.5	10.5	14.6	1.4	0.1	0.0	0.0	0.1
Potency %	—	51.7	99.1	—	9.5	12.0	12.3	13.2
IIS (Johansen and Nielsen, 2009)								
Gauge %	5.3	3.6	3.3	1.2	0.5	0.7	0.0	0.0
Potency %	—	39.5	96.6	—	49.4	88.1	5.8	4.5
Backward elimination, then *Autometrics*								
Gauge %	7.1	3.4	2.4	1.2	0.1	0.3	0.0	0.0
Potency %	—	43.3	98.1	—	24.9	82.4	7.5	6.8
Proposed learning algorithm (*Autometrics*, 'reference' block search)								
Gauge %	8.9	4.9	8.1	1.5	0.2	0.3	0.4	0.4
Potency %	—	48.1	98.5	—	53.0	81.1	33.8	50.9

Table A.3

DGP:L with a break in last τT observations of magnitude $\gamma = 4$. The estimated model is MOD:L, consisting of a constant and T dummies. $T = 100$ observations, $M = 1000$ replications.

	$\tau = 0.02$	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.02$	$\tau = 0.1$	$\tau = 0.2$
Stepwise regression ($p_a = 0.01$)						
Gauge %	1.4	1.2	0.0			
Potency %	92.4	89.2	12.0			
Lasso (BIC, $N_{max} = 50$)				**Lasso (5-fold CV)**		
Gauge %	0.9	3.9	2.3	85.3	82.1	78.2
Potency %	85.5	79.4	16.1	100.0	100.0	100.0
Autometrics (reference block search, $p_a = 0.01$)						
Gauge %	0.7	0.5	0.3			
Potency %	91.6	87.8	81.1			

to IIS and backward elimination for DGP:L. It is somewhat more overgauged, but the benefit is that it has better potency in DGP:S.

Table A.3 provides additional insight by varying the length of the break. DGP:L is used with a break in mean of magnitude four, but the duration of the break is for 2, 10, and 20 observations respectively. The Lasso and stepwise regression both have the gauge and potency changing as the length increases. The block learning algorithm is less sensitive to this. Because structural breaks in time series often persist for extended periods, this is a useful practical aspect of the algorithm. The termination decision for Lasso is problematic in this design.

Table A.4 looks at short data settings with additional variables in the model: DGP:Z with MOD:Z. The first set of experiments has $\beta = 0$, so the empty model is the correct model. The second model has $\beta = 10$, corresponding to an expected t-value of 10. In that case, the significant regressors are so significant (except at $p_a = 0.001$) that their presence should make little difference. We see this for the proposed algorithm.

Table A.4

Gauge of the *Autometrics* algorithm. $T = 100$ observations, $M = 1000$ replications (using $M = 10\,000$ for $p_a = 0.001$). DGP:Z with MOD:Z.

T	m	$\alpha = 0.05$	0.025	0.01	0.001	$\alpha = 0.05$	0.025	0.01	0.001
Autometrics reference				$\beta = 0$		$\beta = 10$			
40	4	0.083	0.034	0.015	0.0013	0.069	0.011	0.002	0.0000
100	8	0.046	0.022	0.010	0.0006	0.052	0.020	0.008	0.0007
250	20	0.034	0.017	0.008	0.0004	0.037	0.016	0.006	0.0007

Table A.5

$T = 139$, $M = 1000$, $p_a = 0.01$, 3 relevant variables. HP7 and HP8 have 37 irrelevant variables, the big versions have 141.

