Robust Discovery of Regression Models

Jennifer L. Castle, Jurgen A. Doornik and David F. Hendry*
Department of Economics, Nuffield College, Magdalen College, and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, UK

April 2020

Abstract

Since complete and correct *a priori* specifications of models for observational data never exist, model selection is unavoidable in that context. The target of selection needs to be the process generating the data for the variables under analysis, while retaining the objective of the study, often a theory-based formulation. Successful selection requires robustness against many potential problems jointly, including outliers and shifts; omitted variables; incorrect distributional shape; non-stationarity; mis-specified dynamics; and non-linearity, as well as inappropriate exogeneity assumptions. The aim is to seek parsimonious final representations that retain the relevant information, are well specified, encompass alternative models, and evaluate the validity of the study. Our approach to doing so inevitably leads to more candidate variables than observations, handled by iteratively switching between contracting and expanding multi-path searches, here programmed in *Autometrics*.

We investigate the ability of indicator saturation to discriminate between measurement errors and outliers, between outliers and large observations arising from non-linear responses (illustrated by artificial data), and apparent outliers due to alternative distributional assumptions. We illustrate the approach by exploring empirical models of the Boston housing market and inflation for the UK (both tackling outliers and non-linearities that can distort other estimation methods). We re-analyze the 'local instability' in the robust method of least median of squares shown by Hettmansperger and Sheather (1992) using indicator saturation to explain their findings.

JEL classifications: C51, C22.
KEYWORDS: Model Selection; Robustness; Outliers; Location Shifts; Indicator Saturation; *Autometrics*.

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 by a 'robust method', while assuming omniscience in all other model aspects. Complete and correct *a priori* specifications almost never exist for models of observational data, so model selection is unavoidable. The target of selection must be discovering the data generating process (DGP) for the variables being modelled while embedding the objective of the

*Financial support from the Robertson Foundation (award 9907422), Institute for New Economic Thinking (grant 20029822), and the ERC (grant 694262, DisCont) is gratefully acknowledged. Presented at the Department of Economics Econometrics Lunch. We wish to thank Vanessa Berenguer-Rico, Andrew Martinez, Bent Nielsen, Daniel Peña, Elvezio Ronchetti, and Kevin Sheppard for helpful comments. email: jennifer.castle@magd.ox.ac.uk, jurgen.doornik@nuffield.ox.ac.uk and david.hendry@nuffield.ox.ac.uk

analysis, which is often a theory-based formulation. Successful selection requires robustifying models against as many contaminating influences as possible which includes:

- C1 omitted variables—tackled by initially including all likely explanatory variables (§3.1);
- C2 inadequate dynamics—by including sufficient lags for a sequential factorization (§3.1);
- C3 misspecified linearity—by a general low-dimensional non-linear representation (§3.1);
- C4 outliers & incorrect distributional assumptions—by impulse-indicator saturation (IIS: §3.3);
- C5 location shifts—by step-indicator saturation (SIS: §3.3);
- C6 stochastic trends—by cointegration and differencing, which are well-established approaches not considered here;
- C7 invalid conditioning—by checking invariance and exogeneity (§3.4).

Crucially, C1–C7 need to be addressed jointly as much as possible to avoid confounding the problems. Given this array of possible mistakes, we proceed by starting selection from a model that is sufficiently general to characterize the target by allowing for these potential problems, yet sustain evaluation of the objective.

The structure of the paper is as follows. After describing the background in Section 2, we briefly rehearse formulating the general unrestricted model (GUM) in Section 3 (problem C1). That requires creating lags to handle dynamics (problem C2), functional-form transformations for non-linearities (problem C3), and indicator saturation for outliers and location shifts (problems C4 & C5). §3.2 describes the evaluation concepts of gauge and potency and §3.3 explains indicator saturation methods, while §3.4 describes testing invariance and exogeneity (problem C7): see Engle, Hendry, and Richard (1983). Section 4 considers cases when indicator saturation provides robustness in selection, including incorrect distributional assumptions, namely fat-tailed distributions, and discriminating between non-linearities and outliers using *Autometrics* with a simulation illustration (see Stillwagon, 2016, for an empirical application). Next, section 5 reconsiders three illustrative case studies, namely re-analyzing the Boston housing market data (§5.1); how to reveal a theoretical artefact induced by failing to model location shifts (§5.2); and an application of all our tools to unweave a problematic case study that had revealed an intrinsic instability in some robust estimation methods (5.3). Finally, section 6 concludes.

2 Empirical model discovery

Empirical modelling of observational data cannot usefully proceed by assuming that *ceteris paribus* conditions can be taken to hold in reality however easily they may be imposed in theories. Economies, societies and environmental systems all evolve and are intermittently hit by natural shocks, wars, crises, policy changes and other, often unanticipated, events. Further, observed time-series data are often uncertain and can be subject to substantial revisions. Both facets lead to potentially misspecified empirical models, inducing forecast failure (a significant deterioration in forecast performance relative to the anticipated outcome), and misleading policy implications. Taken together, if not tackled, these difficulties can cast doubt on the theories used to motivate the objective of modelling, even if those theories are in fact essentially 'correct'. The lack of a complete and correct *a priori* specification of a model for observational data makes model selection unavoidable. Hence, a method of model selection that is robust to many directions of potential misspecifications but still retains relevant theory information is needed in many disciplines.

Hendry and Johansen (2015) provide a solution by merging theory-driven and data-driven approaches to retain the theory objective while ensuring robustness against possible mistakes and omissions in applying that theory to non-stationary data. By orthogonalizing the variables deemed irrelevant by the theory against those thought relevant, when the theory-model is complete and correct, selecting over a vastly larger initial specification that nests the theory will still result in precisely the same estimates as

directly fitting that theory to the data. But if the theory is not complete and correct, commencing from a more general starting point will enable the discovery of a better formulation. Free lunches are rare in economics but this is a win–win scenario. The modeller will get the same result when the theory is correct, and an improved model if the theory is incorrect: see Hendry and Doornik (2014) for details.

In this approach, the theory-model is the null hypothesis to be evaluated stringently against a range of likely alternatives, preferably checking if it simultaneously encompasses all its rivals to ensure severe testing (see Mayo and Spanos, 2006, and Mayo, 2018). Seen from a philosophy of science perspective, if it successfully does so, this provides corroboration of the theory-model in the sense of Karl Popper (1959, 1963), even though that is the null hypothesis. That null could also be strongly rejected if many of the rival candidate variables are found to be highly significant. Of course, the specification of a theory by a given model is not necessarily unique, so there may be potential 'rescue' strategies to protect it (see e.g., Lakatos, 1974), but repeated rejection should trigger a rethink. Conversely, after a strong rejection, an objective researcher may seek to explore why that occurred, and will have immediate access to information on which of the rival candidate variables led to rejection. Although that does not fully overcome the objections to 'accepting the alternative' when the null is rejected, as an excluded source may be responsible (see e.g., Harding, 1976), such information does at least point towards an improved model–and that could be one that may actually still be consistent with the initial theory (see the examples in Hendry and Mizon, 2011, and Hendry, 2018).

3 Automatic model discovery

In this section we describe the initial model formulation and some relevant aspects of automatic model selection. At first sight, the initial models that are created look infeasibly large, but very general specifications are needed if the approach is to be robust to as many forms of potential misspecification as feasible. Moreover, robustness cannot be achieved unless all the complications are tackled as jointly as possible, otherwise misspecification in one direction can lead to another aspect of the model proxying that misspecification and resulting in the wrong inferences. An example is non-modelled outliers (problem C4) that lead a modeller to detect apparent non-linearities that are just an artifact due to misspecification. By modelling the outliers jointly with possible non-linearities the modeller can discriminate between the competing explanations on data evidence rather than *ad hoc* imposed assumptions. Section 4.2 examines this particular example further.

3.1 Formulation of the general unrestricted model (GUM)

Given the r variables $\mathbf{w}_t, t = 1, \ldots, T$, considered by the investigator, three extensions can be automatically implemented to create the GUM, namely dynamics, functional-form transformations for non-linearities and indicator variables to capture outliers and shifts. Before doing so, partition the candidate variables into $\mathbf{w}'_t = (\mathbf{x}'_t : \mathbf{v}'_t)$, where r_1 $\mathbf{x}_t$ are theory specified (their parameters are the 'objective') and are not subject to selection, and the remaining r_2 variables $\mathbf{v}_t$ are then orthogonalized with respect to $\mathbf{x}_t$. The addition of $\mathbf{v}_t$ is to tackle problem C1.

Next, create s lags of $\mathbf{w}_t$ to implement a sequential factorization (see Doob, 1953) and tackle problem C2: see Castle, Doornik, and Hendry (2011). Departures of linearity, C3, can be handled using the approach of Castle and Hendry (2011): including a small set of transformations of the prinicpal components of $\mathbf{w}_t$. We do not use this in the applications below.

Fourth, to tackle problems C4 and C5 of potential outliers and shifts, create T impulse indicators, $1_{\{i=t\}}$ which are zero except for unity at observation t for $t = 1, \ldots, T$ and/or step indicators (depending on the problem under analysis), to be added to the set of candidate variables, as discussed further in §3.3.

The resulting GUM is given by:

$$y_t = \sum_{i=1}^{r_1} \theta_i^F x_{i,t} + \sum_{i=1}^{r_2} \theta_i v_{i,t} + \sum_{i=1}^{r} \sum_{j=1}^{s} \phi_{ij} w_{i,t-j} + \sum_{j=1}^{s} \rho_j y_{t-j} + \mu^F + \sum_{i=1}^{T} \delta_i 1_{\{i=t\}} + \sum_{i=2}^{T-1} \gamma_i 1_{\{i \leq t\}} + \epsilon_t, \tag{1}$$

where $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$ after selection. This leads to $N = r_2 + r(s+1) + 2T - 2$ candidate regressors when both impulse and step indicators are included, so $N > T$. The intercept and r_1 variables $x_{i,t}$ are 'fixed', denoted by the superscript F on the coefficients: they are forced to be present in all estimated models.

3.2 Evaluation of model selection

Selecting from the GUM can be automated in many ways: here we mainly use *Autometrics*[1] (see Doornik, 2009) which seeks parsimonious well-specified final representations that retain theory insights and encompass rival models' results: see Bontemps and Mizon (2008) for an overview of encompassing, and Doornik (2008) for its role in *Autometrics*. Simulations are undertaken to assess the properties of selection, evaluated based on many draws of the data using the criteria of gauge and potency, as well as mean square error measures.

For notational simplicity we stack the parameters in the GUM (1) as $\beta = \left(\theta^{F\prime}, \theta', \phi', \rho', \delta', \gamma'\right)'$. We assume that the DGP is nested within the GUM, so, for a given DGP, let $\beta_1 = \cdots = \beta_n \neq 0$, but $\beta_{n+1} = \cdots = \beta_N = 0$. The estimated coefficient of variable j after selection in replication i is $\widetilde{\beta}_{j,i}$ for $i = 1, 2, ..., M$ replications: $\widetilde{\beta}_{j,i} = 0$ when the corresponding variable is not selected in the final model. The OLS estimates of the DGP are denoted $\widehat{\beta}_{j,i}$.

