William Becker*, Paolo Paruolo and Andrea Saltelli

# Variable Selection in Regression Models Using Global Sensitivity Analysis

https://doi.org/10.1515/jtse-2018-0025
Received August 31, 2018; accepted February 16, 2021

**Abstract:** Global sensitivity analysis is primarily used to investigate the effects of uncertainties in the input variables of physical models on the model output. This work investigates the use of global sensitivity analysis tools in the context of variable selection in regression models. Specifically, a global sensitivity measure is applied to a criterion of model fit, hence defining a ranking of regressors by importance; a testing sequence based on the 'Pantula-principle' is then applied to the corresponding nested submodels, obtaining a novel model-selection method. The approach is demonstrated on a growth regression case study, and on a number of simulation experiments, and it is found competitive with existing approaches to variable selection.

**Keywords:** model selection, Monte Carlo, sensitivity analysis, simulation

**JEL classification:** C52, C53

## 1 Introduction

Model selection in regression analysis is a central issue, both in theory and in practice. Related fields include multiple testing (Bittman et al. 2009; Romano and Wolf 2005), pre-testing (Leeb and Poetscher 2006), information criteria (Hjort and Claeskens 2003; Liu and Yang 2011), model selection based on Lasso (Brunea 2008), model averaging (Claeskens and Hjort 2003), stepwise regression, (Miller 2002), risk

Information and views set out in this paper are those of the authors and do not necessarily reflect the ones of the institutions of affiliation.

***Corresponding author: William Becker**, European Commission, Joint Research Centre, Ispra, VA, Italy, E-mail: william.becker@bluefoxdata.eu. https://orcid.org/0000-0002-6467-4472
**Paolo Paruolo,** European Commission, Joint Research Centre, Ispra, VA, Italy, E-mail: paolo.paruolo@ec.europa.eu. https://orcid.org/0000-0002-3982-4889
**Andrea Saltelli,** Open Evidence Research, Universitat Oberta de Catalunya, Barcelona, Spain, E-mail: andrea.saltelli@gmail.com. https://orcid.org/0000-0003-4222-6975

inflation in prediction, (Foster and George 1994), directed acyclic graphs and causality discovery (Freedman and Humphreys 1999).[1]

Model choice is also of primary concern in many areas of applied econometrics, as witnessed for example by the literature on growth regression (Sala-i-Martin 1997). Controlling for the right set of covariates is central in the analysis of policy impact evaluations; this is embodied in the assumption of unconfoundedness (Imbens and Wooldridge 2009). In economic forecasting, model selection is the main alternative to model averaging, (Hjort and Claeskens 2003).

The analysis of the effects of pretesting on parameter estimation has a long tradition in econometrics (Danilov and Magnus 2004) and in this context Magnus and Durbin (1999) and co-authors proposed the weighted average least squares estimator (WALS) and compared it with model averaging for growth empirics (Magnus, Powell, and Prufer 2010).

Model selection is a major area of investigation also in time-series econometrics (Phillips 1997, 2003). The so-called London School of Economics (LSE) methodology has played a prominent role in this area, advocating the *general-to-specific* (GETS) approach to model selection (Castle, Doornik, and Hendry 2011; Hendry and Krolzig 2005) and references therein. In a widely cited paper, Hoover and Perez (1999) (hereafter HP) 'mechanized'—i.e. translated—the GETS approach into an algorithm for model selection and they then tested the performance of the HP algorithm on a set of time-series regression experiments, constructed along the lines of Lovell (1983).

Model selection is also related to the issue of regression coefficients' robustness (i.e. lack of sensitivity) to the omission/inclusion of additional variables. Leamer (1983) proposed extreme bound analysis, i.e. to report the range of possible parameter estimates of the coefficient of interest when varying the additional regressors included in the analysis, as an application of sensitivity analysis to econometrics. Other applications of sensitivity analysis to econometrics include the

1 Model selection is also associated with current rules of thumb on the maximum number of regression parameters to consider. This literature appears to have been initiated by Freedman (1983), who considered the case of a first screening regression with 50 regressors and 100 data points, where regressors that are significant at 25% significance level are kept in a second regression. Freedman showed that the second regression is troublesome when one acts as if the screening regression had not been performed and the ratio of number of observations to number of regressors in the screening regression is kept in a fixed proportion as the number of observations diverges. This study was followed by Freedman and Pee (1989), Freedman, Pee, and Midthune (1992), who defined the rule of thumb that the ratio of the number of observations per regressor should be at least equal to 4; this rule is included in Harrell (2001), who suggested to have it at least equal to 10.

local sensitivity to model misspecification developed in Magnus and Vasnev (2007) and Magnus (2007).[2]

On the other hand, sensitivity analysis originated in the natural sciences, and is generally defined as 'the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be apportioned to different sources of uncertainty in its inputs', (Saltelli, Tarantola, and Campolongo 2000). The term *global* sensitivity analysis (GSA) is used to refer to sensitivity analysis approaches that fully explore the space of uncertainties, as opposed to 'local' methods which are only valid at a nominal point (Saltelli and Annoni 2010). The main tools used in GSA are based on a decomposition of the variance of the model output (Sobol' 1993).

Despite several uses of sensitivity in econometrics, the present authors are not aware of systematic applications of the techniques of Global Sensitivity Analysis to the problem of model selection in regression. With this in mind, the present paper explores the application of variance-based measures of sensitivity to model selection.

This paper aims to answer the question: "Can GSA methods help in model selection in practice?", rather than to propose a single algorithm with the aim to dominate all alternatives. To this purpose, a simple algorithm is considered as a representative of a novel GSA approach; the new algorithm is found to perform rather well when compared with alternatives. This shows how GSA methods can indeed bring a useful contribution to this field.

In particular, a widely-used measure in the GSA literature, called the 'total sensitivity index' is employed to rank regressors in terms of their importance in a regression model. The information on the ordering of regressors given by GSA methods appears to be somewhat complementary to the one based on $t$-ratios employed in the GETS approach; this suggests to consider viable ordering of regressors combining the two orderings. Based on these insights, a GSA algorithm is constructed which combines the two rankings.

The proposed GSA representative algorithm uses the ordering of the regressors via GSA or the $t$-ratios within a testing strategy based on the 'Pantula-principle', see Pantula (1989). For any ordering of the regressors, this amounts to a single sequence of tests against the full model, starting from the most restricted submodel

**2** They show that local sensitivity measures provide complementary information with respect to standard diagnostic tests for misspecification, i.e. that the two types of statistics are asymptotically independent. In SA a local measure of sensitivity is one focused on a precise point in the space of the input factor, e.g. a partial derivative of the output versus the input. With a global measure of sensitivity the influence of a given input on the output is averaged both on the distribution of the input factor itself and on the distributions of all the remaining factors, see Saltelli, Andres, and Homma (1993).

to the most unrestricted one.[3] This implies both a reduction in the number of tests for each given ordering (with an associated saving of computing times) and the favorable control of the size of the testing sequence. The present application of the 'Pantula-principle' appears novel in the context of model selection.

The GSA algorithm is tested here using several case studies. A detailed investigation of the performance of the GSA algorithm is first performed using the simulation experiments of HP, who defined a set of Data Generating Processes (DGPs) based on real economic data. Simulating these DGPs, one can record how often the algorithm recovers the variables that are included in the DGP. This is compared to the results of HP's GETS algorithm, as well as those of the Autometrics GETS package (Pretis, Reade, and Sucarrat 2018).

In order to further compare the GSA approach to a wider set of model selection procedures, the DGPs in Deckers and Hanck (2014) are also considered; this allows a direct comparison with a number of procedures. Finally, the algorithm is applied to a growth regression case study which is also taken from the same paper.

Overall, results point to the possible usefulness of GSA methods in model selection algorithms. When comparing the optimized GSA and HP algorithms, the GSA method appears to be able to reduce the failure rate in recovering the underlying data generating process from 5 to 1% approximately—a fivefold reduction. When some of the regressors are weak, the recovery of the exact DGP does not appear to be improved by the use of GSA methods.

Comparing the GSA algorithm to a wider set of approaches considered in DH, the results are competitive with alternatives, in the sense that the GSA algorithm is not dominated by alternative algorithms in the Monte Carlo (MC) simulations. In the empirical application on growth regression, not surprisingly, it identifies similar variables to those found by other methods. While these results do not prove the GSA approach to dominate other existing approaches, they show that the GSA approach is not dominated by any single alternative, and that it has the potential to contribute to improve existing algorithms; the present study can hopefully hence pave the way for future advances.

The rest of the paper is organized as follows. Section 2 defines the problem of interest and introduce GSA and variance-based measures. Section 3 presents some theoretical properties of orderings based on the total sensitivity index, while Section 4 presents the GSA algorithm. Results are reported in Sections 5 and 6, where the former is a detailed investigation on datasets generated following the paper of Hoover and Perez (1999), and the latter is a comparison with a wide range of model selection procedures on simulated data sets and on a growth regression, following Deckers and Hanck (2014). Section 7 concludes. Three appendices report

3 This sequence can still be interpreted as compliant to the GETS principle.

proofs of the propositions in the paper, details on the DGP design in HP and a discussion about the identifiability of DGPs. Finally, this paper follows the notational conventions in Abadir and Magnus (2002).

# 2 Model Selection and Global Sensitivity Analysis

This section presents the setup of the problem, and introduces global sensitivity analysis. The details of the proposed algorithm are deferred to Section 4.

## 2.1 Model Selection in Regression

Consider $n$ data points in a standard multiple regression model with $p$ regressors of the form

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \ldots\boldsymbol{X}_p\beta_p + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is $n \times 1$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is $n \times p$, $\boldsymbol{X}_i := (x_{i,1}, \ldots, x_{i,n})'$ is $n \times 1$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is $p \times 1$ and $\varepsilon$ is a $n \times 1$ random vector with distribution $N(0, \sigma^2\boldsymbol{I}_n)$. The symbol $'$ indicates transposition.

Equation (1) describes both the model and the DGP. In the model, the coefficients $\beta_i$ are parameters to be estimated given the observed data $\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X})$. Each DGP is described by Eq. (1) with $\beta_i$ set at some numerical values, here indicated as $\beta_{0i}$, collected in the vector $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})'$.

Some of the true $\beta_{0i}$ may be 0, corresponding to irrelevant regressors. Let $\mathbb{T} := \{i \in \mathbb{J} : \beta_{0,i} \neq 0\}$ be the set of all relevant regressor indices in the DGP, with $r_0$ elements, where $\mathbb{J} := \{1, \ldots, p\}$ indicates the set of the first $p$ integers. Let also $\mathbb{M} := \mathbb{J}\backslash\mathbb{T}$ indicate the set of all regressor indices for irrelevant regressors.[4] Equation (1) also formally nests dynamic specifications, as detailed in Appendix B below; in this case $X_i$ contain lagged dependent variables, and (1) is generated recursively.

Imposing the restriction $\beta_i = 0$ for some regressors $i$, one obtains a submodel[5] of model (1). Each submodel can be characterized by a set $a$, $a \subseteq \mathbb{J}$, containing the indices of the included regressors. For instance, $a = \{1, 5, 9\}$, indicates the submodel including regressors numbered 1, 5, 9. The model without any restriction on $\beta_i = 0$ is called the *general unrestricted model* (GUM).

4 Here $\mathbb{J}\backslash\mathbb{T}$ denotes the set difference $\mathbb{J}\backslash\mathbb{T} := \{i : i \in \mathbb{J}, i \notin \mathbb{T}\}$; sums over empty sets are understood to be equal to 0.

5 In the paper 'submodel' and 'specification' are used as synonyms.

Alternatively, the same information on submodel $a$ can be represented by a $p \times 1$ vector $\boldsymbol{\gamma}_a = (\gamma_1, \ldots, \gamma_p)'$, with $j$-th coordinate $\gamma_j$ with value 1 (respectively 0) that indicates the inclusion (respectively exclusion) of regressor $j$ from the specification, i.e. $\gamma_j = 1\,(j \in a)$ and $1(\cdot)$ is the indicator function.[6] The GUM corresponds to $\boldsymbol{\gamma}$ equal to $\boldsymbol{\iota}$, a vector with all 1s. $\boldsymbol{\gamma}_{\mathbb{T}}$ corresponds to the best selection of regressors, i.e. the same one of the DGP; in the following the notation $\boldsymbol{\gamma}_{\mathbb{T}} = \boldsymbol{\gamma}_0$ is also used.

Let $\Gamma$ be the set of all $p \times 1$ vectors of indicators $\gamma$, $\Gamma = \{0, 1\}^p$. Note that there are $2^p$ different specifications, i.e. $2^p$ possible $\boldsymbol{\gamma}$ vectors in $\Gamma$. When $p = 40$, as some experiments in Section 5, the number of specifications is $2^p \approx 1.0995 \cdot 10^{12}$, a very large number. This is why an exhaustive search of submodels is infeasible in many practical cases, and model selection techniques focus on a search over a limited set of submodels $\Gamma_s \subset \Gamma$.