	HP7	HP8	HP7big	HP8big	HP7	HP8	HP7big	HP8big
Lasso (BIC, $N_{\max} = 50$)					Lasso (10-fold CV)			
Gauge %	19.5	35.1	2.9	2.0	71.6	94.3	86.8	88.9
Potency %	94.4	86.3	71.8	58.0	99.7	100.0	98.9	97.8
adaLasso (BIC, $N_{\max} = 50$)					adaLasso (10-fold CV)			
Gauge %	4.5	3.3	2.9	6.2	2.9	3.9	88.3	76.3
Potency %	99.7	100.0	72.2	98.9	95.6	91.1	95.6	93.4
Autometrics, reference ($p_a = 0.01$)					standard ($p_a = 0.01$)			
Gauge %	1.6	1.6	0.9	0.9	1.8	1.6	1.3	2.2
Potency %	99.2	100.0	99.6	100.0	99.2	100.0	99.5	100.0
blockLasso (BIC, $N_{\max} = 50$)					autoLasso (BIC, $p = 0.01$)			
Gauge %	3.5	3.0	4.1	3.1	2.0	1.8	2.8	2.2
Potency %	100.0	100.0	99.6	100.0	99.9	100.0	99.6	100.0

A2. Experiments with correlated regressors

Further experiments are based on models 7 and 8 from Hoover and Perez (1999), denoted HP7 and HP8 respectively. The regressors are quarterly macro-economic variables, where unit roots are removed by differencing. The DGPs for these experiments are:

$$\text{HP7: } y_{7,t} = 0.75 y_{7,t-1} + 1.33 x_{11,t} - 0.9975 x_{11,t-1} + 6.44 u_t, \, u_t \sim N[0, 1],$$
$$\text{HP8: } y_{8,t} = 0.75 y_{8,t-1} - 0.046 x_{3,t} + 0.0345 x_{3,t-1} + 0.073 u_t, \, u_t \sim N[0, 1].$$

HP7 has $R^2 = 0.58$, and HP8 has $R^2 = 0.93$; all coefficients have very high t-values (in excess of 8). The models have 37 irrelevant macro-economic variables.

We create 'big' versions that have more variables than observations in the initial model by adding 10 independent N[0,1] regressors $z_1, ..., z_{10}$ up to lag 4 to the initial model, making 145 regressors in total. Only 3 matter, and the constant is always included. These experiments are labeled HP7big and HP8big in Table A.5.

The tables now include the adaptive Lasso (adaLasso, Zhou, 2006), where the coefficients in the L1 penalty are scaled down by the OLS estimates. This is undefined for short data: when there are more than $T/2$ regressors, we use coefficients from a ridge regression that implies $T/2$ coefficients.

A3. Lasso for short data

The Lasso both shrinks and selects, but is used here only for selection, with the final model estimated by OLS. Because the Lasso is based on a forward search, it can be applied to short data. However, it has not performed so well in these settings, to a large extent because it is difficult to know when to stop: whether using cross validation or an information criterion, there are usually several minima. Very large models are often selected, and it defeats the purpose of machine learning to have to select the model by visual inspection of a plot of the criterion.

Modeling with more variables than observations is of practical relevance, so our block algorithm could be useful here. One example is the adaLasso, where the coefficients in the L1 penalty are scaled down by the OLS estimates. With too many variables, there are no OLS estimates. In that case, one could use ridge estimates, but instead we propose to run selection in blocks, collecting a set of regressors for the final run. The final run could be another adaLasso, or *Autometrics* if better control of the gauge is required. The leads to two new versions of adaLasso:

blockLasso Uses the block search algorithm with learning from Table 4, with adaLasso(BIC) as the selection device for the algorithm, as well as for the final selection step.