The gauge is defined as the fraction of retained irrelevant variables, and potency as the average retention frequency of DGP variables. Starting from the retention rate of variable j:

$$\begin{aligned} \text{retention rate } \widetilde{\mathsf{p}}_j &= \tfrac{1}{M} \textstyle\sum_{i=1}^{M} 1_{(\widetilde{\beta}_{j,i} \neq 0)}, \quad j = 1, \ldots, N, \\ \text{potency} &= \tfrac{1}{n} \textstyle\sum_{j=1}^{n} \widetilde{\mathsf{p}}_j, \\ \text{gauge} &= \tfrac{1}{N-n} \textstyle\sum_{j=n+1}^{N} \widetilde{\mathsf{p}}_j. \end{aligned}$$

We also calculate the following mean-square error (MSE) measures after model selection:

$$\begin{aligned} \mathsf{MSE}_j &= M^{-1} \textstyle\sum_{i=1}^{M} \left(\widetilde{\beta}_{j,i} - \beta_j\right)^2, \\ \mathsf{CMSE}_j &= \widetilde{\mathsf{p}}_j^{-1} M^{-1} \textstyle\sum_{i=1}^{M} \left(\widetilde{\beta}_{j,i} - \beta_j\right)^2 1_{(\widetilde{\beta}_{j,i} \neq 0)}, \quad (\mathsf{CMSE}_j = \beta_j^2 \text{ if } \widetilde{\mathsf{p}}_j = 0). \end{aligned}$$

The first is unconditional because it includes $\widetilde{\beta}_{j,i} = 0$ when a variable is not selected. The second is the *conditional* MSE that is computed over retained variables only.

As measures of success, a gauge close to α, the significance level for selection, and potency close to the DGP power are sought, as are small MSEs. Hendry and Doornik (2014) discuss these concepts extensively and Johansen and Nielsen (2016) derive the distribution of the estimated gauge.

[1] The *Autometrics* algorithms are available in Doornik and Hendry, 2018 (`www.doornik.com`), the Excel add-in XLModeler (`www.xlmodeler.com`), and in R (Pretis, Reade, and Sucarrat, 2018) All results reported below were obtained with OxMetrics 8.2.

A GUM like (1) is inevitably highly over-parametrized. Indeed, for $r = 10$, $s = 2$, $T = 100$ and $r_1 = 4$, there are $N = 224$ regressors in the GUM. 'Conventional wisdom' might suggest that selection from such a large GUM must be inefficient as the non-null retention frequency of the procedure will be excessive. The average number of null variables retained, k, assuming $n = 0$, is given by:

$$k = \sum_{i=0}^{N} i \frac{N!}{i!\,(N-i)!} \alpha^i \,(1-\alpha)^{N-i} = N\alpha. \tag{2}$$

Provided α is appropriately controlled, the probability of retaining irrelevant variables can be small. Assuming 150 irrelevant variables and using $\alpha = 0.0025$, gives $k = 0.375$, so few will be significant despite 10^{45} possible models. The theoretical gauge is given by $k/N = \alpha$, matching the nominal significance level on average (see Johansen and Nielsen, 2016). Adopting a smaller α reduces potency, but does not affect theory variables as they are always retained.

3.3 Indicator saturation methods

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

IIS impulse-indicator saturation for outliers (Hendry, Johansen, and Santos, 2008, and Johansen and Nielsen, 2009);

SIS step-indicator saturation for location shifts (Castle, Doornik, Hendry, and Pretis, 2015);

TIS trend-indicator saturation for trend breaks (Castle, Doornik, Hendry, and Pretis, 2019, with an application in Walker, Pretis, Powell-Smith, and Goldacre, 2019);

DIS designed-indicator saturation for specific shapes (matching volcanic eruption impacts on temperature in Pretis, Schneider, Smerdon, and Hendry, 2016);

MIS multiplicative saturation for changes in other parameters (Ericsson, 2012, Kitov and Tabor, 2015).

All saturation methods lead to models with more variables than observations. Feasible estimators select from the variables in blocks. In the simplest case of IIS this is the same as partitioning the estimation sample. But, when combined with selection over other variables, it is more convenient to treat the impulse dummy as just another variable. Selection over the blocks proceeds iteratively until convergence under the constraint that fewer than $0.75T$ (say) are selected in total. After this a final model selection step can be undertaken. Appendix C provides details of the *Autometrics* implementation. In general, different ordering of the blocks can lead to different selected models. Efficient implementation of the estimation algorithm is useful considering the huge search space. Random search would also be possible, but has no practical advantage, except perhaps for asymptotic analysis.

Johansen and Nielsen (2009) develop IIS theory for both stationary and unit-root autoregressions, and Johansen and Nielsen (2016) link IIS to robust statistics by showing it is an iterated 1-step Huber-skip M-estimator. They show that for a parameter of interest β in a regression equation, the loss of efficiency of the IIS estimator, $\widetilde{\beta}$, under the null, with respect to the least squares estimator, $\widehat{\beta}$, depends on the selection critical value, c_α, and the error distribution. When the error distribution is symmetric, in a stationary regression with r variables that have coefficient β and scaled asymptotic second moment $\mathbf{\Sigma}$, selecting from T impulse indicators under the null of no outliers leads to:

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

The efficiency of the IIS estimator $\widetilde{\beta}$ with respect to OLS $\widehat{\beta}$ measured by $\mathbf{\Omega}_\alpha$ depends on c_α and the distribution, but is close to $(1-\alpha)^{-1}\mathbf{I}_r$ for small α. Thus even with $N > T$, the usual $\sqrt{T}$ stationary

convergence rate to a normal distribution holds, correctly centered on β and with almost the standard asymptotic variance matrix $\sigma_\epsilon^2 \mathbf{\Sigma}^{-1}$ but weighted by the efficiency matrix. Despite T extra candidates, there is only a small loss of efficiency under the null for small α, against potentially large gains under alternatives of multiple outliers and shifts.

3.4 Testing the validity of conditioning using IIS

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

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

where $\mu_{y,i} = \beta_0 + \beta_1 \mu_{z,i}$ is the theory model of interest. Then the conditional expectation is:

$$\mathsf{E}[y_i \mid z_i] = \mu_{y,i} + \omega_{22}^{-1} \omega_{12} \left(z_i - \mu_{z,i} \right) = \beta_0 + (\beta_1 - \gamma) \mu_{z,i} + \gamma z_i. \tag{3}$$

Thus, unless $\beta_1 = \gamma (= \omega_{22}^{-1} \omega_{12})$, the conditional expectation $\mathsf{E}[y_i | z_i]$ 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 (3), 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 provides a means of doing so: see Hendry and Santos (2010).

4 Robustness to two misspecifications

First we assess the ability of IIS to provide robustness to incorrect distributional assumptions (§4.1) and misspecification due to unknown outliers and non-linearities (§4.2). Monte Carlo simulations are used to examine both cases. Castle and Hendry (2014) consider model selection in under-specified equations when there are location shifts.

4.1 IIS in fat-tailed distributions

Often normality is assumed in economic modelling, but a procedure that is robust to potentially incorrect distributional assumptions would be preferable if its inefficiency under the null was small. IIS can provide robustness to some incorrect distributional assumptions: in particular, a fat-tailed error distribution such as a Student-t distribution with small degrees of freedom can be designed to be closer to normal with impulse indicators capturing the fat tails. The advantage of such a procedure when the fat-tail form is unknown is that approximately 'correct' critical values are used for selection of the regressors. Consider the relationship:

$$\begin{aligned} y_t &= \beta_1 T^{-1/2} z_{1,t} + \cdots + \beta_{12} T^{-1/2} z_{12,t} + \epsilon_t, \ t = 1, ..., T, \\ \mathbf{z}_t &\sim \mathsf{IN}_{12} \left[\mathbf{0}, \mathbf{I}_{12} \right], \qquad (4) \\ \epsilon_t &\sim \mathsf{t}(3), \qquad (5) \end{aligned}$$

where $\mathbf{z}_t' = (z_{1,t}, \cdots, z_{12,t})$, is fixed across replications, and common random numbers are used between settings to reduce simulation variance. We create 3 different DGPs for this relationship:

DGP-A : $\beta_1 = \cdots = \beta_{12} = 0$;
DGP-B : $\beta_1 = 1; \beta_2 = 2; \beta_3 = 3; \beta_4 = 4; \beta_5 = 6; \beta_6 = 8; \beta_7 = \cdots = \beta_{12} = 0$;
DGP-C : as DGP-B, except (4) replaced by $z_{j,t} \sim \mathsf{t}(3), \ j = 1, \ldots, 12$.

Design	A(IIS)	A(IIS+Z)	B(IIS)	B(IIS+Z)	C(IIS)	C(IIS+Z)
Retention rate (impulses)	0.028	0.032	0.028	0.029	0.028	0.029
MCSD (impulses)	0.018	0.020	0.018	0.018	0.018	0.018

Table 1: Average retention rate and MCSD of impulse indicators for $\mathsf{t}(3)$ in IIS with selection at $\alpha = 0.005$. See the main text for the designs. $\mathsf{MCSE} = \mathsf{MCSD}/M^{1/2}$ for $M = 5000$ replications.

DGP-B(.)	NONE	IIS	Z	IIS+Z	Z	IIS+Z	Z	IIS+Z
	Unconditional 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*

Table 2: MCSD and average retention rate of z variables in DGP-B, with and without IIS, and with or without variable selection. Conditional MCSD is measured for variables having been selected; retention amounts to **potency** for relevant variables, *gauge* for irrelevant; $M = 5000$, $\alpha = 0.005$.

Four experiments are considered for these DGPs, based on the initial model

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

(IIS) forces the 12 variables to be retained in all estimated models and selection is conducted for IIS, applied at a significance level $\alpha = 0.005$;

(IIS+Z) applies selection at $\alpha = 0.005$ for both the variables and impulse indicators;

(NONE) no selection (and no IIS);

(Z) selection over the z_i variables but without IIS.

The intercept is included in all estimated models, but omitted from the evaluation. We use $M = 5000$ replications with $T = 100$, so $\beta_1 T^{-1/2} = 0.1$ in DGP-B.

Table 1 records the average retention rates of impulse indicators.[2] The simulations are very precise and show a relatively constant mean retention rate, little affected by the presence of variable selection, and null or alternative. However, the Monte Carlo standard deviation (MCSD) of the average retention rate is quite large, so the distribution of retention rates across replications varies considerably, with a maximum of around 14%. DGP-C with $\mathsf{t}(3)$ distribution for all $z_{i,t}$ demonstrates that IIS does not get confused by irrelevant variables having the same error distribution as the conditional model of interest.

Table 2 records the unconditional and conditional MCSDs for the z variables in DGP-B, which has the first six variables non-null, using selection at $\alpha = 0.005$. Results are averaged separately across relevant and irrelevant variables (with the constant term and impulses ignored in the retention statistics). If the DGP distribution is incorrectly assumed to be normal when it is in fact $t(3)$, the gauge of 3.2% is too high for a significance of $\alpha = 0.5\%$. With IIS the gauge falls to 0.9%, achieved at a higher potency: IIS removes the fat tails, bringing the error distribution closer to normality. The simulations therefore suggest that selection is not too 'over-gauged' with IIS and yet there is little effect on the potency. Although the MCSDs for the irrelevant variables seem large conditional on their selection at 0.36 on average, they are only selected when they are far from the population value of zero, which happens rarely even for $\mathsf{t}(3)$ errors.

Figure 1 plots the unconditional distributions of two parameter estimates for DGP-B. In the first case, the solid line, the errors are standard normal. The second case, the dashed line, is for DGP-

[2]Because the standard algorithm for $N > T$ is somewhat over gauged from the retention of insignificant variables caused by backtesting, we have backtesting in the final model selection switched off. See Appendix C.