Each submodel can be written as model (1) under the restriction

$$\boldsymbol{\beta} = \boldsymbol{H}_{\boldsymbol{\gamma}} \phi, \tag{2}$$

where $\boldsymbol{H}_{\boldsymbol{\gamma}}$ contains the columns of an identity matrix $\boldsymbol{I}_p$ corresponding to elements $\gamma_i$ equal to 1 within $\boldsymbol{\gamma}$. Specification (2) is referred to as the '$\boldsymbol{\gamma}$ submodel' in the following. Also the 'true' vector $\beta_0$ has representation $\boldsymbol{\beta}_0 = \boldsymbol{H}_0 \phi_0$ where $\boldsymbol{H}_0$ is a simplified notation for $\boldsymbol{H}_0 = \boldsymbol{H}_{\boldsymbol{\gamma}_{\mathbb{T}}} = \boldsymbol{H}_{\boldsymbol{\gamma}_0}$.

The least squares estimator of $\boldsymbol{\beta}$ in submodel $\boldsymbol{\gamma}$ can be written as

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \boldsymbol{H}_{\boldsymbol{\gamma}} \left(\boldsymbol{H}_{\boldsymbol{\gamma}}' \boldsymbol{X}'\boldsymbol{X} \boldsymbol{H}_{\boldsymbol{\gamma}}\right)^{-1} \boldsymbol{H}_{\boldsymbol{\gamma}}' \boldsymbol{X}'\boldsymbol{y}. \tag{3}$$

The problem of interest is to retrieve $\mathbb{T}$, or the corresponding $\boldsymbol{\gamma}_{\mathbb{T}}$, given the observed data $\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X})$, i.e. to identify the DGP.[7]

## 2.2 GSA Approach

General-to-specific (GETS) approaches such as the algorithm used by HP (described in detail in Section 5) use $t$-ratios to rank regressors in order of importance, which guides the selection of the set of submodels $\Gamma_s$. This study proposes instead to decompose the selection of models in two stages:

(i) define an ordering of regressors based on their importance;

---

**6** Similarly, the notation $a_{\boldsymbol{\gamma}} := \{i_1, \ldots, i_{k_\gamma}\}'$ is used to indicate the index set corresponding to some vector $\gamma$.

**7** All empirical models are assumed to contain the constant; this is imposed implicitly by de-meaning the $\boldsymbol{y}$ and $\boldsymbol{X}_{\boldsymbol{i}}$ vectors. Hence in the following, the 'empty set of regressors' refers to the regression model with only the constant.

(ii) use a sequence of $p$ tests that compare the GUM with submodels which contain the first $h$ most important regressors, starting from $h = 0, 1, 2, \ldots$ and ending at the first submodel $r$ that does not reject the null hypothesis.

In this paper the ordering in (i) based on the $t$-ratios is complemented with a variance-based measure of importance from GSA, called the 'total sensitivity index'. The proposed algorithm, called the 'GSA algorithm', combines this new ranking with the ranking by $t$-ratios.

The testing sequence is defined based on this new ranking; a 'bottom-up' selection process is adopted, which builds candidate models by adding regressors in descending order of importance. This 'bottom-up' selection process follows the 'Pantula principle' and has well defined theoretical properties, see e. g. (Paruolo 2001), and it can still be interpreted as a GETS procedure.

The total sensitivity index in GSA is based on systematic exploration of the space of the inputs to measure its influence on the system output, as is commonly practiced in mathematical modeling in natural sciences and engineering. It provides a *global* measure of the influence of each input to a system.[8] Reviews of global sensitivity analysis methods used therein are given in Saltelli et al. (2012), Norton (2015), Becker and Saltelli (2015), Wei, Lu, and Song (2015).[9] The total sensitivity index is a variance-based measures of sensitivity, which are the analogue of the analysis of the variance, see Archer, Saltelli, and Sobol (1997).[10]

Given the sample data $\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X})$, consider the $\boldsymbol{y}$ submodel, see eqs. (1), (2) and (3). Let $q(\boldsymbol{y})$ indicate the BIC of model fit of this submodel, $q(\boldsymbol{y}) = \log \hat{\sigma}^2_{\boldsymbol{y}} + k_{\boldsymbol{y}} c_n$, with $c_n := \log(n)/n$.[11] Remark that $q$ is a continuous random variable that depends on the discretely-valued $\boldsymbol{y}$. The idea is to apply the total sensitivity index using $q$ as output, with $\boldsymbol{y}$ as input. Although the BIC is used here as $q$, the measure of model fit, other consistent information criteria or the maximized log-likelihood could be used instead.

---

**8** The 'mechanistic' models in these disciplines are mostly principle-based, possibly involving the solution of some kind of (differential) equation or optimization problem, and the output—being the result of a deterministic calculation—does not customarily include an error term.

**9** Recent applications of these methods to the quality of composite indicators are given in Paruolo, Saltelli, and Saisana (2013) and Becker et al. (2017).

**10** Variance-based methods explore the entire distribution of each factor.

**11** $q$ can be taken to be any consistent information criterion where consistent information criteria replace log$n$ with some other increasing function $f(n)$ of $n$ with the property $c_n = f(n)/n \to 0$. Here the fact that $nc_n$ diverges is not used in the proofs. Note that $q(\boldsymbol{y})$ is a function of $\boldsymbol{Z}$, but this is not indicated in the notation for simplicity.

The objective is to capture both the main effect and the interaction effects of the input factors onto the output $q$, see Saltelli et al. (2012). The following section defines the total sensitivity index.

## 2.3 Sensitivity Measures

Let $\mathbb{E}$ indicate the empirical expectation over $\Gamma$, i.e. $\mathbb{E}(h(\boldsymbol{\gamma})) \coloneqq (\#(\Gamma))^{-1}\sum_{\boldsymbol{\gamma}\in\Gamma}(h(\boldsymbol{\gamma}))$, for any function $h$. Let also $\mathbb{V}$ indicate the variance operator associated with $\mathbb{E}$, $\mathbb{V}(h) \coloneqq \mathbb{E}(h^2) - (\mathbb{E}(h))^2$.

The $\boldsymbol{\gamma}$ vector is partitioned into two components $\gamma_i$ and $\boldsymbol{\gamma}_{-i}$, where $\boldsymbol{\gamma}_{-i}$ contains all elements in $\boldsymbol{\gamma}$ except $\gamma_i$. Let $\mathbb{E}(\cdot|b)$ and $\mathbb{V}(\cdot|b)$ (respectively $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$) indicate the conditional (respectively marginal) expectation and variance operators with respect to a partition $(a, b)$ of $\boldsymbol{\gamma}$, where $a$ and $b$ are taken equal to $\gamma_i$ and to $\boldsymbol{\gamma}_{-i}$.

Two commonly-accepted variance-based measures are reviewed here, the 'first-order sensitivity index' $S_i$, Sobol' (1993), and the 'total-order sensitivity index' $S_{Ti}$, Homma and Saltelli (1996); both rely on decomposing the variance of the output, $V = \mathbb{V}(q)$, into portions attributable to inputs or sets of inputs.

The first-order index measures the contribution to $V = \mathbb{V}(q)$ of varying the $i$-th input alone, and it is defined as $S_i = \mathbb{V}(\mathbb{E}(q|\gamma_i))/V$. This index can be seen as the application of Karl Pearson's correlation ratio $\eta^2$, see Pearson (1905), to the present context. This corresponds to seeing the effect of including or not including a regressor, but averaged over all possible combinations of other regressors. However, this measure does not account for interactions with the inclusion/exclusion of other regressors; hence it is not used in the present paper.

Instead, here the focus is placed on the total effect index, which is defined by Homma and Saltelli (1996) as

$$S_{Ti} = \frac{\mathbb{E}(\mathbb{V}(q|\boldsymbol{\gamma}_{-i}))}{V} = 1 - \frac{\mathbb{V}(\mathbb{E}(q|\boldsymbol{\gamma}_{-i}))}{V}. \tag{4}$$

In the following, the numerator of $S_{Ti}$ is indicated as $\sigma^2_{Ti} = \mathbb{E}(\mathbb{V}(q|\boldsymbol{\gamma}_{-i}))$, and the shorthand $S_T$ for $S_{Ti}$ is often used.

Examining $\sigma^2_{Ti}$, one can notice that the inner term, $\mathbb{V}(q|\boldsymbol{\gamma}_{-i})$, is the variance of $q$ due inclusion/exclusion of regressor $i$, but conditional on a given combination $\boldsymbol{\gamma}_{-i}$ of the remaining regressors. The outer expectation then averages over all values of $\boldsymbol{\gamma}_{-i}$; this quantity is then standardized by $V$ to give the fraction of total output variance caused by the inclusion of $x_i$. The second expression shows that $S_{Ti}$ is 1 minus the first order effect for $\boldsymbol{\gamma}_{-i}$.

These measures are based on the standard variance decomposition formula, or 'law of total variance' (Billingsley 1995), Problem 34.10(b)). In the context of GSA,

these decomposition formulae are discussed in Archer, Saltelli, and Sobol (1997), Saltelli and Tarantola (2002), Sobol' (1993), Brell, Li, and Rabitz (2010). For further reading about GSA in their original setting, see Saltelli et al. (2012).

## 2.4 Estimation of the Total Sensitivity Index

In order to calculate the total sensitivity measure $S_{Ti}$ one should be able to compute $q(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma} \in \Gamma$ (i.e. estimate all possible submodels of the GUM) which is unfeasible or undesirable. Instead, $S_{Ti}$ can be estimated from a random subset of $\Gamma$, i.e. a sample of models. The estimation of $S_{Ti}$ is performed using an estimator and a structured sample constructed as in Jansen (1999), which is a widely used method in GSA.

Specifically, generate a random draw of $\boldsymbol{\gamma}$ in $\Gamma$, say $\boldsymbol{\gamma}_\star$; then consider elements $\boldsymbol{\gamma}_\star^{(i)}$ with all elements equal to $\boldsymbol{\gamma}_\star$ except for the $i$-th coordinate which is switched from 0 to 1 or vice-versa, $\gamma_{\star i}^{(i)} = 1 - \gamma_{\star i}$. Doing this for each coordinate $i$ generates $p$ pairs of $\boldsymbol{\gamma}$ vectors, $\boldsymbol{\gamma}_\star$ and $\boldsymbol{\gamma}_\star^{(i)}$, that differ only in the coordinate $i$. This is then used to calculate $q(\boldsymbol{\gamma})$ and apply an estimator of Jansen (1999).

This process can be formalized as follows: initialize $\ell$ at 1, then,

1. Generate a random draw of $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is a $p$-length vector with each element is randomly selected from {0, 1}. Denote this by $\boldsymbol{\gamma}_\ell$.
2. Evaluate $q_\ell = q(\boldsymbol{\gamma}_\ell)$.
3. Take the $i$th element of $\boldsymbol{\gamma}_\ell$, and switch it to 0 if it is equal to 1, and to 1 if it is 0. Denote this new vector with inverted $i$th element as $\boldsymbol{\gamma}_\ell^{(i)}$.
4. Evaluate $q_{i\ell} = q(\boldsymbol{\gamma}_\ell^{(i)})$.
5. Repeat steps 3 and 4 for $i = 1, 2, \ldots, p$.
6. Repeat steps 1–5 $N$ times, i.e. for $\ell = 1, 2, \ldots, N$.

The estimators for $\sigma_{Ti}^2$ and $V$ are then defined as in Jansen (1999), see also Saltelli et al. (2010):

$$\widehat{\sigma}_{Ti}^2 = \frac{1}{4N} \sum_{\ell=1}^{N} (q_{i\ell} - q_\ell)^2, \quad \widehat{V} = \frac{1}{N-1} \sum_{\ell=1}^{N} (q_\ell - \overline{q})^2, \tag{5}$$

where $\overline{q} = \frac{1}{N} \sum_{\ell=1}^{N} q_\ell$. This delivers the following plug-in estimator for $S_T$, $\widehat{S}_{Ti} = \widehat{\sigma}_{Ti}^2 / \widehat{V}$. Readers familiar with sensitivity analysis may notice that the estimator in (5) is

different by a factor of 2 to the estimator quoted in Saltelli et al. (2010). The reason for this is given in Appendix A.[12]

$\widehat{S}_{Ti}$ is an accurate estimator for $S_{Ti}$ as the number $N$ of models increases;[13] hence, the following discussion is based on the behavior of $S_{Ti}$.

# 3 Properties of Orderings Based on $S_{Ti}$

This section investigates the theoretical properties of ordering of variables in a regression model based on $S_T$, and shows that these orderings satisfy the following minimal requirement. When the true regressors in $\mathbb{T}$ included in the DGP and the irrelevant ones in $\mathbb{M}$ are uncorrelated, the ordering of regressors based on $S_T$ separates the true from the irrelevant regressors in large samples.

Recall that $S_{Ti} = \sigma^2_{Ti}/V = \mathbb{E}(\mathbb{V}(q|\boldsymbol{y}_{-i}))/V$, see (4). The large $n$ properties of $S_{Ti}$ are studied under the following regularity assumptions.

**Assumption 1:** (Assumptions on the DGP). The variables $w_t \coloneqq (y_t, x_{1,t}, \ldots, x_{p,t}, \epsilon_t)'$ are stationary with finite second moments, and satisfy a law of large numbers for large $n$, i.e. the second sample moments of $w_t$ converge in probability to $\Sigma$, the variance covariance matrix of $w_t$.