autoLasso Uses the block search algorithm with learning from Table 4, with adaLasso(BIC) as the selection device for the algorithm. *Autometrics* at p_a is used to select the final model.
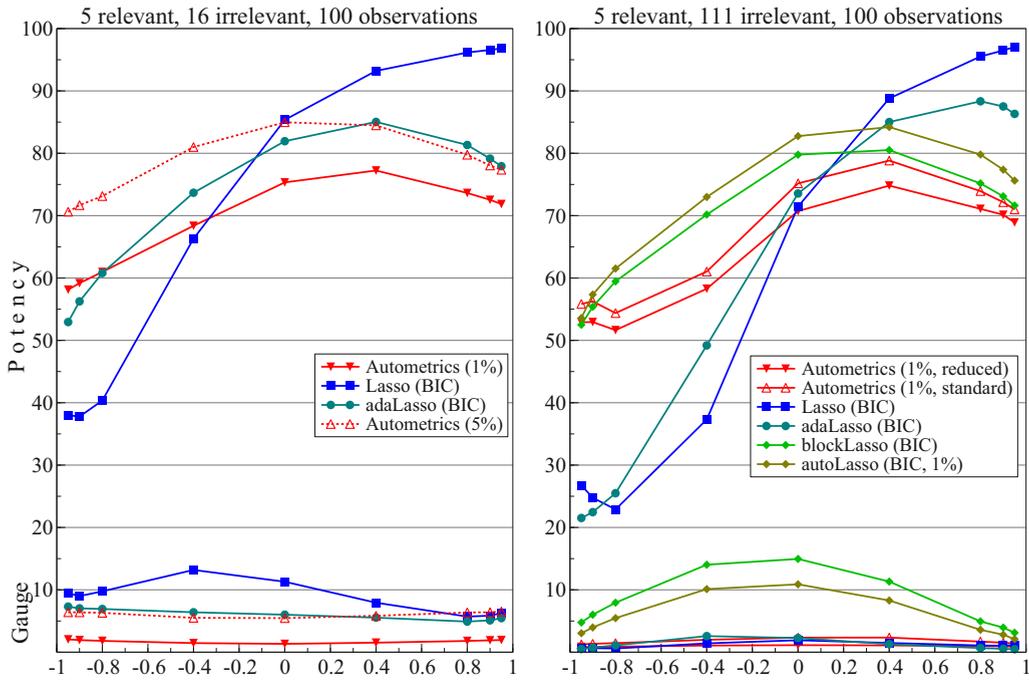
Fig. A.9. JEDC experiment, standard version on left, large version on right.

In both cases the ordering in each block is based on significance in the OLS model using the selected variables. Table A.5 shows that this improves the standard adaLasso: the gauge and potency are now less affected by the addition of the many irrelevant variables.

As a final experiment, we use the simulation settings discussed in Hendry and Doornik (2014), §17.2.2, denoted JEDC:

$$\text{DGP:J } y_t = \mu + \sum_{i=1}^{5} \beta_i T^{-1/2} z_{it} + u_t, \quad u_t \sim N[0, 1], \quad (z_{1t}, ..., z_{Nt}) \sim N[\mathbf{0}, \mathbf{C}_z], \tag{A.6}$$

$$\text{MOD:J } y_t = \psi^R + \sum_{i=1}^{m} \psi_i y_{t-i} + \sum_{j=1}^{N} \sum_{i=0}^{m} \psi_{j,i} z_{j,t-i} + \varepsilon_t, \quad t = 1, ..., 100, \tag{A.7}$$

where $\mathbf{C}_z = (c_{i,j}) = \rho^{|i-j|}$ and $(\beta_1, ..., \beta_5) = (8, 4, 6, 3, 2)$. The standard experiment has lag length of one ($m = 1$) and eleven irrelevant variables ($N = 10$, so irrelevant are $y_{t-1}, z_{6t}, ..., z_{10,t}, z_{1,t-1}, ..., z_{10,t-1}$). The large experiment has $m = 8, N = 12$, adding 111 irrelevant variables when $T = 100$.

Fig. A.9 shows the result for values of $|\rho| = (0, 0.4, 0.8, 0.9, 0.95)$. The left panel is the standard version ($m = 1, N = 10$), while on the right is the large version ($m = 8, N = 12$). Both gauge and potency are plotted, with gauge always at the bottom. For the standard version we see that the gauge of *Autometrics* is close to the target size. The adaLasso gauge is close to that of *Autometrics* at 5%, but then the latter largely dominates in terms of potency. Lasso struggles with negative correlations, possibly for the same reason why stepwise regression fails: negatively correlated variables need to enter jointly as they may not matter much individually.