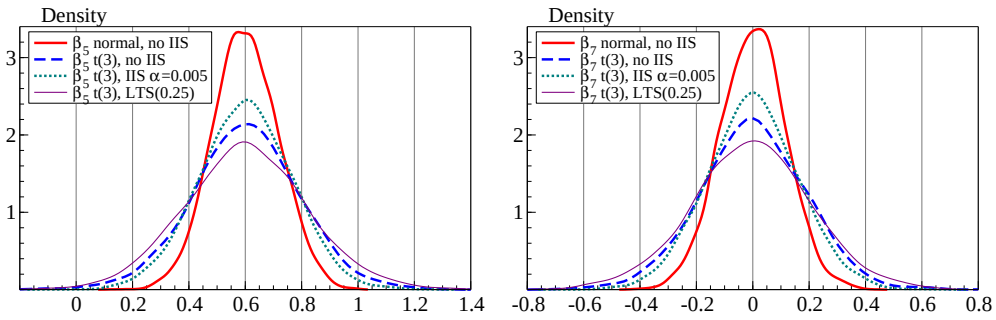

Figure 1: Estimated distribution of coefficients of $z_{5,t}$ and $z_{7,t}$, normal errors (solid line) and $t(3)$ errors.

B(NONE), showing that neglected $t(3)$ errors make inference less precise. Using IIS, as shown in the dashed line, moves the density towards that with normal errors. If the error distribution were known, then selection could use criteria based on the relevant distribution. In practice, with unknown error distributions, selection with IIS offers some robustness using critical values based on normality as a reasonable approximation.

4.1.1 Least trimmed squares

Least trimmed squares (LTS) and Least median of squares (LMS) are robust estimators introduced by Rousseeuw (1984). LTS(0.5) finds that half of the observations that minimizes the residual sum of squares (LMS does the same for the median). LTS is regularly 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. Vîšek (1999) provides asymptotic analysis, while Beringuer-Rico, Johansen, and Nielsen (2019) derive settings where LTS and LMS are the maximum likelihood estimator.

Both IIS and LTS classify observations as outliers. The thin lines in Figure 1 show the distribution of the estimated coefficients when estimating by 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: LTS exacerbates the problem caused by $t(3)$ errors.

4.2 Non-linearities and outliers

We next investigate whether IIS is able to discriminate between non-linearities and outliers. We seek robustness in both directions, so if the DGP contains non-linearities, these should be modelled 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.

Two sources of misspecification are considered, truncation and contamination. Truncation occurs due 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. Figure 2a shows a quadratic function with no outliers, while panel b is a linear function with observations $1, 2, 3$, as outliers, given by a 3 standard deviation shift in the mean of the linear function. In both cases observations $4, \ldots, 13$ are missing

Contamination occurs when some observations are discrepant relative to the DGP. Figure 3 illustrates this by interacting ten consecutive observations with a step dummy. In Figure 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.

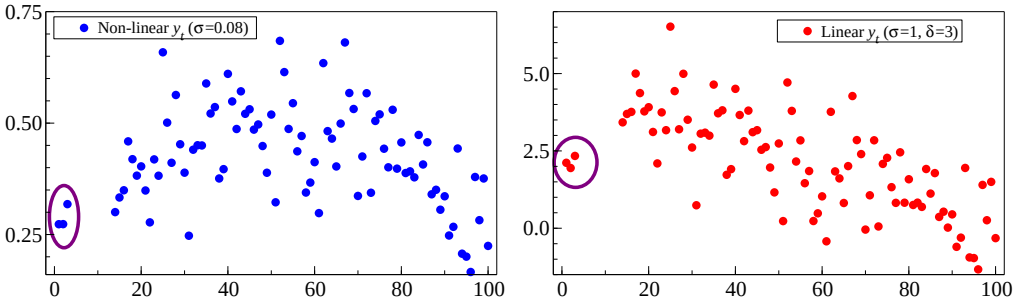

Figure 2: Data from a quadratic DGP (panel a, left) and a linear DGP with outliers in the first three observations (panel b, right), in both cases with observations $4, \ldots, 13$ missing.

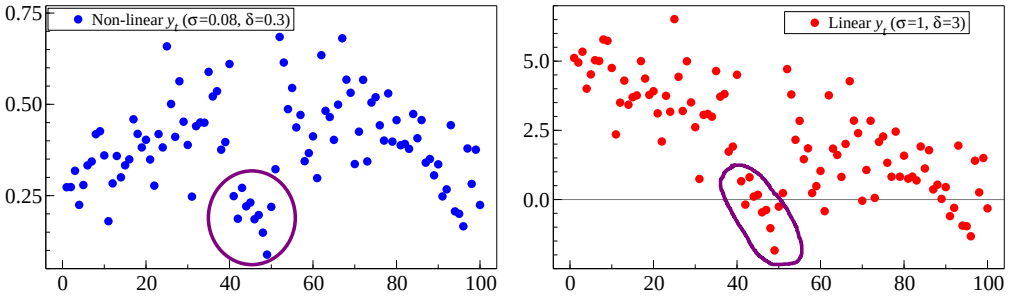

Figure 3: Data from a quadratic DGP (panel a, left) and a linear DGP (panel b, right) with 10 contaminated observations shown in elipses, given by a 3 standard deviation downward shift of observations $41, \ldots, 50$.

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

$$y_i^* = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \sigma u_i, u_i \sim \mathsf{IN}[0,1],\ i = 1, \ldots, T,$$
$$\text{DGP-Q} : \beta_0 = 0.25, \beta_1 = 1, \beta_2 = -1, \tag{6}$$
$$\text{DGP-L} : \beta_0 = 5, \beta_1 = -5, \beta_2 = 0. \tag{7}$$

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

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

   The truncated sample is obtained by dropping observations $i = 4, \ldots, 13$, leaving $T = 90$.
2. DQP-Q with contamination versus DGP-L with contamination, so in both cases:

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

The GUM consists of the intercept (fixed as usual), z_i, z_i^2, and IIS. Table 3 reports the retention rates, biases, and MSEs for selection using *Autometrics* with $\alpha = 0.01$ and $M = 5000$. Two signal-to-noise ratios are considered for each DGP and $\delta = -3$. Average retention is reported over the first three indicators, which correspond to outliers added to DGP-L, so are reflecting potency (bold in the table), but gauge for DGP-Q (italic). Figure 4 demonstrates the results for one draw of the experiment, where retained indicators are labelled.

	Rate	Bias	RMSE	*Rate*	Bias	RMSE	*Rate*	Bias	RMSE	*Rate*	Bias	RMSE
	$\sigma = 0.08$			$\sigma = 0.04$			$\sigma = 1$			$\sigma = 0.5$		
	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*		

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

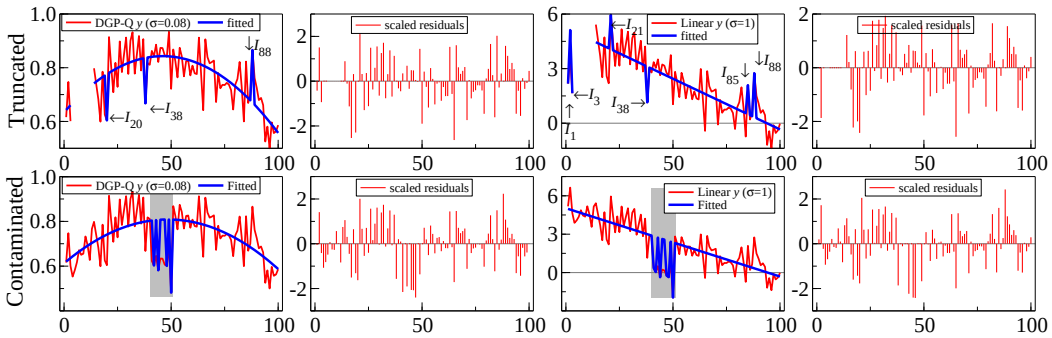

Figure 4: Top row records one truncated experiment and bottom row a contaminated experiment. Model fit from *Autometrics* with IIS and the quadratic function when the DGP is non-linear with no outliers (left panels) or linear with outliers (right panels).

For the non-linear DGP, the quadratic function is retained with unit probability for both large and small signal-to-noise ratio, and the bias and MSE on the quadratic term are small. Hence, accurate estimates of the non-linearity can be obtained despite inclusion of more variables than observations. The three irrelevant impulse indicators in DGP-Q are retained too frequently but still close to the 1% target size. For the linear DGP, the quadratic function is retained too often with a large signal-to-noise ratio (at 25% for $\sigma = 1$), resulting in the indicators not always being retained (at around 50% retention). However, a smaller σ is able to distinguish between the two hypotheses more clearly, and the precision with which the trend and outliers are picked up is much improved.

The second part of Table 3 records the results with contamination. For DGP-Q, the quadratic function is almost always retained and is precisely estimated. The contaminated data is picked up by about half of the outliers modelled by indicator variables. A joint test of equal coefficients would reveal that these could be replaced with a step dummy, increasing power, or SIS could be applied initally. The irrelevant indicators are retained very infrequently, at a lower probability than α, so 'overfitting' is not a concern. For the linear DGP the quadratic function is almost always excluded, so joint selection is not costly. Retention of the indicators for the contaminated data matches that of the non-linear DGP, demonstrating that the properties of IIS do not depend on the functional form of the DGP.

	Rate	Bias	RMSE	*Rate*	Bias	RMSE	*Rate*	Bias	RMSE
	DGP-Q contaminated $\sigma = 0.08$, (SIS+Z)			DGP-L contaminated $\sigma = 1$, (SIS+Z)			DGP-L double $\sigma = 1$, (SIS+Z)		
z_i	**0.693**	−0.306	0.631	**0.711**	1.304	3.496	**0.684**	1.272	3.466
z_i^2	**0.800**	0.271	0.547	*0.085*	−0.244	1.798	*0.110*	−0.064	2.478
$\mathcal{S}_1$	**0.794**			**0.787**			**0.770**		
$\mathcal{S}_1 \pm 1$	**0.951**			**0.942**			**0.922**		
$\mathcal{S}_2$	*0.017*			*0.018*			**0.773**		
$\mathcal{S}_2 \pm 1$							**0.907**		
gauge SIS	*0.023*			*0.024*			*0.023*		

Table 4: SIS estimates. Contamination: observations $\{41, , ..., 50\}$ shifted by $\delta = -3$, with corresponding steps $\mathcal{S}_1 = \{S_{40}, S_{50}\}$. Double contamination also has $\{98, \dots 100\}$ shifted by $\delta = -3$, with $\mathcal{S}_2 = \{S_{97}\}$. *Rate* is the retention rate of variables and impulse indicators: **bold** denotes potency and *italic* gauge. Bias and RMSE are unconditional. $M = 5000$, $\alpha = 0.01$.

	Bias	RMSE	Bias	RMSE
	OLS		LTS	
z_i	−6.163	6.320	5.828	6.416
z_i^2	6.800	6.932	−4.072	4.931

Table 5: OLS and LTS estimates without selection over variables for experiment DGP-L with double contamination.

Overall, the results indicate that jointly selecting impulse indicators and non-linear functions does enable selection to discriminate accurately between the two hypotheses. The costs of testing for both forms of specification are small, particularly with diminishing noise.

In the final set of experiments we investigate how SIS performs in the contaminated setting. The first two cases are for DGP-Q and DGP-L, contaminated as in Table 3 using the higher variance, and fixed intercept. The third and fourth case adds a second contamination of $-\delta\sigma$ to the last three observations. The steps S_i for SIS are constructed with unity up to a designated endpoint i, and zero thereafter. The contamination over observations $\{41, \dots 50\}$ can be captured by selecting $\mathcal{S}_1 = \{S_{40}, S_{50}\}$. Table 4 shows the average potency of detecting these two steps in the line labelled $\mathcal{S}_1$. In general, there is some uncertainty about the exact timing of the break, therefore we add $\mathcal{S}_1 \pm 1$ for the average retention when allowing the start and end to be out by one period. The doubly contaminated experiments have the last three observations contaminated as well: $\mathcal{S}_2 = \{S_{97}\}$.