Notice that these requirements are minimal, and they are satisfied by the HP DGPs as well as the DH DGPs. The following theorem shows that for large $n$, a scree plot on the ordered $S_{Ti}$ allows to separate the relevant regressors from the irrelevant ones when true and irrelevant regressors are uncorrelated.

**Theorem 2:** (Ordering based on $S_{Ti}$ works for uncorrelated regressors in $\mathbb{M}$ and $\mathbb{T}$). *Let Assumption* 1 *hold and assume that the covariance* $\Sigma_{\ell j}$ *between* $x_\ell$ *and* $x_j$ *equals* 0 *for all* $j \in \mathbb{T}$ *and* $\ell \in \mathbb{M}$. *Define* $(S_{T(1)}, S_{T(2)}, \ldots, S_{T(p)})$ *as the set of* $S_{Ti}$ *values in decreasing order, with* $S_{T(1)} \geq S_{T(2)} \geq \cdots \geq S_{T(p)}$. *Then as* $n \to \infty$ *one has*

12 A heuristic reason for this is that the method involves an exploration of models, with equal probability to select $\gamma_i = 0$ or $\gamma_i = 1$. Note that in analyses with continuous variables, it is usually advisable to use low-discrepancy sequences due to their space-filling properties, see Sobol' (1967), which give faster convergence with increasing $N$. However, since $\gamma$ can only take binary values for each element, low-discrepancy sequences offer no obvious advantage over (pseudo-)random numbers.

13 For instance, it is consistent for $S_{Ti}$ for increasing $N$ thanks for Law of Large numbers for i.i.d. sequences applied to its numerator and denominator.

$$(S_{T(1)}, S_{T(2)}, \ldots, S_{T(p)}) \xrightarrow{p} (c_{(1)}, c_{(2)}, \ldots, c_{(r_0)}, 0, \ldots 0)$$

*where* $(c_{(1)}, c_{(2)}, \ldots, c_{(r_0)})$ *is the ordered set of* $c_i > 0$ *values in decreasing order, where*

$$c_i := \frac{1}{4 \cdot 2^{p-1}} \sum_{\boldsymbol{\gamma}_{-i} \in \Gamma_{-i}} \log\left( \frac{\sigma^2 + \sum_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}(i,1)}} \beta_{0,h} \Sigma_{hj.b_{\boldsymbol{\gamma}(i,1)}} \beta_{0,j}}{\sigma^2 + \sum_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}(i,0)}} \beta_{0,h} \Sigma_{hj.b_{\boldsymbol{\gamma}(i,0)}} \beta_{0,j}} \right), \tag{6}$$

*see* Appendix A *for the definition of the relevant quantities in* eq. (6). *Hence the ordered* $S_{Ti}$ *values separate the block of true regressors in* $\mathbb{T}$ *in the first* $r_0$ *positions from the irrelevant ones* $\mathbb{M}$ *in the last* $p - r_0$ *positions of* $(S_{T(1)}, S_{T(2)}, \ldots, S_{T(p)})$.

*Proof*. See Lemma 6 in Appendix A. □

Given the above, one may hence expect this result to apply to other more general situations. However, this turns out not to be necessarily the case. The results in Appendix A also show that one can build examples with correlated regressors across $\mathbb{T}$ and $\mathbb{M}$, where the ordering of regressors based on $S_T$ fails to separate the sets of true and irrelevant regressors in large samples.[14]

In the end, the question of whether the ordering based on $S_T$ can help in selecting regressors is an empirical matter. Section 5 explores the frequency with which this happens in practice, based on simulated data from various DGPs.

# 4 Construction of the Algorithm

In order to construct an algorithm to perform model selection based on $S_T$, an initial investigation was performed to understand to what extent the ranking of regressors provided by $S_T$ is complementary to that given by *t*-ratios. These experiments are based on the MC design by HP; details of these experiments are reported in Section 5. However, since the results provide the basis of the GSA algorithm, they are also summarized here.

In short, 11 different datasets were simulated following the approach and underlying DGPs defined by HP. For each DGP, the regressors were ordered using both $S_T$ and the *t*-ratios. Then a metric was used which measures the success of each ranking in assigning the regressors in the DGP with the highest ranks. This gives a measure of the utility of each ranking in correctly identifying the DGP. It was found that first, $S_T$ gave overall better rankings than *t*-ratios, but for some DGPs *t*-ratios were still more effective.

**14** Worked out examples of this are available from the authors upon request.

This result pointed to the fact that the two measures are in some way complementary, and motivated the GSA algorithm proposed here, which combines the search paths obtained using the $t$-ratios and the $S_T$ measures, and then selects the best model between the two resulting specifications. The combined procedure is expected to be able to reap the advantages of both orderings. For simplicity, this algorithm is called the GSA algorithm, despite the fact that it exploits both the orderings based on GSA and on the $t$-ratios. The rest of this section contains a description of the GSA algorithm in its basic form and with two modifications.

## 4.1 The Basic Algorithm

The procedure involves ranking the regressors by $t$-ratios or $S_T$, then adopting the 'bottom up' approach following the 'Pantula principle', where candidate models are built by successively adding regressors in order of importance. The steps are as follows.

1. Order all regressors by method $m$ (i.e. either the $t$-ratios or $S_T$).
2. Define the initial candidate model as the empty set of regressors (i.e. one with only the constant term).
3. Add to the candidate model the highest-ranking regressor (that is not already in the candidate model).
4. Perform an $F$ test, comparing the validity of the candidate model to that of the GUM.
5. If the $p$-value of the $F$ test in step 4 is below a given significance level $\alpha$, go to step 3 (continue adding regressors), otherwise, go to step 6.
6. Since the $F$-test has not rejected the model in step 4, this is the selected model $\boldsymbol{y}^{(m)}$.

In the following, the notation $\boldsymbol{y}^{(t)}$ is used (respectively $\boldsymbol{y}^{(S)}$) to denote the model selected by this algorithm when $t$-ratios (respectively $S_T$) are used for the ordering. Note that candidate variables are added starting from an empty specification; this is hence a 'bottom up' approach induced by the 'Pantula principle'.

One can observe that this 'bottom up' approach is in line with the GETS philosophy of model selection; in fact it corresponds to the nesting of models known as the 'Pantula-principle' in cointegration rank determination, see Johansen (1996). Every model in the sequence is compared with the GUM, and hence the sequence of tests can be interpreted as an implementation of the GETS philosophy. Moreover, it can be proved that, for large sample sizes, the sequence selects the smallest true model in the sequence with probability equal to $1 - \alpha$,

where $\alpha$ is the size of each test. Letting $\alpha$ tend to 0 as the sample size gets large, one can prove that this delivers a true model with probability tending to 1.[15]

As a last step, the final choice of regressors $\hat{\boldsymbol{y}}$ is chosen between $\boldsymbol{y}^{(t)}$ and $\boldsymbol{y}^{(S)}$ as the one with the fewest regressors (since both models have been declared valid by the $F$-test). If the number of regressors is the same, but the regressors are different, the choice is made using the BIC.

The GSA algorithm depends on some key constants; the significance level of the $F$-test, $\alpha$, is a truly 'sensitive' parameter, in that varying it strongly affects its performance. Of the remaining constants in the algorithm, $N$, the number of points in the GSA sampling, can be increased to improve accuracy; in practice it was found that $N = 128$ provided good results, and further increases made little difference.

In the following two subsections, two extensions to the basic algorithm are outlined with the reasoning explained.

## 4.2 Adaptive-$\alpha$

Varying $\alpha$ essentially dictates how 'strong' the effect of regressors should be to be included in the final model, such that a high $\alpha$ value will tend to include more variables, whereas a low value will cut out variables more harshly. The difficulty is that some DGPs require low $\alpha$ for accurate identification of the true regressors in $\mathbb{T}$, whereas others require higher values. Hence, there could exist no single value of $\alpha$ that is suitable for the identification of all DGPs.

A proposed modification to deal with this problem is to use an 'adaptive-$\alpha$', $\alpha_\phi$, which is allowed to vary depending on the data. This is based on the observation that the $F$-test returns a high $p$-value $p_{\mathrm{H}}$ (typically of the order 0.2–0.6) when the proposed model is a superset of the DGP, but when one or more of the regressors in $\mathbb{T}$ are missing from the proposed model, the $p$-value will generally be low, $p_{\mathrm{L}}$ (of the order $10^{-3}$ say). The values of $p_{\mathrm{H}}$ and $p_{\mathrm{L}}$ will vary depending on the DGP and data set, making it difficult to find a single value of $\alpha$ which will yield good results across all DGPs. However, for a given DGP and data set, the $p_{\mathrm{H}}$ and $p_{\mathrm{L}}$ values are easy to identify.

Therefore, it is proposed to use a value of $\alpha_\phi$, such that for each data set,

$$\alpha_\phi = p_{\mathrm{L}} + \phi\left(p_{\mathrm{H}} - p_{\mathrm{L}}\right) \tag{7}$$

where $p_{\mathrm{H}}$ is taken as the $p$-value resulting from considering a candidate model with

15 See for instance Paruolo (2001). Recall that any model whose set of regressors contains the DGP is 'true'.

all regressors that have $S_{Ti} > 0.01$ against the GUM, and $p_L$ is taken as the $p$-value from considering the empty set of regressors against the GUM. The reasoning behind the definition of $p_H$ is that it represents a candidate model which will contain the DGP regressors with a high degree of confidence. Here $\phi$ is a tuning parameter that essentially determines how far between $p_L$ and $p_H$ the cutoff should be. Figure 1 illustrates this on a data set sampled from DGP 6B. Note that $\alpha_\phi$ is used in the $F$-test for both the $t$-ranked regressors as well as those ordered by $S_T$.

## 4.3 Skipping Regressors

In order to correct situations where the ordering of the regressors is not correct, a different extension of the algorithm is to test discarding "weak" regressors in the selected model. Here, a weak regressor is defined as being one with a value of $S_T$ lower than a certain threshold, which is set as 0.2. When Step 6 is reached, if weak regressors exist in the selected model, they are removed one at a time, each time performing an $F$-test. If the $F$-test is satisfied, the regressor is discarded, otherwise it is retained. This approach is used instead of an exhaustive search of the combinations of remaining regressors, because occasionally there may still be too many regressors left to make this feasible.

## 4.4 Full GSA Algorithm

Adding the extensions discussed in the previous two sections results in the final full algorithm, which can be described as follows.

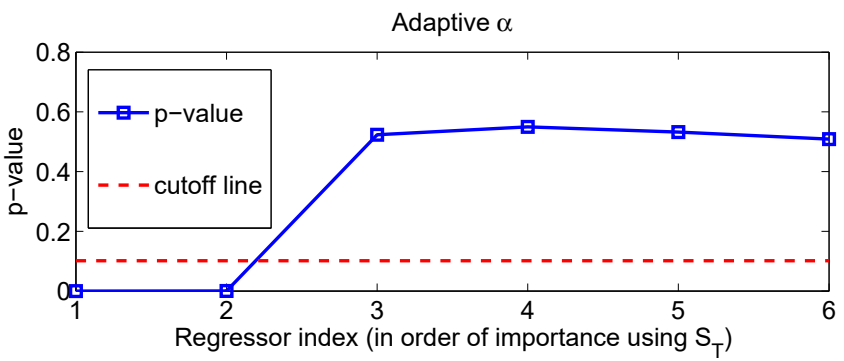


**Figure 1:** $p$-Values from $F$-test comparing candidate models to the GUM in a sample from DGP 6B, for the six highest-ranked regressors. Here $\phi = 0.2$ and $\alpha_\phi$ is marked as a dotted line.

1. Obtain two orderings of regressors, one by the $t$-ratios, and the other by $S_T$, using the BIC as the output (penalized measure of model fit) $q$.
2. Obtain $p_H$ as the $p$-value resulting from considering a candidate model with all regressors that have $S_{Ti} > 0.01$ against the GUM, and $p_L$ as the $p$-value from considering the empty set of regressors against the GUM. Calculate $\alpha_\phi$ using (7), which is used in all subsequent tests.
3. Define the initial candidate model as the empty set of regressors (i.e. one with only the constant term).
4. Add to the candidate model the regressor with the highest $S_T$ (that is not already in the candidate model).
5. Perform an $F$ test, comparing the validity of the candidate model to that of the GUM.
6. If the $p$-value of the $F$ test in step 4 is below $\alpha_\phi$, go to step 4 (continue adding regressors), otherwise, go to step 7.
7. Since the $F$-test has not rejected the model in step 4, this is the selected model $\boldsymbol{\gamma}^{(m)}$.
8. Identify any remaining 'weak' regressors as those with $S_T < 0.2$. Try removing these one at a time: if removing a regressor satisfies the $F$-test, it is discarded; otherwise it is retained. Repeat this procedure for all weak regressors.
9. Repeat steps 3–8, except use the ordering based on $t$-ratios, rather than on $S_T$.
10. Compare between the final model selected by $S_T$ and the final model selected by the $t$-ratios, by selecting the model with the fewest regressors (since both have satisfied the $F$-test). If both final specifications have the same number of regressors, chose the specification with the lowest BIC.