The large case shows that the potency of *Autometrics* is little affected by the many irrelevant variables. The adaLasso is improved by the block search algorithm in the form of blockLasso, at the expense of the gauge—but performance is now more similar to the small case. Adding an *Autometrics* step at the end, as in autoLasso, reduces the gauge, while at the same time increasing potency. This is only possible if the block search retains some relevant variables that the final adaLasso selection removed.

Both the blockLasso and the autoLasso improve the Lasso results when using IIS, but they remain quite strongly over-gauged.

Appendix B. Data definitions for the Boston Housing example

Boston Housing Market data available at lib.stat.cmu.edu/datasets/boston with some transformations used in the table on pp. 244–261 of Belsley et al. (1980). $N = 14$ variables and the sample size is 506 observations. To standardize estimated coefficient values, Zone, Tax and BlkPop (and the corresponding Boston subsample variables) were all rescaled by 100.

Table B.6
Boston Housing Market data

Name	Variable
LmedVal	Log of the median value of owner-occupied homes in $1000s (regressand)
Crime	per capita crime rate by town
Zone	proportion of residential land zoned for lots over 25,000 sq.ft.
Industry	proportion of non-retail business acres per town
Charles	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOx	nitric oxides concentration (parts per 10 million)
Rooms	average number of rooms per dwelling
Age	proportion of owner-occupied units built prior to 1940
Distance	weighted distances to five Boston employment centres
Radial	index of accessibility to radial highways
Tax	full-value property-tax rate per $10,000
PTratio	pupil-teacher ratio by town
BlkPop	$1000(Bk - 0.63)^2$ where Bk is the proportion of black persons by town
LowStat	% lower status of the population.

References

Anderson, T.W., 1962. The choice of the degree of a polynomial regression as a multiple-decision problem. Annals of Mathematical Statistics 33, 255–265.

Belsley, D.A., Kuh, E., Welsh, R.E., 1980. Regression diagnostics. Identifying influential data and sources of collinearity. In: Wiley series in probability and mathematical statistics. John Wiley, New York.

Berenguer-Rico, V., Johansen, S., & Nielsen, B. (2019). Models where the least trimmed squares and least median of squares estimators are maximum likelihood. Working paper 2019w05Oxford University Nuffield College.

Bontemps, C., Mizon, G.E., 2008. Encompassing: Concepts and implementation. Oxford Bulletin of Economics and Statistics 70, 721–750.

Castle, J.L., Doornik, J.A., Hendry, D.F., 2011. Evaluating automatic model selection. Journal of Time Series Econometrics 3 (1). DOI: 10.2202/1941–1928.1097

Castle, J.L., Doornik, J.A., Hendry, D.F., Pretis, F., 2015. Detecting location shifts during model selection by step-indicator saturation. Econometrics 3 (2), 240–264.

Castle, J. L., Doornik, J. A., Hendry, D. F., & Pretis, F. (2019). Trend-indicator saturation. Working paperOxford University Nuffield College.

Castle, J.L., Hendry, D.F., 2010. A low-dimension portmanteau test for non-linearity. Journal of Econometrics 158, 231–245.

Castle, J.L., Hendry, D.F., 2011. Automatic selection of non-linear models. In: Wang, L., Garnier, H., Jackman, T. (Eds.), System identification, environmental modelling and control. Springer, New York, pp. 229–250.

Castle, J.L., Hendry, D.F., Martinez, A.B., 2017. Evaluating forecasts, narratives and policy using a test of invariance. Econometrics 5 (39). doi:10.3390/econometrics5030039.

Cox, D.R., 1962. Further results on tests of separate families of hypotheses. Journal of the Royal Statistical Society B, 24, 406–424.

Doob, J.L., 1953. Stochastic Processes. John Wiley Classics Library, New York. 1990 edition

Doornik, J.A., 2008. Encompassing and automatic model selection. Oxford Bulletin of Economics and Statistics 70, 915–925.

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

Doornik, J.A., 2016. An example of instability: Discussion of the paper by Søren Johansen and Bent Nielsen. Scandinavian Journal of Statistics 43, 357–359.