Comparing Tables 4 and 3, we see that the potency of SIS is higher than IIS, improving from around 50% to almost 80%, and well over 90% when allowing some leeway. However, this is at the expense of a higher gauge for steps than impulses, and some reduction in potency for z_i, z_i^2. This last effect is reflected in the biases and RSMEs, but as the experiment without SIS (or IIS) shows, there remain large improvements over ignoring the contamination.

Without saturation, the quadratic term is almost always selected into the model for the linear DGP: in that case the results are almost identical to those without selection in Table 5. LTS estimates are a small improvement over OLS without selection, but not as good as using SIS: compare the last two columns of Table 5 to Table 4.

5 Empirical applications of robust model selection

A range of case studies are reviewed to illustrate various aspects of robust model selection, including a demonstration of the importance of SIS to isolate subsample effects in Boston housing market (§5.1), and saturation techniques resulting in a rejection of theory in §5.2.

5.1 Re-analyzing the Boston Housing Market data

The Boston Housing Market data were originally from Harrison and Rubinfeld (1978), also used by Kuh, Belsley, and Welsh (1980) and most recently by Peña (2019). They are summarized in Appendix A. The data consists of 506 observations, and Kuh, Belsley, and Welsh (1980) notes that observations 357–488 correspond to Boston, whereas the rest correspond to the suburbs. Peña (2019) finds very different regression estimates in those sub-samples.

As a baseline for a non-robust equation, we first record in (B1) the regression of the log of the median value of owner-occupied homes, denoted *LmedVal*, on the 13 regressors listed in the appendix:

$$\begin{aligned}\widehat{\textit{LmedVal}} = \underset{(0.20)}{4.1} &- \underset{(0.001)}{0.010}\textit{Crime} + \underset{(0.055)}{0.117}\textbf{Zone} + \underset{(0.002)}{0.002}\textbf{Industry} + \underset{(0.034)}{0.101}\textit{Charles} \\ &- \underset{(0.153)}{0.778}\textit{NOx} + \underset{(0.017)}{0.091}\textit{Rooms} + \underset{(0.0005)}{0.0002}\textbf{Age} - \underset{(0.008)}{0.049}\textit{Distance} + \underset{(0.003)}{0.014}\textit{Radial} \\ &- \underset{(0.015)}{0.063}\textit{Tax} - \underset{(0.005)}{0.038}\textit{PTratio} + \underset{(0.011)}{0.041}\textit{BlkPop} - \underset{(0.002)}{0.029}\textit{LowStat} \\ \widehat{\sigma} = 0.190 \; \mathsf{R}^2 = 0.79 \; \mathsf{F}_{\mathsf{Het}}(25,480) &= 7.24^{**} \; \chi^2_{\mathsf{nd}}(2) = 92.5^{**} \; \mathsf{F}_{\mathsf{Reset}}(2,490) = 15.4^{**}\end{aligned} \tag{B1}$$

All the misspecification tests[3] strongly reject, and three of the regressors would not be judged significant at any reasonable level, shown in bold. The first row in Figure 5 shows respectively the scatter plot of actual against fitted values, the residuals, scaled to unit variance, and the QQ plot. These confirm the serious mismatch.

Applying IIS with *Autometrics* at $\alpha = 0.001$ to (B1), while fixing all its regressors, found 58 outliers with $\widehat{\sigma} = 0.11$ — many of these are insignificant, but retained because of backtesting or diagnostic testing in the final selection. The scaled residuals showed no further outliers, but revealed blocks of zeroes, mainly occurring over the subsample corresponding to Boston. The three misspecification tests also still rejected, so the formulation remained unsatisfactory. Consequently, an investigator might try combining IIS and SIS to capture the steps, albeit that such a modelling order is from simple to general. Doing so at 0.001 selected 9 impulse indicators and 32 step indicators despite commencing from 1010 candidates, with $\widehat{\sigma} = 0.09$.

As two misspecification tests rejected, we next discriminate between 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 the additional Boston variables to (B1), and selecting at 5%, but

[3]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, $\mathsf{F}_{\mathsf{Het}}$ is a test for residual heteroskedasticity (see White, 1980), $\chi^2_{\mathsf{nd}}(2)$ is a test for normality (see Doornik and Hansen, 2008), and $\mathsf{F}_{\mathsf{Reset}}$ is the RESET test (see Ramsey, 1969). Results are obtained with PcGive's models for cross-section in OxMetrics 8.20.

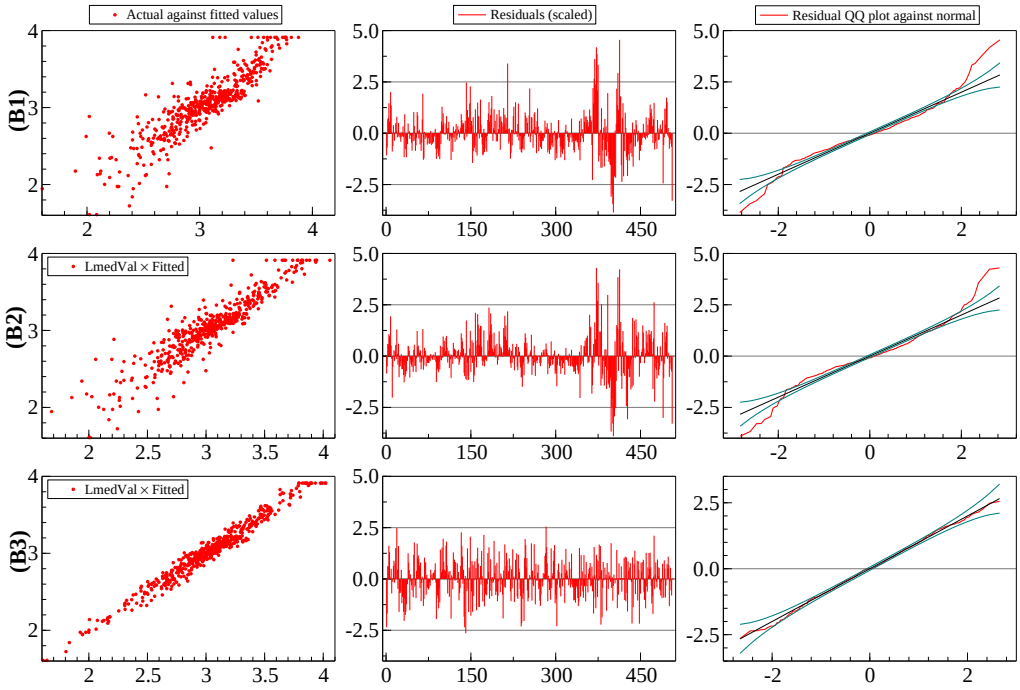

Figure 5: Graphical results for the three models of Boston house prices. In rows: (B1), (B2), (B3). In columns: scatter plots of actual against fitted values, scaled residuals, QQ plots against the normal distribution.

without any saturation, leads to:

$$\begin{aligned}\widehat{LmedVal} = {} & \underset{(0.186)}{2.02} + \underset{(0.019)}{0.285}\,Rooms - \underset{(0.0005)}{0.0023}\,Age - \underset{(0.005)}{0.027}\,Distance + \underset{(0.005)}{0.012}\,Radial \\ & - \underset{(0.011)}{0.059}\,Tax - \underset{(0.004)}{0.027}\,PTratio + \underset{(0.022)}{0.075}\,BlkPop - \underset{(0.002)}{0.009}\,LowStat \\ & - \underset{(0.001)}{0.010}\,BosCrime + \underset{(0.064)}{0.327}\,BosCharles - \underset{(0.257)}{1.79}\,BosNOx - \underset{(0.028)}{0.377}\,BosRooms \\ & + \underset{(0.001)}{0.006}\,BosAge - \underset{(0.024)}{0.057}\,BosBlkPop - \underset{(0.004)}{0.034}\,BosLowStat + \underset{(0.280)}{3.62}\,isBoston \\ \widehat{\sigma} = 0.159 \; & R^2_d = 0.79 \; \mathsf{F}_{\mathsf{Het}}(30,475) = 4.76^{**} \; \chi^2_{\mathsf{nd}}(2) = 108.1^{**} \; \mathsf{F}_{\mathsf{Reset}}(2,487) = 8.23^{**}\end{aligned} \tag{B2}$$

R^2_d is the R^2 relative to all deterministic terms, i.e. the intercept with all impulses and steps, which is the constant and *isBoston* in (B2). Although the fit is substantially improved, all the misspecification tests still strongly reject, and consequently it is difficult to judge which variables really influence house prices. There are noticeable differences from (B1): the overall negative effect of *Crime* has been replaced by *BosCrime*; *Charles* by *BosCharles* and *NOx* by *BosNOx*, both with much larger coefficients. As five of the Boston interactive variables are constant over the subsample, their effects relative to the full sample are all captured by *isBoston*. The middle row of Figure 5 shows that there is little improvement over (B1).

Finally, and what in a general to specific modelling exercise should have been the starting point, the outcome when selecting over all variables with IIS+SIS is recorded in (B3). With the overall and Boston intercepts fixed, 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 the final selection, after adding all the regressors back, to 81. The

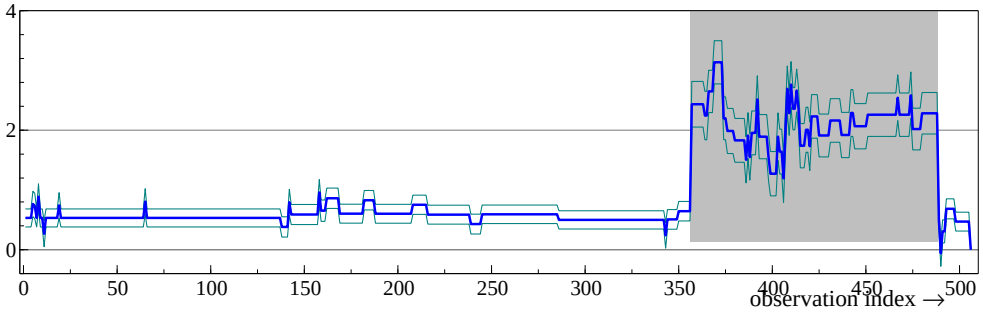

Figure 6: Contribution to the intercept for each observation from the constant term, impulse, and step indicators. Dotted line is plus/minus two standard errors. Boston is the shaded area.

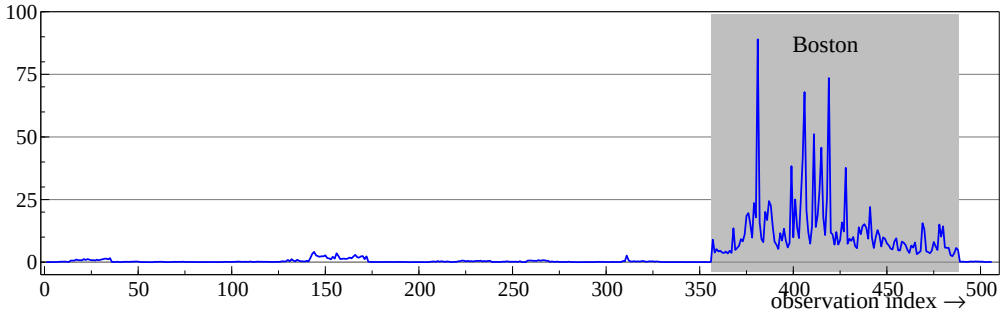

Figure 7: Graph of *Crime* in Boston (shaded area) and suburbs.

entire procedure took just over two minutes, and found (indicators not reported):

$$\begin{aligned}\widehat{LmedVal} = \underset{(0.12)}{1.39} &- \underset{(0.012)}{0.046}\mathit{Crime} + \underset{(0.025)}{0.071}\mathit{Zone} + \underset{(0.009)}{0.267}\mathit{Rooms} - \underset{(0.0002)}{0.0023}\mathit{Age} - \underset{(0.004)}{0.033}\mathit{Distance} \\ &+ \underset{(0.003)}{0.012}\mathit{Radial} - \underset{(0.007)}{0.038}\mathit{Tax} - \underset{(0.003)}{0.018}\mathit{PTratio} + \underset{(0.007)}{0.069}\mathit{BlkPop} - \underset{(0.0012)}{0.0083}\mathit{LowStat} \\ &+ \underset{(0.012)}{0.043}\mathit{BosCrime} - \underset{(0.176)}{0.724}\mathit{BosNOx} - \underset{(0.016)}{0.227}\mathit{BosRooms} + \underset{(0.17)}{1.79}\mathit{isBoston} \qquad \text{(B3)} \\ \widehat{\sigma} = 0.074 \; R_d^2 = 0.88 \; &F_{\mathsf{Het}}(66, 413) = 1.32 \; \chi^2_{\mathsf{nd}}(2) = 3.71 \; F_{\mathsf{Reset}}(2, 424) = 1.64\end{aligned}$$

No misspecification test now rejects, so no outliers, or hidden shifts, remain. Hence inference should be more reliable and the residual standard deviation is very much smaller. The bottom row of Figure 5 confirms the major improvements in the residual distribution.