In the following section the performance of the algorithm is examined compared to some benchmark test cases, with and without the extensions introduced in previous sections. In the following, $S_{T\text{full}}$ indicates the full procedure as described above; $S_{T\text{no-skip}}$ refers to the same procedure without the skipping extension (i.e. without step 8); finally $S_{T\text{simple}}$ is the one without step 8, and also without step 2 (adaptive-$\alpha$). For $S_{T\text{simple}}$ a fixed value of $\alpha$ is used.

# 5 The Experiments of Hoover and Perez

This section tests the GSA algorithm on a suite of DGP simulation experiments developed by HP. These experiments consider a possibly dynamic regression equation with $n = 139$ and exogenous variables, fixed across experiments, taken from real-world, stationary, macroeconomic time series, in the attempt to represent typical macroeconomic data. Several papers have used HP's experiments to test the

performance of other methods (Castle, Doornik, and Hendry 2011; Hendry and Krolzig 1999). HP's experiments are of varying degree of difficulty for model search algorithms. Details on the design of HP DGPs are reported in Appendix B.

The features of HP's experiments prompt a number of considerations. First, because sample size is limited and fixed, consistency of model-selection algorithms cannot be the sole performance criterion. Secondly, some of the DGPs in HP's experiments are characterized by a low signal-to-noise ratio for some coefficients; the corresponding regressors are labeled 'weak'. This situation makes it very difficult for statistical procedures to discover if the corresponding regressors should be included or not. This raises the question of how to measure selection performance in this context.

This paper observes that, in the case of weak regressors, one can measure performance of model-selection algorithms also with respect to a simplified DGP, which contains the subset of regressors with sufficiently high signal-to-noise ratio; this is called the 'Effective DGP' (EDGP). The definition of the EDGP is made operational using the 'parametricness index' introduced in Liu and Yang (2011)—this concept is described in detail in Appendix C. For full transparency, results are presented also relative to the original DGPs in cases where the EDGP is different.

## 5.1 Orderings Based on *t* and GSA

As mentioned in Section 4, the DGPs of HP were used as the basis for an initial investigation into the comparative rankings of $S_T$ and the $t$ ratios. Here, these numerical experiments are described in more detail.

For each of the 11 DGPs under investigation, $N_R = 500$ replications for $\boldsymbol{Z}$ were generated; on each sample, regressors were ranked by the $t$-ratios and $S_T$, using $N = 128$ in (5). Both for the $t$-ratios ranking and the $S_T$ ranking, the ordering is from the best-fitting regressor to the worst-fitting one.

In order to measure how successful the two methods were in ranking regressors, the following measure $\delta$ of minimum relative covering size is defined. Indicate by $\varphi_0 = \{i_1, \ldots, i_{r_0}\}$ the set containing the positions $i_j$ of the true regressors in the list $i = 1, \ldots, p$; i.e. for each $j$ one has $\gamma_{0,i_j} = 1$. Recall also that $r_0$ is the number of elements in $\varphi_0$. Next, for a generic replication $j$, let $\varphi_\ell^{(m)} = \{i_1^{(m)}, \ldots, i_\ell^{(m)}\}$ be the set containing the first $\ell$ positions $i_j^{(m)}$ induced by the ordering of method $m$, $m$ equal $t$, $S_T$. Let $b_j^{(m)} = \min\{\ell : \varphi_0 \subseteq \varphi_\ell^{(m)}\}$ be the minimum number of elements $\ell$ for which $\varphi_\ell^{(m)}$ contains all the true regressors. Observe that $b_j^{(m)}$ is well defined, because at least for $\ell = p$ one always has $\varphi_0 \subseteq \varphi_p^{(m)} = \{1, \ldots, p\}$. $\delta$ is defined to equal $b_j^{(m)}$ divided by its minimum;

this corresponds to the (relative) minimum number of elements in the ordering $m$ that covers the set of true regressors.

Observe that, by construction, one has $r_0 \le b_j^{(m)} \le p$, and that one wishes $b_j^{(m)}$ to be as small as possible; ideally one would like to have to have $b_j^{(m)} = r_0$. Hence for $\delta_j^{(m)}$ defined as $b_j^{(m)}/r_0$ one has $1 \le \delta_j^{(m)} \le p/r_0$. Finally $\delta^{(m)}$ is defined as the average $\delta_j^{(m)}$ over $j = 1, \ldots, N_R$, i.e. $\delta^{(m)} = \frac{1}{N_R}\sum_{j=1}^{N_R}\delta_j^{(m)}$.

For example, if the regressors, ranked in descending order of importance by method $m$ in replication $j$, were $x_3, x_{12}, x_{21}, x_{11}, x_4, x_{31}, \ldots$, and the true DGP were $x_3$, $x_{11}$ the measure $\delta_j$ would be 2; in fact the smallest-ranked set containing $x_3$, $x_{11}$ has four elements $b_j^{(m)} = 4$, and $r_0 = 2$.

The results over the $N_R = 500$ replications are summarized in Table 1. Overall $S_T$ appears to perform better than $t$-ordering. For some DGPs (such as DGP 2 and 5) both approaches perform well ($\delta = 1$ indicating correct ranking for all 500 data sets). There are other DGPs where the performance is significantly different. In particular the $t$-ratios is comparatively deficient on DGPs 3 and 6A, whereas $S_T$ performs worse on DGP 8. This suggests that there are some DGPs in which $S_T$ may offer an advantage over the $t$-ratios in terms of ranking regressors in order of importance. This implies that a hybrid approach, using both measures, may yield a more efficient method of regressor selection.

**Table 1:** Values of $\delta$ (average over 500 data replications per DGP), using $t$-test and $S_T$. Mean refers to average across DGPs. Comparatively poor rankings are in boldface.

| DGP | 1 | 2 | 3 | 4 | 5 | 6 | 6A | 6B | 7 | 8 | 9 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $S_T$ | | 1.00 | 1.01 | 1.00 | 1.00 | 1.00 | 1.12 | 1.02 | **1.15** | **1.64** | **1.13** | 1.11 |
| $t$-ratios | | 1.00 | **1.53** | **1.04** | 1.00 | **1.06** | **3.95** | **1.14** | 1.04 | 1.00 | 1.01 | **1.38** |

## 5.2 Measures of Performance

The performance of algorithms was measured by HP via the number of times the algorithm selected the DGP as a final specification. Here use is made of measures of performance similar to the ones in HP, as well as of additional ones proposed in Castle, Doornik, and Hendry (2011).

Recall that $\boldsymbol{\gamma}_{\mathbb{T}} = \boldsymbol{\gamma}_0$ is the true set of included regressors and let $\widehat{\boldsymbol{\gamma}}_j$ indicate the one produced by a generic algorithm in replication $j = 1, \ldots, N_R$. Define $r_j$ to be number of correct inclusions of components in vector $\widehat{\boldsymbol{\gamma}}_j$, i.e. the number of

regression indices $i$ for which $\widehat{\gamma}_{j,i} = \gamma_{0,i} = 1$, $r_j = \sum_{i=1}^{p} 1(\widehat{\gamma}_{j,i} = \gamma_{0,i} = 1)$. Recall that $r_0$ indicates the number of true regressors.

The following exhaustive and mutually exclusive categories of results can be defined:

- $C_1$: exact matches;
- $C_2$: the selected model is correctly specified, but it is larger than necessary, i.e. it contains all relevant regressors as well as irrelevant ones;
- $C_3$: the selected model is incorrectly specified (misspecified), i.e. it lacks relevant regressors.

$C_1$ matches correspond to the case when $\widehat{\boldsymbol{\gamma}}_j$ coincides with $\boldsymbol{\gamma}_{\mathbb{T}} = \boldsymbol{\gamma}_0$; the corresponding frequency $C_1$ is computed as $C_1 = \frac{1}{N_R}\sum_{j=1}^{N_R} 1(\widehat{\boldsymbol{\gamma}}_j = \boldsymbol{\gamma}_{\mathbb{T}})$. The frequency of $C_2$ cases is given by $C_2 = \frac{1}{N_R}\sum_{j=1}^{N_R} 1(\widehat{\boldsymbol{\gamma}}_j \neq \boldsymbol{\gamma}_0, r_j = r_0)$. Finally, $C_3$ cases are the residual category, and the corresponding frequency is $C_3 = 1 - C_1 - C_2$.[16]

The performance can be further evaluated through measures taken from Castle, Doornik, and Hendry (2011), known as potency and gauge. First the retention rate $\tilde{p}_i$ of the $i$-th variable is defined as, $\tilde{p}_i = \frac{1}{N_R}\sum_{j=1}^{N_R} 1(\widehat{\gamma}_{j,i} = 1)$. Then, potency and gauge are defined as follows:

$$\text{potency} = \frac{1}{r_0}\sum_{i:\beta_{0,i}\neq 0} \tilde{p}_i, \quad \text{gauge} = \frac{1}{p - r_0}\sum_{i:\beta_{0,i}=0} \tilde{p}_i.$$

Potency therefore measures the average frequency of inclusion of regressors belonging to the DGP, while gauge measures the average frequency of inclusion of regressors not belonging to the DGP. An ideal performance is thus represented by a potency value of 1 and a gauge of 0.

In calculating these measures, HP chose to discard MC replications for which a preliminary application of the battery of misspecification tests defined in (15) in Appendix A reported a rejection.[17] This choice is called in the following 'pre-search elimination' of MC replications.

---

**16** $C_1$ corresponds to Category 1 in HP; $C_2$ corresponds to Category 2 + Category 3 − Category 1 in HP; finally $C_3$ corresponds to Category 4 in HP.

**17** The empirical percentage of samples that were discarded in this way was found to be proportional to the significance level $\alpha$. This fact, however, did not influence significantly the number of $C_1$ catches. Hence the HP procedure was allowed to discard replications as in the original version. For the GSA algorithm no pre-search elimination was performed.

## 5.3 Benchmark

The performance of HP's algorithm is taken as a benchmark. The original MATLAB code for generating data from HP's experiments was downloaded from HP's home page.[18] The original scripts were then updated to run on the current version of MATLAB. A replication of the results in Tables 4, 6 and 7 in HP is reported in the first panel of Table 2, using a nominal significance level of $\alpha$ = 1, 5, 10% and $N_R = 10^3$ replications. The results do not appear to be significantly different from the ones reported in HP.

When checking the original code, an incorrect coding was found in the original HP script for the generation of the AR series $u_t$ in Eq. (13), which produced simulations of a moving average process of order 1, MA(1), with MA parameter 0.75 instead of an AR(1) with AR parameter 0.75.[19] The script was hence modified to produce $u_t$ as an AR(1) with AR parameter 0.75; this is called the 'modified script' in the following.

Re-running the DGP simulation experiments using this modified script, the results in the second panel in Table 2 were obtained; for this set of simulations $N_R = 10^4$ replications were used. Comparing the first and second panel in the table for the same nominal significance level $\alpha$, one observes a significant increase in $C_1$ catches in DGP 2 and 7. One reason for this can be that when the modified script is employed, the regression model is well-specified, i.e. it contains the DGP as a special case.[20] Table 2 documents how HP's algorithm depends on $\alpha$, the significance level chosen in the test $R$ in (15).

## 5.4 Alternative Algorithms

This section presents results using the performance measures introduced in Section 5.2. The results compare the three variations of the $S_T$ algorithm with the

**18** http://www.csus.edu/indiv/p/perezs/Data/data.htm.

**19** This means that the results reported in HP for DGP 2, 3, 7, 8, 9 refer to a misspecified model. The MA process can be inverted to obtain a AR($\infty$) representation; substituting from the $y_t$ equation as before, one finds that the DGP contains an infinite number of lags on the dependent variable and of the $x^*_{it}$ variables, with exponentially decreasing coefficients. The entertained regression model with four lags on the dependent variable and two lags on the $x^*_{it}$ variables can be considered an approximation to the DGP.

**20** This finding is similar to the one reported in Hendry and Krolzig (1999), section 6; they re-run HP experiments using PcGets, and they document similar increases in $C_1$ catches in DGP 2 and 7 for their modified algorithms. Hence, it is possible that this result is driven by the correction of the script for the generation of the AR series.

**Table 2:** Percentages of Category 1 matches $C_1$ for different values of $\alpha$. Original script: data generated by the original script, $N_R = 10^3$ replications. The frequencies are not statistically different from the ones reported in HP (Tables 4, 6, 7). Modified script: data from modified script for the generation of AR series, $N_R = 10^4$ replications.

| | Original script | | | Modified script | | |
|---|---|---|---|---|---|---|
| **DGP** | **$\alpha$ = 0.01** | **0.05** | **0.1** | **0.01** | **0.05** | **0.1** |
| 1 | 81.1 | 28.6 | 6.8 | 79.3 | 30.0 | 7.3 |
| 2 | 1.2 | 0.0 | 0.0 | 77.0 | 27.0 | 6.9 |
| 3 | 71.4 | 27.2 | 9.1 | 71.7 | 27.3 | 6.9 |
| 4 | 78.2 | 31.2 | 6.4 | 81.8 | 31.1 | 7.0 |
| 5 | 80.9 | 30.1 | 7.4 | 80.7 | 29.9 | 6.4 |
| 6 | 0.2 | 1.0 | 0.7 | 0.4 | 0.5 | 0.6 |
| 6A | 68.0 | 27.8 | 7.8 | 70.6 | 27.8 | 7.6 |
| 6B | 80.8 | 30.7 | 7.8 | 81.1 | 31.4 | 8.0 |
| 7 | 23.6 | 4.7 | 0.3 | 75.7 | 26.7 | 7.6 |
| 8 | 80.6 | 31.0 | 8.0 | 79.2 | 30.3 | 9.4 |
| 9 | 0.1 | 0 | 0 | 0 | 0 | 0 |

modified HP code. To compare with a similar but more recent GETS implementation, the Autometrics package 'gets', see Pretis, Reade, and Sucarrat (2018), is also added as an additional algorithm in the comparison.