Doornik, J.A., Castle, J.L., Hendry, D.F., 2020. Short-term forecasting of the coronavirus pandemic. International Journal of Forecasting doi:10.1016/j.ijforecast.2020.09.003.

Doornik, J.A., Hansen, H., 2008. An omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics 70, 927–939.

Doornik, J.A., Hendry, D.F., 2015. Statistical model selection with big data. Cogent Economics and Finance doi:10.1080/23322039.2015. 1045216.

Doornik, J.A., Hendry, D.F., 2018. OxMetrics: An Interface to Empirical Modelling, edition 8th. Timberlake Consultants Press, London.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

Ericsson, N. R. (2012). Detecting parameter nonconstancy and changes in regime. Working paperWashington, D.C Federal Reserve Board of Governors.

Ericsson, N.R., MacKinnon, J.G., 2002. Distributions of error correction tests for cointegration. Econometrics Journal 5, 285–318.

Harding, S. G. (1976). Can theories be refuted?Dordrecht, Holland: D. Reidel Publishing Company.

Harrison, D., Rubinfeld, D.L., 1978. Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management 5, 81–102.

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

Hendry, D.F., Johansen, S., 2015. Model discovery and Trygve Haavelmo's legacy. Econometric Theory 31, 93–114.

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

Hendry, D.F., Mizon, G.E., 2011. Econometric modelling of time series with outlying observations. Journal of Time Series Econometrics 3 (1). 10.2202/1941–1928.1100

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

Hettmansperger, T.P., Sheather, S.J., 1992. A cautionary note on the method of least median squares. The American Statistician 46, 79–83.

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

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

Johansen, S., Nielsen, B., 2016. Asymptotic theory of outlier detection algorithms for linear time series regression models. Scandinavian Journal of Statistics 43, 321–348.

Kitov, O. I., & Tabor, M. N. (2015). Detecting structural changes in linear models: A variable selection approach using multiplicative indicator saturation. Unpublished paper University of Oxford.

Koenker, R., 1982. Robust methods in econometrics. Econometrics Reviews 1, 213–255.

Mason, R.L., Gunst, R.F., Hess, J.L., 1989. Statistical Design and Analysis of Experiments. John Wiley, New York.

Mayo, D. G. (2018). Statistical inference as severe testing. Cambridge: Cambridge University Press.

Mayo, D.G., Spanos, A., 2006. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. British Journal for the Philosophy of Science 57, 323–357.

Mizon, G.E., 1995. A simple message for autocorrelation correctors: Don't. Journal of Econometrics 69, 267–288.

Peña, D., 2019. Detecting outliers and influential and sensitive observations in linear regression. In: Handbook of engineering statistics. Springer, New York. Forthcoming

Popper, K.R., 1959. The Logic of Scientific Discovery. Basic Books, New York.

Popper, K.R., 1963. Conjectures and Refutations. Basic Books, New York.

Pretis, F., Schneider, L., Smerdon, J.E., Hendry, D.F., 2016. Detecting volcanic eruptions in temperature reconstructions by designed break-indicator saturation. Journal of Economic Surveys 30, 403–429.

Ramsey, J.B., 1969. Tests for specification errors in classical linear least squares regression analysis. Journal of the Royal Statistical Society B, 31, 350–371.

Ronchetti, E., 1985. Robust model selection in regression. Statistics and Probability Letters 3, 21–23.

Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.

Stillwagon, J.R., 2016. Non-linear exchange rate relationships: An automated model selection approach with indicator saturation. North American Journal of Economics and Finance 37, 84–109.

Víšek, J.A., 1999. The Least Trimmed Squares – random carriers. Bulletin of the Czech Econometric Society 6, 1–30.

Walker, A., Pretis, F., Powell-Smith, A., Goldacre, B., 2019. Variation in responsiveness to warranted behaviour change among NHS clinicians: a novel implementation of change-detection methods in longitudinal prescribing data. British Medical Journal 367, l5205.

White, H., 1980. A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.

Zhou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101:476, 1418–1429.