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 is 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 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. Figure 6 shows the combined magnitude 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 Figure 7 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 §3.4 to test for super exogeneity. First we modelled *BosCrime* by the Boston regressors and IIS+SIS at 1%, finding thirteen indicators in the marginal model that are not in (B3). Adding them in the conditional model (B3), yielded the insignificant outcome $\mathsf{F}_{\mathsf{valcond}}(13, 413) = 1.5$. Thus, we conclude that *BosCrime* is super-exogenous and therefore a valid conditioning variable.

This empirical application demonstrates how SIS can reveal important differences in subsamples of data that are essential to model to obtain a congruent specification. Selection with saturation ensured a model specification that was robust to the different data properties of subsamples of the cross section and resulted in a well-specified and economically interpretable model.

5.2 Undoing a theoretical artefact: UK inflation

Many models of inflation, denoted Δp_t, such as the New-Keynesian Phillips curve (NKPC), include expected future inflation to explain current inflation, written as:

$$\Delta p_t = \underset{\geq 0}{\gamma_1} \mathsf{E}_t \left[\Delta p_{t+1} \mid \mathcal{I}_t\right] + \underset{\geq 0}{\gamma_2} \Delta p_{t-1} + \underset{\geq 0}{\gamma_3} s_t + u_t, \tag{8}$$

where $\mathsf{E}_t \left[\Delta p_{t+1} | \mathcal{I}_t\right]$ denotes expected inflation one-period ahead given today's information, denoted $\mathcal{I}_t$, and the real marginal costs facing firms, denoted s_t, where lower case letters are in logs. The anticipated signs of the coefficients from the theory model are shown: see Galí and Gertler (1999), Galí, Gertler, and Lopez-Salido (2001), and Castle, Doornik, Hendry, and Nymoen (2014).

To make (8) operational, the expectations are replaced by actual future inflation plus an error:

$$\mathsf{E}_t \left[\Delta p_{t+1} \mid \mathcal{I}_t\right] = \Delta p_{t+1} + \nu_{t+1}. \tag{9}$$

Taking expectations on both sides of (9):

$$\mathsf{E}_t \left[\Delta p_{t+1} | \mathcal{I}_t\right] = \mathsf{E}_t \left[\Delta p_{t+1} | \mathcal{I}_t\right] + \mathsf{E}_t[\nu_{t+1} \mid \mathcal{I}_t], \tag{10}$$

so $\mathsf{E}_t[\nu_{t+1} \mid \mathcal{I}_t] = 0$, and hence ν_{t+1} must be unpredictable from available information. Then, substituting (9) into (8):

$$\Delta p_t = \gamma_1 \Delta p_{t+1} + \gamma_2 \Delta p_{t-1} + \gamma_3 s_t + \epsilon_t, \tag{11}$$

where $\epsilon_t \sim \mathsf{D}[0, \sigma_\epsilon^2]$. Although it is unpredictable from $\mathcal{I}_t$, the error ν_{t+1} in (9) is not independent of Δp_{t+1}, so neither is ϵ_t in (11), and hence instrumental variables estimation is required, based on a set of valid exogenous and predetermined variables $\mathbf{z}_t$.

Lacking accurate data on aggregate real marginal costs, we use real unit labour costs, $c_t = (w - p - g + l)_t$, for s_t, where w, g, l are respectively the wage bill, GDP, and employment. s_t is also equal to the wage share and, being contemporaneous, treated as endogenous. Appendix B reports the data measures. Figure 8 shows the historical annual time series for Δp_t, $c_t = (w - p - g + l)_t$, $\Delta(g - l)_t$, measuring changes in labour productivity (output per person per year), and the long-term bond rate $R_{L,t}$.

The estimates of (11) use as additional instruments c_{t-1}, c_{t-2}, $\Delta(g - l)_{t-1}$, $\Delta(g - l)_{t-2}$, $R_{L,t-1}$, and $R_{L,t-2}$, which delivers:

$$\begin{aligned} \widehat{\Delta p_t} = \underset{(0.08)}{0.63} \widehat{\Delta p}_{t+1} + \underset{(0.06)}{0.025} \widehat{c}_t + \underset{(0.08)}{0.51} \Delta p_{t-1} - \underset{(0.07)}{0.11} \Delta p_{t-2} - \underset{(0.06)}{0.02} \qquad \text{(U1)} \\ \widehat{\sigma} = 2.89\% \;\; \mathsf{F}_{\mathsf{ar}}(2, 141) = 1021^{**} \;\; \mathsf{F}_{\mathsf{Het}}(8, 139) = 5.61^{**} \;\; \chi^2_{\mathsf{nd}}(2) = 69^{**} \; \chi^2_{\mathsf{Sar}}(4) = 9.59 \end{aligned}$$

Estimation is over 1866–2013 ($T = 148$), F_{ar} is the test for up to second order residual autocorrelation, and $\chi^2_{\mathsf{Sar}}(k)$ is a test of the validity of the instruments (see Sargan, 1964), with a p-value of 4.8% here.

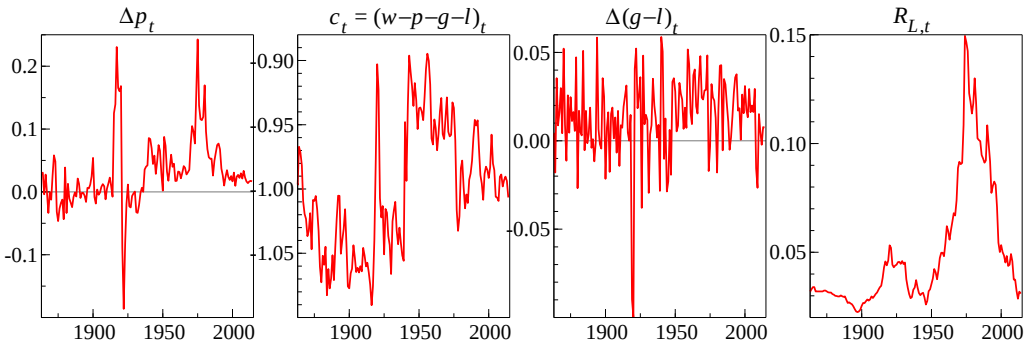

Figure 8: From left to right: UK annual inflation, real unit labour costs (in logs), changes in output per person per year, long-term bond rate.

The signs of coefficients are as anticipated from (8), although the coefficients of the inflation variables reveal considerable inertia and add to more than unity, and the coefficient of c_t is insignificant. Both the normality and heteroskedasticity tests strongly reject, which reveals that the residuals have those problems, but there are many possible sources thereof, including unmodelled location shifts given the turbulence in the middle of the 20th century. Despite the very different sample period and data frequency, such estimates are similar to those reported using more recent quarterly data and for other countries.

To check for robustness to possible location shifts, SIS was applied at 1%, fixing the three regressors in (U1). Then 22 significant step indicators, denoted $\{S_{i,t}\}$, are selected:

$$\widehat{\Delta p_t} = \underset{(0.05)}{0.50}\,\widehat{\Delta p}_{t+1} + \underset{(0.09)}{0.40}\,\widehat{c}_t + \underset{(0.05)}{0.35}\,\Delta p_{t-1} - \underset{(0.05)}{0.04}\,\Delta p_{t-2} - \underset{(0.09)}{0.40} + \{S_{i,t}\} \tag{U2}$$

$$\widehat{\sigma} = 1.67\%\ \ \mathsf{F}_{\mathsf{ar}}(2,120) = 3.3^{*}\ \ \mathsf{F}_{\mathsf{het}}(18,118) = 1.1\ \ \chi^2_{\mathsf{nd}}(2) = 4.4\ \ \chi^2_{\mathsf{Sar}}(15) = 31^{**}$$

Now $\widehat{\gamma}_1$ is smaller, c_t is highly significant, and inflation inertia has fallen. The misspecification tests are also insignificant. The rejection at 1% on $\chi^2_{\mathsf{Sar}}(15)$ is largely alleviated by removing the insignificant Δp_{t-2}. The important determinants of inflation found in Hendry (2015, Ch. 6), reduce $\widehat{\sigma}$ to 1.1%.

As Castle, Doornik, Hendry, and Nymoen (2014) report for their analysis of NKPC models fitted to more recent quarterly data and using IIS, the apparent significance of Δp_{t+1} in equations like (U1) appears to be an artefact due to the future value acting as a proxy for the unmodelled shifts. It would have taken remarkable prescience for anyone during 1914 to have anticipated the dramatically higher inflation of 1915, or even in 1916 anticipating a doubling during 1917, as well as foreseeing the introduction of price controls restraining inflation during the Second World War and their later removal, not to mention the Oil Crises of the 1970s.

5.3 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.

	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

Table 6: LMS estimates from Hettmansperger and Sheather. Our LTS(0.5) estimates with standard errors in parentheses.

Hettmansperger and Sheather (1992) took data from Mason, Gunst, and Hess (1989, p. 529), aimed to predict 'engine *knock*' from a constant and four regressors called '*spark* timing', '*Air*/fuel ratio', '*intake* temperature', and '*exhaust* temperature', where italic denotes the legend below. There are 16 observations on each. However, on inputting the data, they had inadvertently entered the second observation for *Air* as 15.1 rather than the correct 14.1 (denoted *WAir*, for wrong air), 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 it 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.