The performance is measured with respect to the true DGP or with respect to the Effective DGP (EDGP) that one can hope to recover, given the signal to noise ratio. Because the HP, GSA and Autometrics algorithms depend on tunable constants, results are given for various values of these constants.

The procedure employed to define the EGDP is discussed in Appendix C; it implies that the only EDGPs differing from the true DGP are DGP 6 and DGP 9. DGP 6 contains regressors 3 and 11, but regressor 3 is weak and hence EDGP 6 contains only regressor 11. DGP 9 contains regressors 3, 11, 21, 29 and 37 but regressors 3 and 21 are weak and they are dropped from the corresponding EDGP 9. More details are given in Appendix C.

Both the HP algorithm and Autometrics depend on the significance levels $\alpha$, whereas the GSA algorithm depends on the threshold $\phi$ (which controls $\alpha_\phi$) for $S_{T\text{no-skip}}$ and $S_{T\text{full}}$ and on $\alpha$ for $S_{T\text{simple}}$. Because the values of $\alpha$ and $\phi$ can seriously affect the performance of the algorithms, a fair comparison of the performance of the algorithms may be difficult, especially since the true parameter values will not be known in practice. To deal with this problem, the performance of the algorithms was measured at a number of parameter values within a plausible range.

This allowed two ways of comparing the algorithms: first, the 'optimized' performance, corresponding to the value of $\alpha$ or $\phi$ that produced the highest $C_1$ score, averaged over the 11 DGPs. This can be viewed as the 'potential performance'. In practice, the optimization was performed with a grid search on $\alpha$ and $\phi$ with $N_R = 10^3$ replications, averaging across DGPs.

Secondly, a qualitative comparison was performed between the algorithms of comparing their average performance over the range of parameter values. This latter comparison gives some insight into the more realistic situation, where the optimum parameter values are not known.

## 5.5 Results for optimal values of tuning coefficients

Table 3 shows the classification results in terms of $C_1$ matches, as well as the potency and gauge measures, for all algorithms at their optimal parameter values, using $N_R = 10^4$. Note that the value of $\alpha = 4 \times 10^{-4}$ for Autometrics represents the lowest value of $\alpha$ that it was possible to assign without errors occurring due to singular matrices—likely due to issues with numerical precision. Results for the $S_T$ algorithm are shown with and without the extensions discussed in Section 4. Recovery of the true specification is here understood in the EDGP sense.

The $C_1$ column measures the percentage frequency with which the algorithms identified the EDGP. One notable fact is that the performance of the HP algorithm has been vastly improved (compared to the results in HP's original paper) simply by setting $\alpha$ to a better value, in this case $\alpha = 4 \times 10^{-4}$, compare with Table 2.

The comparison shows that with the full $S_T$ algorithm, the correct classification rate ($C_1$) is 98.9%, compared with 94.3% for HP, and 88.6% with Autometrics. It is presumed that if it were possible to reduce further the value of $\alpha$ for Autometrics, the mean $C_1$ value would increase still further. However, it was not possible to test this conjecture. Removing the 'skipping' extension, the average performance falls to 96.7%, and further to 92.6% without the adaptive-$\alpha$ feature.

Examining the DGPs individually, the GSA algorithm performs well on all DGPs, although there are slightly lower $C_1$ values in DGPs 3 and 6A (around 96%). For HP, these differences are more marked, with $C_1 = 62\%$ for DGP 3, and $C_1 = 85.3\%$ for DGP 7. Autometrics also has a lower success rate of 74% for DGP 3, and 87% for DGP 7. It is evident though that the adaptive-$\alpha$ and the skipping extensions contribute significantly to the performance of the GSA algorithm in these DGPs.

The potency and gauge measures (also in Table 3) reveal a little more about the nature of the errors made by the algorithms. Gauge is very low for the GSA and HP algorithms, but higher for Autometrics. Higher gauge measures are found in DGP 6A, particularly for HP, indicating the inclusion of irrelevant regressors. The full

**Table 3:** Percentage $C_1$, percentage gauge (Gge) and percentage potency (Pot) by EDGP. Optimized parameter values used.

| | $S_{T\text{simple}}$ | | | $S_{T\text{no-skip}}$ | | | $S_{T\text{full}}$ | | | $HP_{\text{opt}}$ | | | Autometrics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha = 0.0371$ | | | $\phi = 0.3$ | | | $\phi = 0.3$ | | | $\alpha = 4 \times 10^{-4}$ | | | $\alpha = 4 \times 10^{-4}$ | | |
| EDGP | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot |
| 1 | 98.7 | 0.11 | 100 | 99.8 | 0.01 | 100 | 99.8 | 0 | 100 | 99.2 | 0.02 | 100 | 85.0 | 0.48 | 100 |
| 2 | 98.5 | 0.09 | 100 | 99.4 | 0.02 | 100 | 99.4 | 0.02 | 100 | 98.9 | 0.03 | 100 | 93.0 | 0.18 | 96.0 |
| 3 | 79.4 | 0.8 | 94.7 | 95.2 | 0.10 | 98.5 | 96.0 | 0.06 | 98.5 | 62.0 | 0.05 | 81.2 | 74.0 | 0.42 | 90.0 |
| 4 | 98.6 | 0.09 | 100 | 99.2 | 0.03 | 100 | 99.2 | 0.02 | 100 | 99.3 | 0.02 | 99.9 | 89.0 | 0.28 | 98.0 |
| 5 | 98.8 | 0.08 | 100 | 99.9 | 0 | 100 | 99.9 | 0 | 100 | 99.3 | 0.02 | 100 | 90.0 | 0.28 | 98.0 |
| 6 | 98.7 | 0.09 | 100 | 99.2 | 0.03 | 100 | 99.2 | 0.02 | 100 | 99.2 | 0.03 | 99.8 | 89.0 | 0.31 | 98.0 |
| 6A | 65.3 | 0.46 | 87.9 | 78.4 | 0.66 | 97.9 | 96.2 | 0.05 | 98.5 | 85.3 | 0.55 | 92.9 | 87.0 | 0.47 | 94.0 |
| 6B | 97.6 | 0.1 | 100 | 98.6 | 0.04 | 100 | 99.4 | 0.02 | 100 | 98.4 | 0.07 | 99.5 | 91.0 | 0.24 | 96.0 |
| 7 | 92.7 | 0.13 | 98.6 | 97.1 | 0.09 | 99.9 | 99.5 | 0.01 | 99.9 | 98.8 | 0.03 | 99.8 | 93.0 | 0.24 | 96.7 |
| 8 | 98.4 | 0.07 | 100 | 99.9 | 0 | 100 | 99.9 | 0 | 100 | 99.1 | 0.03 | 100 | 94.0 | 0.05 | 96.0 |
| 9 | 91.4 | 0.18 | 98.6 | 96.5 | 0.11 | 99.9 | 99.6 | 0.01 | 99.9 | 98.2 | 0.04 | 99.8 | 90.0 | 0.24 | 96.0 |
| Mean | 92.6 | 0.2 | 98.2 | 96.7 | 0.10 | 99.7 | 98.9 | 0.02 | 99.7 | 94.3 | 0.08 | 97.5 | 88.6 | 0.29 | 96.2 |

GSA algorithm has gauges of at most 0.06% across these data sets. Autometrics has gauge values which are relatively consistent around 0.2–0.5% in most cases, but as low as 0.05% for DGP 8. This partially confirms the authors' assertion that gauge is close to constant across a variety of models, see Doornik (2009).

The potency measures show that the true regressors are being identified nearly all the time for all three approaches. However, overall the GSA method has the highest potency values of 98% or above, whereas HP has a lower value of 80% for DGP 3. Autometrics yields potency values generally over 95%, with the exception of DGP 3, which has a potency value of 90%.

## 5.6 Recovering the DGP

Although it is argued here that the signal-to-noise ratio in DGPs 6 and 9 is too low for certain regressors to be identified, it is still worth looking at the results with respect to the true DGP, shown in Table 4. All algorithms failed to identify the true DGP 9 even once out of the $10^4$ runs. This fact is reflected in the potency, which drops from 100 to 50% (DGP 6), and about 60% (DGP 9). These results are mirrored in the original results of HP. This suggest that GSA may not help when regressors are very 'weak', but the same is true for HP and Autometrics. Put simply, the signal-to-noise ratio is too low to identify the effect of these regressors (see the discussion of the EDGP in Appendix C).

## 5.7 Robustness of Algorithms

As discussed earlier, the results in Table 3 are obtained after optimization of the tuning parameters $\alpha$ and $\phi$. This provides a measure of potential performance, but in reality the best $\alpha$ and $\phi$ will not be known. For this reason it is indicative to show the results when varying the tuning parameter.

The upper panel in Figure 2 shows how the categorization of the final model varies with $\phi$ in the full GSA algorithm. While the value of $\phi$ varies between 0.1 and 0.5, the value of $C_1$ (exact matches) is generally above 95%, and $C_2$ (correct specification, but with irrelevant regressors) and $C_3$ (misspecification) are consistently very low.

In contrast, the middle panel shows the effect of varying $\alpha$ for the HP algorithm. It is clear that the peak performance of the algorithm is obtained in a small neighborhood around a rather sharp maximum at a low $\alpha$ value—increasing $\alpha$ from this value results in a rapid increase in $C_2$, whereas decreasing it sharply increases $C_3$.

Autometrics exhibits a similar sensitivity to $\alpha$, although the decrease in $C_1$ is even steeper, when $\alpha$ is increased. As noted previously, Autometrics may yield

**Table 4:** Percentage $C_1$, gauge and potency by DGP. Optimized parameter values used. The mean frequency is taken over all DGPs, but only the results for DGPs 6 and 9 are shown since the remaining results are identical to Table 3.

| | STsimple | | | STno-skip | | | STfull | | | HPopt | | | Autometrics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha = 0.0371$ | | | $\phi = 0.3$ | | | $\phi = 0.3$ | | | $\alpha = 4 \times 10^{-4}$ | | | $\alpha = 4 \times 10^{-4}$ | | |
| DGP | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot | $C_1$ | Gge | Pot |
| 6 | 1 | 0.09 | 50.0 | 1 | 0.03 | 50.0 | 2 | 0.02 | 50.0 | 3 | 0.03 | 49.9 | 0 | 0.29 | 49.5 |
| 9 | 0 | 0.19 | 59.2 | 0 | 0.11 | 60.0 | 0 | 0.01 | 60.0 | 0 | 0.04 | 59.9 | 0 | 0.20 | 58.0 |
| Mean | 0.5 | 0.14 | 54.6 | 0.5 | 0.07 | 55.0 | 1 | 0.02 | 55.0 | 1.5 | 0.03 | 54.9 | 0 | 0.25 | 53.8 |

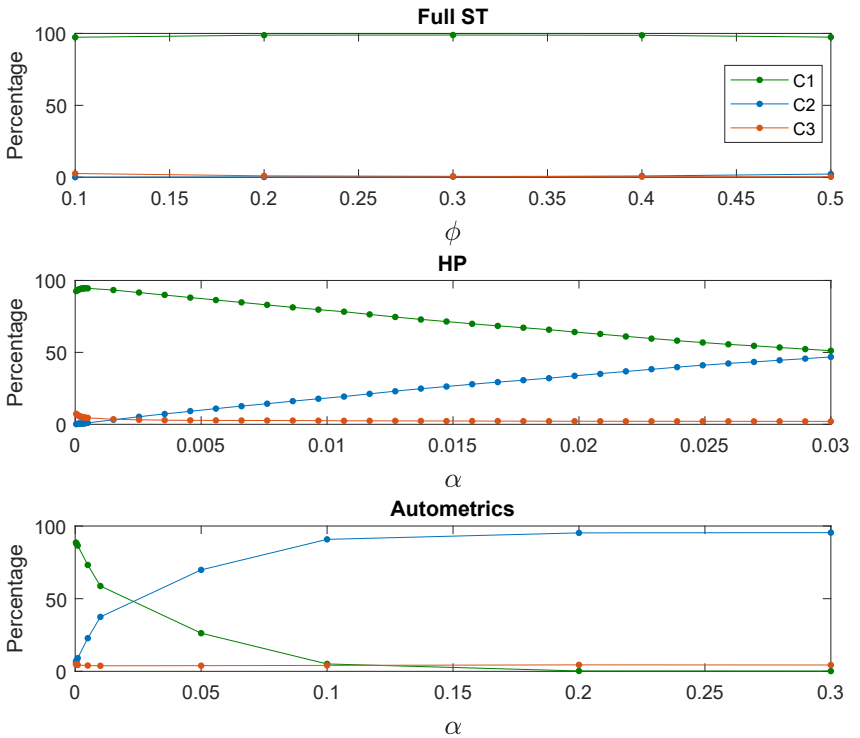


**Figure 2:** Optimization of algorithms with respect to tuning parameters; upper panel: full ST algorithm; middle panel: HP algorithm, lower panel: Autometrics. Percentages correspond to averages over all EDGPs.

even higher $C_1$ values than those observed here, if it were possible to further reduce $\alpha$ without numerical issues.