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 fixed, so all estimated models have an intercept. For correct air IIS finds four outliers, retaining *Air* and *intake*:

$$\widehat{knock}_i = \underset{(0.3)}{3.2}\, Air_i + \underset{(0.09)}{0.35}\, intake_i + \underset{(0.5)}{6.5}\ \mathbf{1}_{\{5\}} + \underset{(0.5)}{1.6}\ \mathbf{1}_{\{9\}} + \underset{(0.5)}{3.0}\ \mathbf{1}_{\{13\}} + \underset{(0.5)}{3.4}\ \mathbf{1}_{\{15\}} + \underset{(2.4)}{28.4} \tag{E1}$$

$$\widehat{\sigma} = 0.5\ \mathsf{R}^2_d = 0.99\ \ \mathsf{F}_{\mathsf{Het}}(4,7) = 0.6\ \ \chi^2_{\mathsf{nd}}(2) = 2.8\ \ \mathsf{F}_{\mathsf{Reset}}(2,7) = 5.0^*\ \mathsf{F}_{\mathsf{nl}}(2,7) = 0.3$$

where the indicator $\mathbf{1}_{\{5\}}$ reveals that observation 5 is selected as an outlier.[4] None of the diagnostic tests is significant at 5%, except for RESET which has a p-value of 4.5%. For 20 candidates, using (2): $20\alpha = 1$, so 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:

$$\widehat{knock}_i = \underset{(0.6)}{2.1}\, WAir_i + \underset{(0.2)}{0.9}\, intake_i + \underset{(1.5)}{6.3}\ \mathbf{1}_{\{5\}} + \underset{(6.8)}{27.3} \tag{E2}$$

$$\widehat{\sigma} = 1.4\ \ \mathsf{R}^2_d = 0.90\ \ \mathsf{F}_{\mathsf{Het}}(4,10) = 1.4\ \ \chi^2_{\mathsf{nd}}(2) = 0.2\ \ \mathsf{F}_{\mathsf{Reset}}(2,10) = 1.6\ \ \mathsf{F}_{\mathsf{nl}}(6,6) = 1.5$$

The two regression estimates are now similar, and both detect that observation 5 is an outlier. Figure 9 records actual and fitted values by OLS and IIS for the two measures of '*Air*' showing their closeness for the former. 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 $\mathbf{1}_{\{9\}}$.

[4] See footnote 7 for details of regression output and tests, with F_{nl} a test for non-linearity (see Castle and Hendry, 2010).

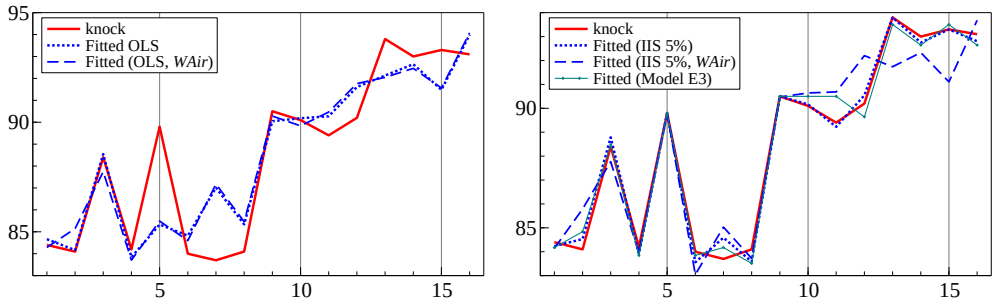

Figure 9: Actual and fitted values from OLS (left), and IIS (right) (E1) and (E2) for the two measures of '*Air*'

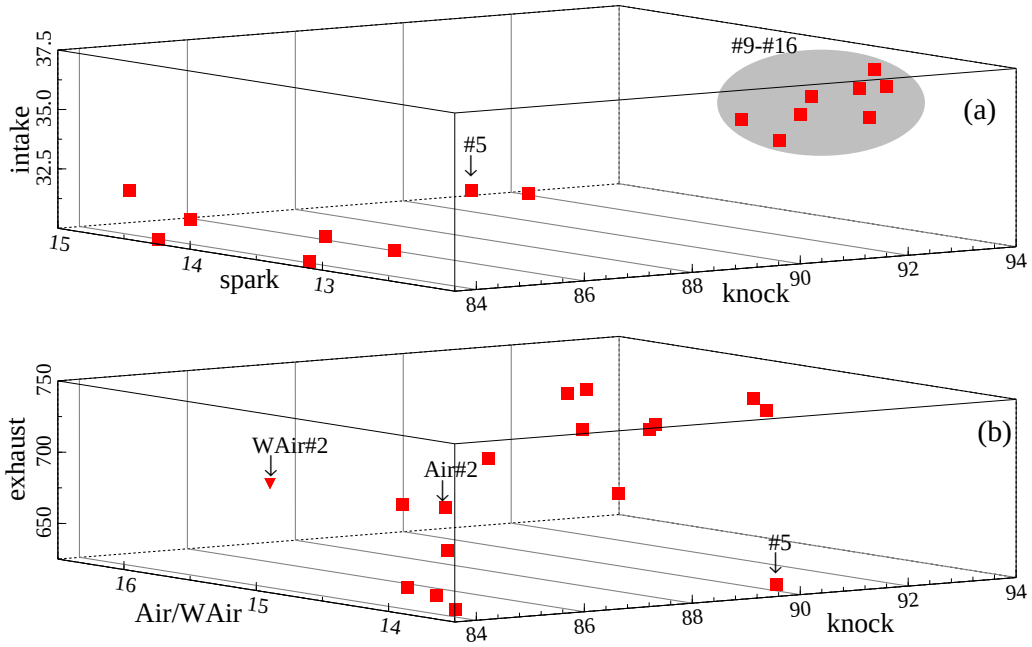

Figure 10: 3D plot of the data (a) *knock* against *spark* & *intake*; (b) *knock* against *exhaust* & *Air/WAir*.

5.3.1 Unweaving the findings

What actually caused the instability in LMS and LTS, and has IIS resolved it? The upper 3D plot in Figure 10(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 Figure 10(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' apparent in the upper graph (Figure 10a) 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

	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

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

'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 Figure 10a. $S_{\{i\geq 9\}}$ is interacted with the four regressors other than *exhaust*. The general model then comprises the fixed 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 with opposing sign. Imposing this simplification we find:

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

$$\widehat{\sigma} = 0.54\ \ \mathsf{R}^2_d = 0.98\ \ \mathsf{F}_{\mathsf{Het}}(4,10) = 0.7\ \ \chi^2_{\mathsf{nd}}(2) = 2.9\ \ \mathsf{F}_{\mathsf{Reset}}(2,10) = 1.6\ \ \mathsf{F}_{\mathsf{nl}}(6,6) = 2.4$$

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 Figure 10(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, 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.

5.3.2 Valid conditioning

We next test for valid conditioning using the method described in §3.4. Modelling *spark* by *Air*, *exhaust*, *intake* and a fixed 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 which 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 $\mathsf{F}_{\mathsf{valcond}}(2,10) = 0.46$, 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 $\mathsf{F}_{\mathsf{Het}}$ and $\mathsf{F}_{\mathsf{Reset}}$. Fixing all the regressors without selection when redoing IIS at 1% yields:

$$\begin{aligned}\widehat{knock}_i = &\underset{(0.23)}{3.2}\ WAir_i S_{\{i\leq 8\}} + \underset{(0.26)}{4.1}\ spark_i S_{\{i\geq 9\}} + \underset{(0.42)}{6.4}\ \mathbf{1}_{\{5\}} + \underset{(3.2)}{40.2} - \underset{(0.47)}{4.0}\ \mathbf{1}_{\{2\}} - \underset{(0.41)}{1.3}\ \mathbf{1}_{\{11\}} \\ \widehat{\sigma} = 0.37\ \ \mathsf{R}^2_d = 0.99\ \ &\mathsf{F}_{\mathsf{Het}}(4,8) = 0.37\ \ \chi^2_{\mathsf{nd}}(2) = 2.6\ \ \mathsf{F}_{\mathsf{Reset}}(2,8) = 0.7 \end{aligned} \tag{E4}$$

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 $\mathsf{F}_{\mathsf{valcond}}(2,8) = 0.33$, 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. There appear to be 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.

5.3.3 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 C confirms this problem in simulation experiments. Reducing this set with *Autometrics* at 2.5% finds the following model:

$$\begin{aligned}\widehat{knock}_i = &\underset{(0.28)}{3.2}\ Air_i + \underset{(0.09)}{0.35}\ intake_i + \underset{(0.52)}{6.5}\ \mathbf{1}_{\{5\}} - \underset{(0.54)}{1.6}\ \mathbf{1}_{\{9\}} - \underset{(0.51)}{3.0}\ \mathbf{1}_{\{13\}} - \underset{(0.52)}{3.4}\ \mathbf{1}_{\{15\}} + \underset{(2.4)}{28.4} \\ \widehat{\sigma} = 0.47\ \ \mathsf{R}^2_d = 0.99\ \ &\mathsf{F}_{\mathsf{Het}}(4,8) = 0.63\ \ \chi^2_{\mathsf{nd}}(2) = 2.8\ \ \mathsf{F}_{\mathsf{Reset}}(2,8) = 5.9^* \end{aligned} \tag{E5}$$

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.

The appendix suggests that the *autoLasso* provides a better approach to adaLasso estimation of models that have more variables than observation. The first step applies the block learning algorithm (as used in *Autometrics* for IIS) 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 give 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 seek a general notion of robustness in model selection which requires that methods used to select models have acceptable

performance when there are possible outliers and shifts leading to an incorrect distributional shape, omitted variables, misspecified dynamics, non-linearity, and non-stationarity, as well as 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 misspecification to a more general sense of robust model discovery. As a consequence, to tackle all these forms of potential misspecification jointly, general methods are needed, while at the same time retaining relevant subject matter insights. Hendry and Doornik (2014) call this empirical model discovery and theory evaluation. In this review paper, we describe that approach to achieving robustness against the seven potential problems just noted, all of which are empirically testable: see Hendry (1995).

The paper outlines that model selection approach, and the role of indicator saturation methods therein as designed to match the likely problem. A range of empirical examples demonstrates how the approach delivers robust selection and, hence, viable inference. Robustness can only be achieved if all modelling 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 misspecification 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.

References

Beringuer-Rico, V., S. Johansen, and B. Nielsen (2019). Models where the least trimmed squares and least median of squares estimators are maximum likelihood. Working paper 2019w05, Nuffield College.

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

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

Castle, J. L., J. A. Doornik, D. F. Hendry, and R. Nymoen (2014). Mis-specification testing: Non-invariance of expectations models of inflation. *Econometric Reviews 33*, 553–574. DOI:10.1080/07474938.2013.825137.

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

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

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

Castle, J. L. and D. F. Hendry (2011). Automatic selection of non-linear models. In L. Wang, H. Garnier, and T. Jackman (Eds.), *System Identification, Environmental Modelling and Control*, pp. 229–250. New York: Springer.

Castle, J. L. and D. F. Hendry (2014). Model selection in under-specified equations with breaks. *Journal of Econometrics 178*, 286–293.

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

Doob, J. L. (1953). *Stochastic Processes*. New York: John Wiley Classics Library. 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. See 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. and H. Hansen (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics 70*, 927–939.

Doornik, J. A. and D. F. Hendry (2015). Statistical model selection with big data. *Cogent Economics and Finance*, DOI:10.1080/23322039.2015.1045216. http://www.tandfonline.com/doi/full/10.1080/23322039.2015.1045216#.VYE5bUYsAsQ.

Doornik, J. A. and D. F. Hendry (2018). *Empirical Econometric Modelling using PcGive: Volume I.* (8th ed.). London: Timberlake Consultants Press.

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

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica 51*, 277–304. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000; in Ericsson, N. R. and Irons, J. S. (eds.) *Testing Exogeneity*, Oxford: Oxford University Press, 1994; and in Campos, J., Ericsson, N.R. and Hendry, D.F. (eds.), *General to Specific Modelling*. Edward Elgar, 2005.

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

Galí, J. and M. Gertler (1999). Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics 44*, 195–222.

Galí, J., M. Gertler, and J. D. Lopez-Salido (2001). European inflation dynamics. *European Economic Review 45*, 1237–1270.

Harding, S. G. (Ed.) (1976). *Can Theories be Refuted?* Dordrecht, Holland: D. Reidel Publishing Company.

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

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

Hendry, D. F. (2015). *Introductory Macro-econometrics: A New Approach*. London: Timberlake Consultants. http://www.timberlake.co.uk/macroeconometrics.html.

Hendry, D. F. (2018). Deciding between alternative approaches in macroeconomics. *International Journal of Forecasting 34*, 119–135, with 'Response to the Discussants', 142–146.

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

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

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

Hendry, D. F. and H.-M. Krolzig (2004). Resolving three 'intractable' problems using a Gets approach. Unpublished paper, Economics Department, University of Oxford.

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

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

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

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

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

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

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

Kitov, O. I. and M. N. Tabor (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.

Kuh, E., D. A. Belsley, and R. E. Welsh (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley.

Lakatos, I. (1974). Falsification and the methodology of scientific research programmes. In I. Lakatos and A. Musgrave (Eds.), *Criticism and the Growth of Knowledge*, pp. 91–196. Cambridge: Cambridge University Press.

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

Mayo, D. G. (Ed.) (2018). *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.

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

Peña, D. (2019). Detecting outliers and influential and sensitive observations in linear regression.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Popper, K. R. (1963). *Conjectures and Refutations*. New York: Basic Books.

Pretis, F., J. J. Reade, and G. Sucarrat (2018). Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. *Journal of Statistical Software 68, 4*, https://www.jstatsoft.org/article/view/v086i03.

Pretis, F., L. Schneider, J. E. Smerdon, and D. F. Hendry (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.

Sargan, J. D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology (with discussion). In P. E. Hart, G. Mills, and J. K. Whitaker (Eds.), *Econometric Analysis for National Economic Planning*, Volume 16 of *Colston Papers*, pp. 25–63. London: Butterworth Co. Reprinted as pp. 275–314 in Hendry D. F. and Wallis K. F. (eds.) (1984). *Econometrics and Quantitative Economics*. Oxford: Basil Blackwell, and as pp. 124–169 in Sargan J. D. (1988), *Contributions to Econometrics*, Vol. 1, Cambridge: Cambridge University Press.

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., F. Pretis, A. Powell-Smith, and B. Goldacre (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 21*, 1418–1429.

A Data definitions for the Boston Housing example

The Boston Housing Market data are available at `lib.stat.cmu.edu/datasets/boston` with some transformations used in the table on pp. 244–261 of Kuh, Belsley, and Welsh (1980). The data consists of variables and 506 observations. To standardize estimated coefficient values, Zone, Tax and BlkPop (and the corresponding Boston subsample variables) were all rescaled by 100.

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	$1000(\text{Bk} - 0.63)^2$ where Bk is the proportion of black persons by town
LowStat	% lower status of the population.

B Data definitions for the UK inflation example

Sources for the annual UK inflation data are reported in `www.timberlake.co.uk/macroeconometrics.html`.

Name	Variable
P_t	implicit deflator of GDP, (1860=1)
G_t	real GDP, £ million, 1985 prices
W_t	average weekly wage earnings index, (1860=1)
L_t	employment count (thousands)
$R_{L,t}$	long-term bond interest rate
Δx_t	$(x_t - x_{t-1})$ for any variable x_t

C Estimation of short data models

The aim is to extend the estimation of regression models to the situation where there are more variables than observations. At the same time we assume sparsity: not all variables matter, and we wish to select those that do. We assume that the final models are sufficiently small to be estimated using standard regression methods. In practice, we wish to start using our method already when the number of candidate variables, N, gets close to the sample size T. 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 note 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 modelling developing economies, or allowing for many effects with generous lag lengths (our main interest is in models for time series). They also naturally arise with saturation estimators, e.g. when adding an impulse indicator ('dummy') for every observation. The full model can then not be estimated. However, feasible estimates exist if combined with selection of indicators. This is called impulse indicator saturation (IIS).

When selection is jointly over indicators and variables, it is more convenient to treat them in the same way. Unfortunately, the standard split sample algorithm can no longer be used. An alternative is proposed below. The candidate set is still partitioned in blocks; 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 where we also select over variables. Hendry and Krolzig (2005) and Hendry and Krolzig (2004) propose algorithms for short data, but provide no evidence on their practical performance.

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 $256T^2$.

C.1 A learning algorithm for short data

C.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 of $\mathcal{X}$ is retained in the model, say $\mathcal{X}_F$, then the presence of $\mathcal{X}_F$ is fixed, but not their coefficients. Write $\overline{\mathcal{X}}$ for the free variables, i.e. $\mathcal{X}$ with $\mathcal{X}_F$ removed: $\overline{\mathcal{X}} = \mathcal{X}/\mathcal{X}^F$. M is used to select (the target $\mathbf{Y}$ is fixed throughout):

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

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

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:

$$\overline{\mathcal{X}} = \overline{\mathcal{X}}_1 \cup ... \cup \overline{\mathcal{X}}_B, \tag{12}$$

and apply model selection to each block:

$$\mathcal{S} = \cup_{i=1}^{B} M(\overline{\mathcal{X}}_i \mid \mathcal{X}^F). \tag{13}$$

$\mathbf{E}_1(\overline{\mathcal{X}} \mid \mathcal{C};\ \alpha, N^B)$:
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 $\mathcal{S} = \mathbf{E}_1(\overline{\mathcal{X}} \mid \mathcal{C};\ \alpha, N^B)$.
2. If $\dim\mathcal{S} < N^{\min}$ reselect $\mathcal{S} = \mathbf{E}_1(\overline{\mathcal{X}} \mid \mathcal{C};\ f(\alpha, \lambda), N^B)$;
   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 8: Algorithms for the expansion step

$\mathbf{L}(\mathcal{X} \mid \mathcal{C};\ \widetilde{\mathcal{C}}, \alpha_e, \alpha_r, N^{\min}, N^B, N^{\max}, j^{\max}, \lambda)$:
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 $\dim\mathcal{C}^{(j+1)} \geq N^{\max}$: finish with $\mathcal{C}^{(j+1)}, \widetilde{\mathcal{C}}^{(j+1)}$,
   else increment j and return to *expansion*. ■

Table 9: Learning from expansion and reduction

Now (13) enables 'short data' estimation with $N \geq T$, provided we keep $\dim(\overline{\mathcal{X}}_i \cup \mathcal{X}^F) < \eta T$, which allows for the use of a standard estimation method ($0 < \eta < 1$).

An algorithm for estimating short data models can be built upon (12) and (13). Some form of iteration will be needed, e.g. when the most important variable is in the final block, all previous block estimates are effectively voided. We will propose an algorithm that learns from previous rounds of estimation, and investigate its properties through some Monte Carlo experiments.

In short, 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.

C.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 8 is a procedure $\mathbf{E}_1$ that splits the free candidates $\overline{\mathcal{X}}$ into blocks and selects given the currently fixed 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 $\mathbf{E}_1$. The size of the selected set is limited: we reselect, and then, if still too large, select the N^B most significant variables. The significance is based on the estimates in each block. On the other hand, if the selection is too small, an optional *boost* of size λ is applied (only if $N^{\min} > 0$) and selection repeated with a more relaxed significance level.[5] 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.

C.1.3 Expansion and reduction

The 'learning' algorithm 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 given in Table 8. As

[5] $f(\alpha, s) = s\alpha$ if $0.94 \leq 1 + \alpha(s-1) \leq 1.06$; $f(\alpha, s) = s\alpha[1 + \alpha(s-1)]^{-1}$ otherwise.

A. $[\mathcal{C}^A, \widetilde{\mathcal{C}}^A] = \mathbf{L}(\mathcal{X} \mid \emptyset; \quad \emptyset, \quad \alpha, \alpha, N^{\min} = \min(N^B/8, 8), N^B, N^{\max}, 1, \lambda)$;
B. $[\mathcal{C}^B, \widetilde{\mathcal{C}}^B] = \mathbf{L}(\mathcal{X} \mid \mathcal{C}^A; \widetilde{\mathcal{C}}^A, \alpha, \alpha, N^{\min} = \min(N^B/8, 8)^*, N^B, N^{\max}, 10, \lambda)$;
C. $[\mathcal{C}^C, \widetilde{\mathcal{C}}^C] = \mathbf{L}(\mathcal{X} \mid \mathcal{C}^B; \mathcal{C}^B, f(\alpha, 2), f(\alpha, 1/2), 0, N^B, N^{\max}, 1, \lambda)$;
D. $[\mathcal{C}^D, \widetilde{\mathcal{C}}^D] = \mathbf{L}(\mathcal{X} \mid \mathcal{C}^C; \widetilde{\mathcal{C}}^C, \alpha, \alpha, 0, N^B, N^{\max}, 10, \lambda)$.

Table 10: Block search algorithm with learning

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 9.

A separate significance level is specified for expansion and for reduction, α_e and α_r, respectively, but they are usually the same. A maximum number of iterations is set through $j^{\max}$. Note that, in general, the outcome will be sensitive to the ordering of the variables. To facilitate replication, we always sort the expansion step by database index of the variable, and within that by lag length. 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.

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. Then stage C is another single iteration with a small boost (unless stage B hits the iteration upper limit), as the expansion significance level is approximately doubled, offset in the reduction. The history at that point is reset to the current model. Finally, stage D continues for up to ten iterations. This leads to our block search algorithm with learning, see Table 10.

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

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

where $\delta = 0.8$ by default, leaving space for the largest expansion set based on the default N^B. N^F is the variables that are fixed 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. In the last case, the boost at the start of C is also omitted. The boost at the start of B is only applied in the first iteration, provided stage A did not receive a boost.

There are a few additional adjustments that are specific to using *Autometrics* for the selection procedure M: diagnostic testing is switched off in expansion of stages C and D; the maximum number of terminals for expansion is set to one in stage A; backtesting is switched off in stages A and B, as there is no clear GUM to test against. Diagnostic testing is postponed if the model from the reduction step passes the tests. These choices make the procedure faster, at the cost of increased complexity. Finally, there is no lag presearch, and the model obtained by *Autometrics* is the union of the terminal models that it found: there does not seem to be a reason to select the best (penalized) fitting at this stage. The final union $\widetilde{\mathcal{C}}^D$ is the input to a normal selection using the default *Autometrics* settings. However, because the procedure is somewhat overgauged from backtesting and boosts, a better gauge is achieved if this final step is run without backtesting.

C.2 Experiments with independence

Monte Carlo experiments are used to compare the power of the algorithm described above:

$$\text{DGP:L } y_t^L = \mu + \gamma\left(I_{\tau T+1} + ... + I_T\right) + u_t, \qquad u_t \sim \mathsf{N}(0,1), \tag{14}$$

$$\text{DGP:S } y_t^S = \mu + \gamma\left(I_1 + I_{1+S} + I_{1+2S} + ...\right) + u_t, \qquad u_t \sim \mathsf{N}(0,1), \tag{15}$$

$$\text{DGP:Z } y_t^Z = \mu + \beta T^{-1/2}\left(z_{1t} + ... + z_{12,t}\right) + u_t, \qquad z_t, u_t \sim \mathsf{N}(0,1). \tag{16}$$

Setting $T = 100$ and $\tau = 0.8$ in DGP:L means that twenty percent of the sample is in the break period. Defining $S = \lfloor (1-\tau)^{-1} \rfloor$ then DGP:S with $T = 100$ and $\tau = 0.8$ has its mean shifted 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.

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

$$\text{MOD:L } y_t = \alpha^F + \alpha_1 I_1 + ... + \alpha_T I_T + \varepsilon_t, \tag{17}$$

$$\text{MOD:Z } y_t = \alpha^F + \sum_{i=1}^{m} \alpha_m y_{t-m} + \sum_{i=1}^{12}\sum_{i=0}^{m} \alpha_{im} z_{i,t-m} + \varepsilon_t. \tag{18}$$

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

Several other approaches that are feasible in this setting are included in the comparison:

1. *Stepwise regression* at significance level p_a;
2. *IIS algorithm* of Johansen and Nielsen (2009) at significance level p_a;
3. *Lasso* of Efron, Hastie, Johnstone, and Tibshirani (2004) 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;
4. *Backward elimination* in blocks, followed by *Autometrics* at significance level p_a;
5. *Standard Autometrics*: block search algorithm with learning at p_a (Table 10), followed by *Autometrics* at p_a;
6. *Reduced Autometrics*: block search algorithm with learning (Table 10) at p_a, followed by *Autometrics* at p_a but without backtesting.