Overall, while it is difficult to make a perfectly fair comparison of the robustness of the three algorithms, due to the incomparable scales of the optimizing parameters, the GSA algorithm seems to be considerably less sensitive to variation in its tuning coefficient on these data sets, since it indirectly specifies $\alpha$ through $\phi$. This has the advantage that the tuning parameter, $\phi$, is somewhat problem-independent.

# 6 Comparisons with Other Methods

In the previous section, a rather detailed comparison was given with the approach of HP and Autometrics, since those approaches share some similarities with the

proposed GSA algorithm. This section compares with a wider set of approaches and also includes a real case study; specifically, the experiments in Deckers and Hanck (2014) (henceforth DH) are considered. They consist of a comparison of a nine-model selection approach (Section 6.1) in the DH DGPs, as well as an application to a growth regression model (Section 6.2).

## 6.1 Simulation Experiments

To give a comparison of performance with a wider range of model selection procedures, the DGP simulation experiment of Deckers and Hanck (2014) is used as a test case. In their paper, the authors use a simple cross-section regression framework to compare the performance of nine-model selection approaches, additionally investigating the effect of altering tuning parameters. Here, the GSA algorithm is applied to their test case, which allows a comparison with a number of competing model selection procedures. These other methods are the "classical" hypothesis testing procedure, which simply uses $p$-values (classical); the same procedure but with the Bonferroni correction Bonferroni (1936) (Bonferroni); the step-up method of Benjamini and Hochberg (1995) (BH); the bootstrap step-down method of Romano, Shaikh, and Wolf (2008) (Boot); the PcGets/Autometrics software package in Krolzig and Hendry (2001); the HP approach as investigated in the previous section (HP); Bayesian model averaging as in Ley and Steel (2009) (BMA); the "two million regressions" approach in Sala-i-Martin (1997) (S-i-M); and the least absolute shrinkage and selection operator (Lasso) of Tibshirani (1996). For full details on these approaches the reader is referred to Deckers and Hanck (2014), and references therein.

The DGPs are defined using (1) with $p = 50$ and $n = 100$ with $X_1, \ldots, X_{50}$ distributed as a multivariate normal distribution with mean zero, variance one and common correlation $\rho$ that can take values in $\{0, 0.3, 0.5\}$.

DGPs are classified according to how many $\beta_i$ are set to non-zero values.

1. Every tenth $\beta_i$ is set to 0.5, with the rest is set to zero (5 regressor indices in $\mathbb{T}$).
2. Every fifth $\beta_i$ is set to 0.5, with the rest is set to zero (10 regressor indices in $\mathbb{T}$).
3. Every second $\beta_i$ is set to 0.5, with the rest is set to zero (25 regressor indices in $\mathbb{T}$).

DH point out that the value of $\beta_i = 0.5$ is used because it results in population $R^2$ values that are realistic for data sets encountered in growth econometrics. The experiments are run here with 2000 replications for each case.

Tables 5, 6 and 7 give the results of the GSA algorithm added to the existing results of the other methods. For consistency with DH, the performance measures from their work are used. The false discovery rate (FDR) is defined as the expected

number of falsely rejected hypotheses divided by the total number of falsely rejected hypotheses—in this sense, it is similar to the gauge, but is normalized by the number of correct rejections rather than the actual number of false hypotheses. CR denotes the average number of correct rejections—this means that it is simply the potency without the normalization by the number of false hypotheses.

The results of the model selection approaches apart from the GSA algorithm are discussed in Deckers and Hanck (2014) in some detail—for this reason the discussion here is limited to the relative performance of the GSA algorithm. Table 5 shows that in the case of five false hypotheses, the GSA algorithm has a CR of 4.27, 3.99 and 3.60 for $\rho = 0$, 0.3, 0.5 respectively. This puts it on a similar level, albeit

**Table 5:** Results of Monte Carlo experiment with five false hypotheses.

| | $\rho = 0$ | | $\rho = 0.3$ | | $\rho = 0.5$ | |
|---|---|---|---|---|---|---|
| | **FDR** | **CR** | **FDR** | **CR** | **FDR** | **CR** |
| Classical: $\alpha = 0.01$ | 0.083 | 3.96 | 0.097 | 3.08 | 0.125 | 2.25 |
| Classical: $\alpha = 0.05$ | 0.273 | 4.61 | 0.302 | 4.12 | 0.336 | 3.48 |
| Classical: $\alpha = 0.10$ | 0.426 | 4.81 | 0.457 | 4.49 | 0.473 | 3.98 |
| Bonferroni: $\alpha = 0.01$ | 0.003 | 1.76 | 0.003 | 0.94 | 0.006 | 0.46 |
| Bonferroni: $\alpha = 0.05$ | 0.01 | 2.57 | 0.014 | 1.65 | 0.02 | 0.96 |
| Bonferroni: $\alpha = 0.10$ | 0.021 | 3.03 | 0.031 | 2.06 | 0.035 | 1.27 |
| BH: $\alpha = 0.01$ | 0.01 | 2.27 | 0.007 | 1.21 | 0.009 | 0.55 |
| BH: $\alpha = 0.05$ | 0.038 | 3.35 | 0.043 | 2.29 | 0.042 | 1.32 |
| BH: $\alpha = 0.10$ | 0.086 | 3.81 | 0.092 | 2.85 | 0.08 | 1.8 |
| Bootstrap: $\alpha = 0.01$ | 0.011 | 2.36 | 0.009 | 1.22 | 0.01 | 0.59 |
| Bootstrap: $\alpha = 0.05$ | 0.052 | 3.45 | 0.05 | 2.39 | 0.048 | 1.35 |
| Bootstrap: $\alpha = 0.10$ | 0.1 | 3.94 | 0.102 | 2.95 | 0.092 | 1.88 |
| PcGets/Autometrics | 0.334 | 4.95 | 0.344 | 4.83 | 0.363 | 4.51 |
| HP | 0.087 | 4.87 | 0.123 | 4.63 | 0.165 | 4.12 |
| Bayesian model averaging | | | | | | |
| $m = k/2$, $g = k^{-2}$, random $\theta$ | 0.005 | 3.97 | 0.02 | 3.97 | 0.047 | 3.34 |
| $m = 7$, $g = k^{-2}$, random $\theta$ | 0.004 | 3.91 | 0.02 | 3.94 | 0.047 | 3.31 |
| $m = k/2$, $g = k^{-2}$, fixed $\theta$ | 0.056 | 4.85 | 0.066 | 4.58 | 0.093 | 4.01 |
| $m = 7$, $g = k^{-2}$, fixed $\theta$ | 0.007 | 4.38 | 0.021 | 4.12 | 0.05 | 3.48 |
| $m = k/2$, $g = n^{-1}$, random $\theta$ | 0.042 | 4.75 | 0.049 | 4.46 | 0.07 | 3.82 |
| $m = 7$, $g = n^{-1}$, random $\theta$ | 0.036 | 4.72 | 0.042 | 4.42 | 0.066 | 3.77 |
| $m = k/2$, $g = n^{-1}$, fixed $\theta$ | 0.272 | 4.97 | 0.266 | 4.83 | 0.279 | 4.49 |
| $m = 7$, $g = n^{-1}$, fixed $\theta$ | 0.04 | 4.81 | 0.047 | 4.5 | 0.072 | 3.88 |
| Sala-i-Martin | 0.488 | 4.98 | 0.67 | 4.99 | 0.689 | 4.97 |
| Lasso | 0.246 | 4.85 | 0.434 | 4.97 | 0.5 | 4.9 |
| GSA algorithm | 0.029 | 4.27 | 0.057 | 3.99 | 0.102 | 3.60 |

**Table 6:** Results of Monte Carlo experiment with 10 false hypotheses.

| | $\rho = 0$ | | $\rho = 0.3$ | | $\rho = 0.5$ | |
|---|---|---|---|---|---|---|
| | **FDR** | **CR** | **FDR** | **CR** | **FDR** | **CR** |
| Classical: $\alpha = 0.01$ | 0.039 | 7.94 | 0.047 | 6.19 | 0.062 | 4.57 |
| Classical: $\alpha = 0.05$ | 0.156 | 9.26 | 0.176 | 8.27 | 0.199 | 6.9 |
| Classical: $\alpha = 0.10$ | 0.268 | 9.61 | 0.284 | 8.95 | 0.299 | 8 |
| Bonferroni: $\alpha = 0.01$ | 0.003 | 3.48 | 0.003 | 1.87 | 0.003 | 0.94 |
| Bonferroni: $\alpha = 0.05$ | 0.005 | 5.28 | 0.008 | 3.34 | 0.01 | 1.96 |
| Bonferroni: $\alpha = 0.10$ | 0.01 | 6.06 | 0.014 | 4.11 | 0.019 | 2.51 |
| BH: $\alpha = 0.01$ | 0.008 | 5.39 | 0.008 | 2.96 | 0.006 | 1.41 |
| BH: $\alpha = 0.05$ | 0.04 | 7.68 | 0.038 | 5.45 | 0.036 | 3.27 |
| BH: $\alpha = 0.10$ | 0.079 | 8.5 | 0.076 | 6.77 | 0.072 | 4.63 |
| Bootstrap: $\alpha = 0.01$ | 0.011 | 5.5 | 0.01 | 3.1 | 0.008 | 1.52 |
| Bootstrap: $\alpha = 0.05$ | 0.05 | 7.89 | 0.046 | 5.71 | 0.044 | 3.42 |
| Bootstrap: $\alpha = 0.10$ | 0.101 | 8.67 | 0.096 | 7.03 | 0.091 | 4.74 |
| PcGets/Autometrics | 0.202 | 9.88 | 0.211 | 9.55 | 0.234 | 8.83 |
| HP | 0.05 | 9.66 | 0.071 | 9.04 | 0.106 | 7.9 |
| Bayesian model averaging | | | | | | |
| $m = k/2$, $g = k^{-2}$, random $\theta$ | 0.007 | 6.57 | 0.022 | 7.9 | 0.044 | 6.51 |
| $m = 7$, $g = k^{-2}$, random $\theta$ | 0.006 | 6.23 | 0.021 | 7.79 | 0.044 | 6.41 |
| $m = k/2$, $g = k^{-2}$, fixed $\theta$ | 0.037 | 9.5 | 0.044 | 8.93 | 0.064 | 7.64 |
| $m = 7$, $g = k^{-2}$, fixed $\theta$ | 0.008 | 7.01 | 0.021 | 7.72 | 0.045 | 6.39 |
| $m = k/2$, $g = n^{-1}$, random $\theta$ | 0.059 | 9.59 | 0.047 | 8.92 | 0.057 | 7.41 |
| $m = 7$, $g = n^{-1}$, random $\theta$ | 0.05 | 9.48 | 0.04 | 8.82 | 0.053 | 7.28 |
| $m = k/2$, $g = n^{-1}$, fixed $\theta$ | 0.164 | 9.88 | 0.14 | 9.56 | 0.142 | 8.62 |
| $m = 7$, $g = n^{-1}$, fixed $\theta$ | 0.025 | 9.29 | 0.029 | 8.55 | 0.047 | 7.06 |
| Sala-i-Martin | 0.328 | 9.44 | 0.761 | 10 | 0.78 | 10 |
| Lasso | 0.336 | 9.85 | 0.425 | 9.96 | 0.451 | 9.89 |
| GSA algorithm | 0.018 | 8.50 | 0.035 | 8.15 | 0.064 | 7.26 |

slightly less, than the other methods. Particularly high CR values result from the Lasso, Sala-i-Martin, and Autometrics methods, for example.

However, the FDR of the GSA algorithm is lower than most of the other methods apart from the classical and Bonferroni approaches, with one or two exceptions. Therefore the GSA algorithm could be viewed as giving a good, but conservative performance. In fact, only in three experiments different versions of BMA outperform the GSA algorithm in both CR and FDR simultaneously for $\rho = 0.3$ and 0.5. While the GSA algorithm is not an outright winner in this experiment, it appears to be competitive with other approaches, even without altering the tuning parameters from default values; this appears promising.