Table 11 gives the gauge and potency for selected values of γ, i.e. the size 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 power. The Lasso does not use a significance level, and termination is based on an information criterion. The potency is low, except for larger γ, but then the gauge shoots up as well. The termination decision for Lasso is problematic in this design. An alternative is to use cross validation, but that does not work here either.

IIS and backward elimination are similar, with the former having best control of the size when there is no break, courtesy of the bias correction that is used. The proposed learning algorithm, as used by *Autometrics*, is close 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. When $\gamma = 0$, both the reduced and the standard version are overgauged. Seriously so for the latter, which is caused by the large initial boost at $p_a = 0.05$ (the reduced version has it switched off at 5%, but not for $p_a = 0.01$). The standard version retains too many insignificant variables that are kept in backtesting, and these are counted in the gauge.

Table 12 provides more insight by varying the length of the break. DGP:L is used with a break in mean of size four, but the duration of the break is for 2,10, and 20 observations respectively. The Lasso

	5% target DGP:L			1% target DGP:L			1% DGP:S	
	$\gamma=0$	$\gamma=2$	$\gamma=4$	$\gamma=0$	$\gamma=3$	$\gamma=4$	$\gamma=4$	$\gamma=5$
Stepwise regression								
Gauge %	15.5	10.5	14.6	1.4	0.1	0.0	0.0	0.0
Potency %	—	51.7	99.1	—	9.5	12.0	10.3	10.7
Lasso (BIC, $N_{max} = 50$)								
Gauge %				1.2	0.2	2.3	2.0	14.2
Potency %				—	6.1	16.5	14.5	77.9
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
Block learning with Autometrics, reduced								
Gauge %	9.0	4.8	8.2	1.6	0.2	0.4	0.5	0.5
Potency %	—	47.9	98.5	—	52.0	86.2	34.7	53.0
Block learning with Autometrics, standard								
Gauge %	20.0	8.3	9.2	4.0	0.4	0.5	0.6	0.5
Potency %	—	56.6	98.7	—	68.9	90.3	39.2	57.9

Table 11: DGP:L and DGP:S have a break in mean of size γ in 20% of observations. The estimated model is MOD:L, consisting of a constant and T dummies. $T = 100$ observations, $M = 1000$ replications, $p_a = 0.05, 0.01$.

	$\tau=0.02$	$\tau=0.1$	$\tau=0.2$	$\tau=0.02$	$\tau=0.1$	$\tau=0.2$
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	54.5	48.3	40.6
Potency %	85.5	79.4	16.0	100.0	99.9	99.7
Lasso (BIC, no N_{max})						
Gauge %	94.9	95.5	95.3			
Potency %	100.0	100.0	99.7			
Autometrics, reduced ($p_a = 0.01$)				*Autometrics, standard ($p_a = 0.01$)*		
Gauge %	0.8	0.7	0.4	1.3	1.0	0.5
Potency %	92.6	90.1	86.2	95.6	93.1	90.3

Table 12: DGP:L with break in last τT observations of size $\gamma = 4$. The estimated model is MOD:L, consisting of a constant and T dummies. $T = 100$ observations, $M = 1000$ replications.

and stepwise regression both have the gauge falling 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.

Table 13 looks at short data settings involving variables. In the first set $\beta = 0$, 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 reduced version of *Autometrics*. The standard version is somewhat overgauged, more so when the correct model is empty, because then the initial boost is likely to be used. Lasso selects models that are much too large when $T = 40$, which will to a large extent be the consequence of using an information criterion. It also has more difference between $\beta = 0$ and $\beta = 10$.

T	m	α=0.05	0.025	0.01	0.001	α=0.05	0.025	0.01	0.001
Autometrics reduced				$\beta = 0$				$\beta = 10$	
40	4	0.080	0.043	0.026	0.0056	0.077	0.035	0.012	0.0024
100	8	0.042	0.023	0.014	0.0026	0.044	0.022	0.009	0.0013
250	20	0.032	0.018	0.009	0.0014	0.033	0.018	0.009	0.0006
Autometrics standard									
40	4	0.104	0.060	0.032	0.0057	0.087	0.038	0.013	0.0028
100	8	0.087	0.060	0.027	0.0028	0.065	0.041	0.018	0.0015
250	20	0.087	0.066	0.030	0.0017	0.066	0.056	0.029	0.0010
Lasso (BIC, $N_{max} = 50$)				$\beta = 0$			$\beta = 10$		
40	4			0.459			0.514		
100	8			0.010			0.046		
250	20			0.004			0.012		
Lasso (CV, 5 fold)				$\beta = 0$			$\beta = 10$		
40	4			0.518			0.447		
100	8			0.482			0.404		
250	20			0.205			0.154		

Table 13: Gauge of the *Autometrics* algorithm. $T = 100$ observations, $M = 1000$ replications, $M = 10\,000$ for $p_a = 0.001$. DGP:Z with MOD:Z.

C.3 Experiments with correlation

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:

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

HP7 has $R^2 = 0.58$, and HP8 has $R^2 = 0.93$; all coefficients have very high t-values (in excess of 8). We create versions that have more variables than observations by adding 10 $\mathsf{IIN}(0,1)$ 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 labelled HP7big and HP8big in Table 14.

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.

C.4 Lasso for short data

The experiments in this paper focus on selection. 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, moreover it defeats the purpose of machine learning to have to select the model by visual inspection of a plot of the criterion.

Modelling with more variables than observations is practically relevant, and 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, we could use ridge estimates, but instead we propose to run selection in blocks, collecting a set of regressors for

	HP7	HP8	HP7big	HP8big	HP7	HP8	HP7big	HP8big
Stepwise regression ($p_a = 0.01$)								
Gauge %	1.4	1.2	0.9	1.7				
Potency %	92.4	89.2	99.9	50.9				
Lasso (BIC, $N_{max} = 50$)					*Lasso (10-fold CV)*			
Gauge %	19.5	35.1	2.9	2.0	64.5	89.3	36.8	36.5
Potency %	94.4	86.3	71.8	58.0	99.7	99.9	99.3	76.5
adaLasso (BIC, $N_{max} = 50$)					*adaLasso (10-fold CV)*			
Gauge %	4.5	3.3	2.9	6.2	70.7	38.9	92.1	63.7
Potency %	99.7	100.0	72.2	98.9	99.1	98.0	94.3	92.3
Autometrics, reduced ($p_a = 0.01$)					*Autometrics, standard ($p_a = 0.01$)*			
Gauge %	1.6	1.6	0.7	0.9	1.6	1.6	1.3	2.2
Potency %	99.2	100.0	99.4	100.0	99.2	100.0	99.5	100.0

Table 14: $T = 139$, $M = 1000$, $p_a = 0.01$, 3 relevant variables. HP7 and HP8 have 37 irrelevant variables, the big versions have 141.

	HP7	HP8	HP7big	HP8big	HP7	HP8	HP7big	HP8big
adaLasso (BIC, $N_{max} = 50$)					*adaLasso (10-fold CV)*			
Gauge %	4.5	3.3	2.9	6.2	70.7	38.9	92.1	63.7
Potency %	99.7	100.0	72.2	98.9	99.1	98.0	94.3	92.3
blockLasso (BIC, $N_{max} = 50$)					*autoLasso (BIC, $p = 0.001$)*			
Gauge %	3.5	3.0	4.1	3.0	0.9	0.8	2.0	1.5
Potency %	100.0	100.0	99.6	100.0	99.3	100.0	99.6	100.0

Table 15: $T = 139$, $M = 1000$, $p_a = 0.01$, 3 relevant variables. HP7 and HP8 have 37 irrelevant variables, the big versions have 141.

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 10, 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 10, with adaLasso (BIC) as the selection device for the algorithm. *Autometrics* at p_a is used to select the final model.

In both cases the ordering in each block is based on significance in the OLS model using the selected variables.

Table 15 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 JEDC setting as discussed in Hendry and Doornik (2014, §17.2.2).

$$\text{DGP:J } y_t^J = \mu + \sum_{i=1}^{5} \beta_i T^{-1/2} z_{it} + u_t, \quad u_t \sim \mathsf{N}(0,1), (z_{1t}, ..., z_{Nt}) \sim \mathsf{N}(0, C_z), \tag{19}$$

$$\text{MOD:J } y_t = \alpha^F + \sum_{i=1}^{m} \alpha_m y_{t-m} + \sum_{i=1}^{N} \sum_{i=0}^{m} \alpha_{im} z_{i,t-m} + \varepsilon_t, \quad , t = 1, ..., 100, \tag{20}$$

where $C_z = (c_{i,j}) = \rho^{|i-j|}$ 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 the sample size is $T = 100$.

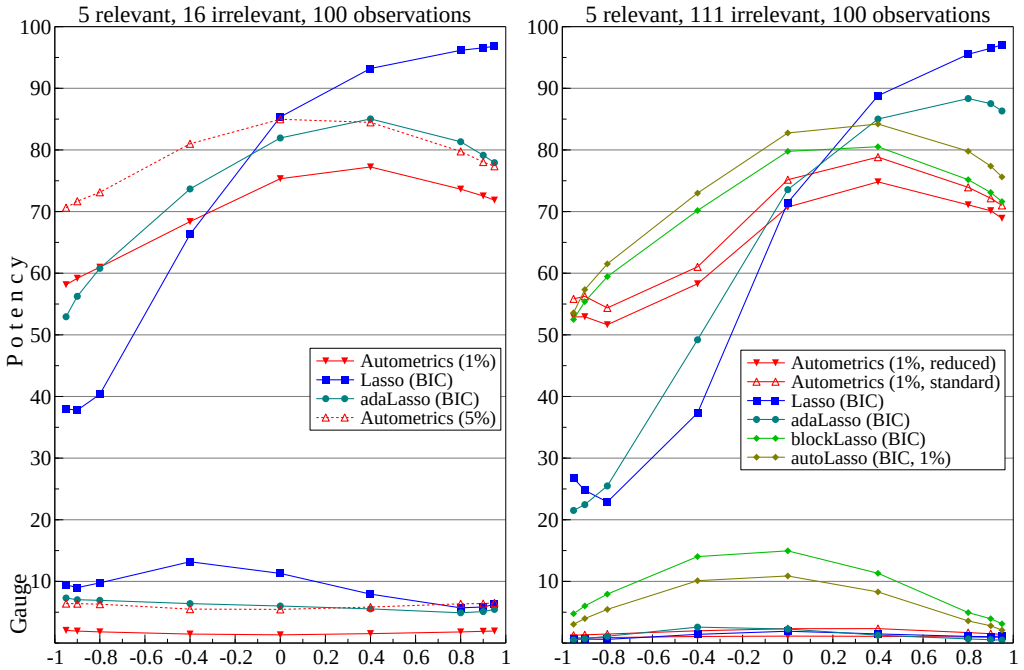

Figure 11: JEDC experiment, standard version on left, large version on right.

Figure 11 shows the result for values of $\rho = -0.95, -0.9, -0.8, 0.4, 0, 0.4, 0.8, 0.9, 0.95$. The left panel is the standard version ($m = 1, N = 10$), while on the left is the large version ($m = 8, N = 12$). Both gauge and potency are plotted, with gauge always at the bottom. For the standard experiment we see that the gauge of *Autometrics* is close to the target size. The adaLasso gauge is close to the 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: two variables need to enter jointly but do 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 overgauged.