**Table 7:** Results of Monte Carlo experiment with 25 false hypotheses.

| | $\rho = 0$ | | $\rho = 0.3$ | | $\rho = 0.5$ | |
|---|---|---|---|---|---|---|
| | **FDR** | **CR** | **FDR** | **CR** | **FDR** | **CR** |
| Classical: $\alpha = 0.01$ | 0.011 | 19.83 | 0.013 | 15.39 | 0.017 | 11.32 |
| Classical: $\alpha = 0.05$ | 0.048 | 23.11 | 0.054 | 20.61 | 0.063 | 17.27 |
| Classical: $\alpha = 0.10$ | 0.09 | 24.01 | 0.098 | 22.38 | 0.103 | 19.98 |
| Bonferroni: $\alpha = 0.01$ | 0 | 8.79 | 0.004 | 4.55 | 0.001 | 2.28 |
| Bonferroni: $\alpha = 0.05$ | 0.001 | 13.2 | 0.002 | 8.27 | 0.003 | 4.78 |
| Bonferroni: $\alpha = 0.10$ | 0.003 | 15.2 | 0.004 | 10.2 | 0.005 | 6.4 |
| BH: $\alpha = 0.01$ | 0.005 | 16.67 | 0.004 | 10.18 | 0.004 | 5.16 |
| BH: $\alpha = 0.05$ | 0.025 | 21.61 | 0.026 | 17.34 | 0.024 | 11.81 |
| BH: $\alpha = 0.10$ | 0.05 | 23.07 | 0.05 | 20.1 | 0.048 | 15.77 |
| Bootstrap: $\alpha = 0.01$ | 0.008 | 18.15 | 0.008 | 11.32 | 0.006 | 5.6 |
| Bootstrap: $\alpha = 0.05$ | 0.046 | 22.8 | 0.042 | 19.31 | 0.034 | 13.18 |
| Bootstrap: $\alpha = 0.10$ | 0.095 | 23.93 | 0.087 | 21.82 | 0.075 | 17.27 |
| PcGets/Autometrics | 0.072 | 24.31 | 0.087 | 22.97 | 0.109 | 20.56 |
| HP | 0.022 | 22.19 | 0.037 | 20.01 | 0.059 | 17.19 |
| Bayesian model averaging | | | | | | |
| $m = k/2$, $g = k^{-2}$, random $\theta$ | 0.005 | 4.17 | 0.035 | 17.26 | 0.044 | 13.97 |
| $m = 7$, $g = k^{-2}$, random $\theta$ | 0.004 | 3.02 | 0.035 | 16.52 | 0.045 | 13.45 |
| $m = k/2$, $g = k^{-2}$, fixed $\theta$ | 0.028 | 17.05 | 0.036 | 19.04 | 0.046 | 16.06 |
| $m = 7$, $g = k^{-2}$, fixed $\theta$ | 0.008 | 2.86 | 0.037 | 12.82 | 0.047 | 11.45 |
| $m = k/2$, $g = n^{-1}$, random $\theta$ | 0.107 | 24.13 | 0.034 | 19.7 | 0.034 | 14.26 |
| $m = 7$, $g = n^{-1}$, random $\theta$ | 0.077 | 23.45 | 0.031 | 18.75 | 0.033 | 13.56 |
| $m = k/2$, $g = n^{-1}$, fixed $\theta$ | 0.068 | 23.98 | 0.036 | 20.4 | 0.037 | 16.57 |
| $m = 7$, $g = n^{-1}$, fixed $\theta$ | 0.02 | 12.15 | 0.03 | 13.93 | 0.035 | 10.83 |
| Sala-i-Martin | 0.155 | 17.21 | 0.494 | 24.96 | 0.499 | 25 |
| Lasso | 0.285 | 24.8 | 0.255 | 24.93 | 0.263 | 24.8 |
| GSA algorithm | 0.013 | 17.58 | 0.019 | 19.42 | 0.031 | 17.14 |

The results of the same experiment with 10 false hypotheses are shown in Table 6. Similar to the previous results, the GSA algorithm performs a good-quality, but slightly conservative model selection. The CR values show that it is correctly rejecting 7.3–8.5 hypotheses depending on the correlation between variables. Other methods, such as the Lasso and BMA, tend to have a higher CR. However the GSA algorithm again has a very low FDR, which is only bettered by other approaches at the expense of a low CR. The exceptions to this are three instances of the BMA approach, one at $\rho = 0.3$ and two at $\rho = 0.5$, where CR is marginally higher and FDR marginally lower. Consider however that the BMA results here reflect the performance at eight different values of tuning parameters, while the GSA algorithm uses only default values.

Finally, Table 7 gives the results in the case of 25 false hypotheses. Again the results are similar to the other two cases. The GSA algorithm gives slightly lower rates of correct rejection than Autometrics/HP, BMA, Sala-i-Martin and Lasso; however it has a FDR which is lower than most of these approaches. The GSA algorithm is outperformed both on CR and FDR only in two experiments by the classical procedure and a bootstrap one.

The BMA approach actually performs quite well on this case study. However as Decker and Hanck note, there is no one configuration of the BMA approach that performs consistently the best—particular tuning parameter values are suited to particular cases.

Overall, in spite of the fact that GSA was run with default tuning coefficients, the results are encouraging for the GSA algorithm. While it does not outperform all competitors, it has a performance which is competitive with other approaches and can be adjusted to give higher potency by changing the values of its tuning parameters.

## 6.2 An Empirical Growth Model

The final test case is also taken from DH and represents a real case study on economic data. This case study (more details of which can be found in their paper) takes the data set of Fernandez, Ley, and Steel (2001) to build an empirical growth model. Growth models attempt to explain the differences in economic growth across a set of countries, over a fixed period of time, in terms of the number of candidate explanatory variables. This problem is of course different from the previous test cases, in that $y$ has not been generated from the candidate regressors and therefore there is no "true DGP" contained within the set of candidates. Indeed, cross-country growth may well be affected by many other variables outside of the list considered here.

The data set consists of $n = 72$ countries, whose growth is measured over the period 1960–1992, as well as $p = 41$ explanatory variables. Table 8 shows the results of the growth regression using the same list of model selection approaches considered in the previous example, with the addition of the GSA algorithm. For the classical, "boot" and BH columns, the number represents the level of significance at which the hypothesis is rejected; if no number is reported the regressor is not included in the final model. The BMA column can be interpreted as rejecting the null hypothesis when the probability of inclusion is greater than 0.5. The S-i-M column gives frequency of inclusion, with an asterisk denoting a significant relation to growth, and a double asterisk indicating that variables are always included. The FDP column refers to a method of Romano et al. (2006)

**Table 8:** Results for growth regression case study.

| | Regressor | $\widehat{\beta}_l$ | *p*-value | Classical (%) | BH (%) | Boot (%) | BMA | S-i-M | H&K | Lasso | FDP (%) | GSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GDP level 1960 | −0.017 | 0.00001 | 1 | 1 | 1 | 1 | 1.00** | 1 | 1 | 1 | 1 |
| 2 | Fraction confucian | 0.075 | 0.00003 | 1 | 1 | 1 | 0.995 | 1.00* | 1 | 1 | 1 | 1 |
| 3 | Life expectancy | 0.001 | 0.003 | 1 | 5 | 5 | 0.946 | 0.999** | 1 | 1 | 20 | 1 |
| 4 | Equipment investment | 0.127 | 0.008 | 1 | 5 | 5 | 0.942 | 1.00* | 1 | 1 | – | 1 |
| 5 | Sub-Saharan dummy | −0.02 | 0.006 | 1 | 5 | 5 | 0.757 | 0.997* | 1 | 1 | – | 1 |
| 6 | Fraction Muslim | 0.011 | 0.227 | – | – | – | 0.656 | 1.00* | 0 | 1 | – | 0 |
| 7 | Rule of law | 0.012 | 0.068 | 10 | – | – | 0.516 | 1.00* | 0 | 1 | – | 1 |
| 8 | Number of years open economy | −0.003 | 0.62 | – | – | – | 0.502 | 1.00* | 1 | 0 | – | 1 |
| 9 | Degree of capitalism | 0.001 | 0.284 | – | – | – | 0.471 | 0.987* | 0 | 0 | – | 0 |
| 10 | Fraction protestant | −0.003 | 0.677 | – | – | – | 0.461 | 0.966* | 0 | 0 | – | 1 |
| 11 | Fraction GDP in mining | 0.04 | 0.008 | 1 | 5 | 5 | 0.441 | 0.994* | 1 | 1 | – | 1 |
| 12 | Non-equipment investment | 0.037 | 0.081 | 10 | – | – | 0.431 | 0.982* | 0 | 1 | – | 0 |
| 13 | Latin American dummy | −0.013 | 0.039 | 5 | – | 10 | 0.19 | 0.998* | 1 | 1 | – | 1 |
| 14 | Primary school enrolment, 1960 | 0.02 | 0.045 | 5 | – | 10 | 0.184 | 0.992** | 1 | 1 | – | 1 |
| 15 | Fraction Buddhist | 0.007 | 0.276 | | – | – | 0.167 | 0.964* | 0 | 0 | – | 0 |
| 16 | Black-market premium | −0.007 | 0.075 | 10 | – | – | 0.157 | 0.825 | 0 | 0 | – | 0 |
| 17 | Fraction catholic | 0.003 | 0.593 | – | – | – | 0.11 | 0.963* | 0 | 0 | – | 0 |
| 18 | Civil liberties | −0.002 | 0.321 | – | – | – | 0.1 | 0.997* | 0 | 0 | – | 0 |
| 19 | Fraction Hindu | −0.097 | 0.001 | 1 | 1 | 1 | 0.097 | 0.654 | 1 | 1 | 5 | 1 |
| 20 | Political rights | 0.0002 | 0.934 | – | – | – | 0.071 | 0.998* | 0 | 0 | – | 0 |
| 21 | Primary exports, 1970 | −0.006 | 0.421 | – | – | – | 0.069 | 0.990* | 0 | 0 | – | 1 |
| 22 | Exchange rate distortions | −0.00002 | 0.538 | – | – | – | 0.06 | 0.968* | 0 | 0 | – | 0 |
| 23 | Age | −0.00001 | 0.774 | – | – | – | 0.058 | 0.903 | 0 | 0 | – | 0 |
| 24 | War dummy | −0.001 | 0.548 | – | – | – | 0.052 | 0.984* | 0 | 0 | – | 0 |
| 25 | Size labor force | 3.00E-07 | 0.004 | 1 | 5 | 5 | 0.047 | 0.835 | 1 | 1 | – | 0 |
| 26 | Fraction speaking foreign language | −0.002 | 0.468 | – | – | – | 0.047 | 0.831 | 0 | 0 | – | 0 |

**Table 8:** (continued)

| | Regressor | $\widehat{\beta}_l$ | *p*-value | Classical (%) | BH (%) | Boot (%) | BMA | S-i-M | H&K | Lasso | FDP (%) | GSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | Fraction of pop speaking English | −0.007 | 0.131 | – | – | – | 0.047 | 0.91 | 0 | 0 | – | 0 |
| 28 | Ethnologic fractionalization | 0.014 | 0.012 | 5 | 5 | 5 | 0.035 | 0.643 | 1 | 1 | – | 0 |
| 29 | Spanish colony dummy | 0.013 | 0.022 | 5 | 10 | 10 | 0.034 | 0.938* | 1 | 0 | – | 0 |
| 30 | SD of black-market premium | −0.000001 | 0.892 | – | – | – | 0.031 | 0.993* | 0 | 0 | – | 0 |
| 31 | French colony dummy | 0.009 | 0.038 | 5 | – | 10 | 0.031 | 0.702 | 1 | 0 | – | 0 |
| 32 | Absolute latitude | −0.0001 | 0.521 | – | – | – | 0.024 | 0.980* | 0 | 0 | – | 0 |
| 33 | Ratio of workers to population | −0.001 | 0.945 | – | – | – | 0.024 | 0.766 | 0 | 0 | – | 0 |
| 34 | Higher education enrolment | −0.129 | 0.002 | 1 | 5 | 5 | 0.024 | 0.579 | 1 | 1 | 20 | 0 |
| 35 | Population growth | −0.119 | 0.609 | – | – | – | 0.022 | 0.807 | 0 | 0 | – | 0 |
| 36 | British colony dummy | 0.007 | 0.072 | 10 | – | – | 0.022 | 0.579 | 1 | 0 | – | 0 |
| 37 | Outward orientation | −0.005 | 0.036 | 5 | – | 10 | 0.021 | 0.634 | 0 | 0 | – | 0 |
| 38 | Fraction Jewish | −0.001 | 0.942 | – | – | – | 0.019 | 0.747 | 0 | 0 | – | 0 |
| 39 | Revolutions and coups | 0.003 | 0.503 | – | – | – | 0.017 | 0.995* | 0 | 0 | – | 0 |
| 40 | Public education share | 0.137 | 0.249 | – | – | – | 0.016 | 0.58 | 0 | 0 | – | 0 |
| 41 | Area (scale effect) | 3.00E-07 | 0.637 | – | – | – | 0.016 | 0.532 | 0 | 0 | – | 0 |

which aims to ensure that the probability of the proportion of false rejections does not exceed a set value. Finally, the H&K, Lasso and GSA columns use a 1 to indicate that the variable is included, and a zero to indicate that it is not included.

The results show that the GSA algorithm agrees with the majority of other approaches in selecting the regressors GDP level, Fraction Confucian and Fraction Hindu as being robustly related to growth. The results are in fact broadly similar with H&K and the Lasso, with a few exceptions. For example, the GSA algorithm identifies fraction protestant as an additional predictor of growth, which is not selected by other algorithms apart from S-i-M. Similarly, primary exports in 1970 is found by the GSA algorithm to be significant. However, the GSA algorithm does not select variables such as ethnologic fractionalization and higher education enrollment, which are included by most other approaches. Overall the performance of GSA algorithm is slightly more parsimonious than other methods, and this agrees with the results from the simulation experiments in the previous section, although the degree of conservatism can be adjusted by the tuning parameter.

# 7 Conclusions

In the model selection problem, one has to choose which candidate regressors to include in a regression model. The approach in this paper is to view the problem as a sensitivity analysis of a measure of fit in the space of candidate variables. One can therefore calculate the sensitivity, e.g. of the BIC with respect to the presence (or absence) of each candidate variable.

These interactions are in principle relevant, as the importance of including a given regressor is conditioned by inclusion or exclusion of the other regressors. For this reason it is appropriate to use $S_T$, a sensitivity measure capable of appreciating the sensitivity of a trigger for the presence of one regressor, inclusive of its interaction effects with triggers for all other regressors.

The proposed algorithm uses the ordering of the regressors via GSA or the *t*-ratios within a testing strategy based on the 'Pantula-principle', see Pantula (1989). This implies a reduction in the number of tests for each given ordering (with an associated saving of computing times) and the favorable control of the size of the testing sequence.

When compared to the general-to-specific algorithm described by HP, and to Autometrics, the GSA algorithm performs well on the simulation experiments investigated, both in the theoretical case where tuning parameters were known, and in average performance in the practical situation when tuning parameters are unknown. The robustness of the algorithm to its main tuning parameter is a particularly positive feature, since the optimal values would not be known in a

practical case. When compared to a wider range of approaches, the GSA approach performs competitively, giving comparable performance to a range of well-established approaches, without adjusting its tuning parameters in any way. This shows that GSA methods can help in model selection.

This study is a first exploration of the use of GSA in the world of model selection; it shows that GSA can be fruitfully used to order regressors by importance. These results call for more research on the use of GSA methods in model selection.

# Appendix A: Proofs

In this section, Assumption 1 is maintained throughout, $\sigma^2_{Ti}$ is first expressed as a sum of terms involving $\widehat{\sigma}^2_{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma} \in \Gamma$ in Lemma 3; next the large $n$ behavior of $\widehat{\sigma}^2_{\boldsymbol{\gamma}}$ is discussed in Lemma 4. Lemma 5 states the probability limit of $\sigma^2_{Ti}$. Lemma 6 shows that, in case the true regressors in the DGP and the irrelevant ones are uncorrelated, $\sigma^2_{Ti} \xrightarrow{p} 0$ for an irrelevant regressor $i$, while $\sigma^2_{Ti} \xrightarrow{p} c_i > 0$ for a relevant one, where $\xrightarrow{p}$ indicates convergence in probability as $n \to \infty$. Under the same conditions, Lemma 6 proves Theorem 2, which shows that for large samples, a scree plot on the ordered $S_{Ti}$ allows to separate the relevant regressors from the irrelevant ones.

Let $\gamma_i = \boldsymbol{e}_i' \boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_{-i} = \boldsymbol{A}_i' \boldsymbol{\gamma}$, where $\boldsymbol{e}_i$ is the $i$-th column of the identity matrix of order $p$, $I_p$ and $A_i$ is a $p \times (p-1)$ matrix containing all the columns of $I_p$ except the $i$-th one. Next indicate $\boldsymbol{q}(\gamma)$ as $q(\gamma_i, \boldsymbol{\gamma}_{-i})$ or, more simply as $q_{-i}(\gamma_i)$. Denote by $\boldsymbol{\gamma}^{(i,0)}$ the vector corresponding to $\gamma_i = 0$, with the remaining coordinates equal to $\boldsymbol{\gamma}_{-i}$, and let $\boldsymbol{\gamma}^{(i,1)}$ the vector corresponding to $\gamma_i = 1$ with the remaining coordinates equal to $\boldsymbol{\gamma}_{-i}$. Finally let $\Gamma_{-i} := \{\boldsymbol{\gamma}_{-i} = \boldsymbol{A}_i' \boldsymbol{\gamma}, \boldsymbol{\gamma} \in \Gamma\}$.

**Lemma 3** ($\sigma^2_{Ti}$ *as an average over* $\boldsymbol{\gamma}_{-i}$). *One has*

$$\sigma^2_{Ti} = \mathbb{E}(\mathbb{V}(q|\boldsymbol{\gamma}_{-i})) = \frac{1}{4 \cdot 2^{p-1}} \sum_{\boldsymbol{\gamma}_{-i} \in \Gamma_{-i}} \left(q_{-i}(1) - q_{-i}(0)\right)^2 \tag{8}$$

*and for q equal to BIC* (*or any other consistent information criterion*)

$$q_{-i}(1) - q_{-i}(0) = \log\left(\frac{\widehat{\sigma}^2_{\boldsymbol{\gamma}^{(i,1)}}}{\widehat{\sigma}^2_{\boldsymbol{\gamma}^{(i,0)}}}\right) + o(1), \tag{9}$$

*where* $o(1)$ *is a non-stochastic term tending to* 0 *for large n and* $\widehat{\sigma}^2_{\boldsymbol{\gamma}} := n^{-1} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{\gamma}}' \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{\gamma}}$ *where* $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{\gamma}}$ *are the residuals of model* $\boldsymbol{\gamma}$.

*Proof*. Note that for $h = 1,2$ one has $\mathbb{E}(q^h|\boldsymbol{\gamma}_{-i}) = \frac{1}{2}(q^h_{-i}(1) + q^h_{-i}(0))$ so that

$$\begin{aligned}\mathbb{V}(q|\boldsymbol{\gamma}_{-i}) &= \mathbb{E}\left(q^2|\boldsymbol{\gamma}_{-i}\right) - \left(\mathbb{E}(q|\boldsymbol{\gamma}_{-i})\right)^2 \\ &= \frac{1}{2}(q^2_{-i}(1) + q^2_{-i}(0)) - \frac{1}{4}(q^2_{-i}(1) + q^2_{-i}(0) + 2q_{-i}(1)q_{-i}(0)) \\ &= \frac{1}{4}(q_{-i}(1) - q_{-i}(0))^2.\end{aligned}$$

Hence one finds (8). When $q$ is BIC, $q(\boldsymbol{\gamma}) = \log\widehat{\sigma}^2_{\boldsymbol{\gamma}} + k_{\boldsymbol{\gamma}}c_n$ with $c_n := \log(n)/n$. Other consistent information criteria replace log$n$ with some other increasing function $f(n)$ of $n$ with the property $c_n = f(n)/n \to 0$, see Paulsen (1984) Theorem 1. Note also that $k_{\boldsymbol{\gamma}^{(i,1)}} - k_{\boldsymbol{\gamma}^{(i,0)}} = 1$, and that one has

$$q_{-i}(1) - q_{-i}(0) = \log\left(\frac{\widehat{\sigma}^2_{\boldsymbol{\gamma}^{(i,1)}}}{\widehat{\sigma}^2_{\boldsymbol{\gamma}^{(i,0)}}}\right) + (k_{\boldsymbol{\gamma}^{(i,1)}} - k_{\boldsymbol{\gamma}^{(i,0)}})c_n = \log\left(\frac{\widehat{\sigma}^2_{\boldsymbol{\gamma}^{(i,1)}}}{\widehat{\sigma}^2_{\boldsymbol{\gamma}^{(i,0)}}}\right) + c_n.$$

Because $c_n \to 0$, one finds (9). □

The asymptotic behavior of $\widehat{\sigma}^2_{\boldsymbol{\gamma}}$ is next discussed. Let $w_t := (y_t, x_{1,t}, \ldots, x_{p,t}, \epsilon_t)'$, where, without loss of generality, it is assumed that all variables have mean zero. Denote $\Sigma := E(w_t w_t')$, where

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{y\boldsymbol{x}} & \sigma^2 \\ & \Sigma_{\boldsymbol{xx}} & 0 \\ & & \sigma^2 \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{y1} & \cdots & \Sigma_{yp} & \sigma^2 \\ & \Sigma_{11} & & \Sigma_{1p} & 0 \\ & & \ddots & & \\ & & & \Sigma_{pp} & 0 \\ & & & & \sigma^2 \end{pmatrix}.$$

Let $\Sigma_{ij.v} := \Sigma_{ij} - \Sigma_{iv}\Sigma^{-1}_{vv}\Sigma_{vj}$ indicate partial covariances, where $v := \{i_1, \ldots, i_s\}$ indicates a set of indices. Note that $\Sigma_{\boldsymbol{x}\epsilon} = 0$.

For each $\boldsymbol{\gamma}$, let $a_{\boldsymbol{\gamma}} := \{i_1, \ldots, i_{k_{\boldsymbol{\gamma}}}\}'$ indicate the set of indices $i_j$ such that $\gamma_{i_j} = 1$ in $\boldsymbol{\gamma}$. Similarly let $b_{\boldsymbol{\gamma}} := \{i_1, \ldots, i_s\}'$ indicate the set of indices $i_j$ that belong to $a_{\boldsymbol{\gamma}} \setminus \mathbb{T}$. The representation $\boldsymbol{\beta}_0$ as $\boldsymbol{\beta}_0 = \boldsymbol{H}_0\boldsymbol{\phi}_0$ is used here, where $\boldsymbol{H}_0$ contains the $r_0$ columns of $\boldsymbol{I}_p$ corresponding to $\mathbb{T}$, and $\boldsymbol{\phi}_0$ contains the corresponding $\beta_{0,i}$ coefficients. Moreover the matrix of regressors in the $\boldsymbol{\gamma}$ specification is written as $\boldsymbol{X}\boldsymbol{H}_{\boldsymbol{\gamma}}$, where $\boldsymbol{H}_{\boldsymbol{\gamma}}$ contains the columns of $\boldsymbol{I}_p$ with column indices $a_{\boldsymbol{\gamma}}$. Define also $\boldsymbol{M}_{\boldsymbol{\gamma}} := \boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{H}_{\boldsymbol{\gamma}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{H}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{H}_{\boldsymbol{\gamma}}'\boldsymbol{X}'$.

**Lemma 4** (*Large sample behavior of* $\widehat{\sigma}^2_{\gamma}$). *As* $n \to \infty$, *one has*

$$\widehat{\sigma}^2_{\boldsymbol{\gamma}} \xrightarrow{p} \sigma^2 + \sum_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}}} \beta_{0,h}\Sigma_{hj.b_{\boldsymbol{\gamma}}}\beta_{0,j}, \tag{10}$$

*where* $\mathbb{T} \setminus a_{\boldsymbol{\gamma}}$ *is the set of indices of the true regressors omitted from the* $\boldsymbol{\gamma}$ *specification, and* $b_{\boldsymbol{\gamma}}$ *is the set of indices* $a_{\boldsymbol{\gamma}} \setminus \mathbb{T}$ *of the regressors included in the* $\boldsymbol{\gamma}$ *specification except the ones that belong to the DGP. Remark that the sum in* (10) *is equal to* 0 *when* $\boldsymbol{\gamma}$ *is correctly specified (i.e. it contains all regressors in the DGP) i.e.* $\mathbb{T} \setminus a_{\boldsymbol{\gamma}} = \emptyset$.

*Proof*. Because $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{H}_0\boldsymbol{\phi}_0 + \boldsymbol{\varepsilon}$ one has

$$\widehat{\sigma}^2_{\boldsymbol{\gamma}} = n^{-1}\boldsymbol{y}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{y} = n^{-1}\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{\varepsilon} + 2n^{-1}\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{X}\boldsymbol{H}_0\boldsymbol{\phi}_0 + n^{-1}\boldsymbol{\phi}_0'\,\boldsymbol{H}_0'\,\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{X}\boldsymbol{H}_0\boldsymbol{\phi}_0$$

Because $\Sigma_{\boldsymbol{x}\epsilon} = 0$, by the law or large numbers for stationary linear processes, see e.g. Anderson (1971), one finds

$$n^{-1}\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{\varepsilon} \xrightarrow{p} \sigma^2 - \Sigma_{\epsilon\boldsymbol{x}}\boldsymbol{H}_{\boldsymbol{\gamma}}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{H}_{\boldsymbol{\gamma}}\right)^{-1}\boldsymbol{H}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\epsilon} = \sigma^2,$$

$$n^{-1}\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{X} \xrightarrow{p} \Sigma_{\epsilon\boldsymbol{x}}\left(\boldsymbol{I}_p - \boldsymbol{H}_{\boldsymbol{\gamma}}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{H}_{\boldsymbol{\gamma}}\right)^{-1}\boldsymbol{H}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\boldsymbol{x}}\right) = 0.$$

Similarly

$$n^{-1}\boldsymbol{H}_0'\,\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{X}\boldsymbol{H}_0 \xrightarrow{p} \boldsymbol{H}_0'\,\left(\Sigma_{\boldsymbol{x}\boldsymbol{x}} - \Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{H}_{\boldsymbol{\gamma}}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{H}_{\boldsymbol{\gamma}}\right)^{-1}\boldsymbol{H}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\boldsymbol{x}}\right)\boldsymbol{H}_0$$

$$= \boldsymbol{H}_0'\,\boldsymbol{V}_{\boldsymbol{\gamma}}\left(\boldsymbol{V}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\boldsymbol{x}}^{-1}\,\boldsymbol{V}_{\boldsymbol{\gamma}}\right)^{-1}\boldsymbol{V}_{\boldsymbol{\gamma}}'\,\boldsymbol{H}_0$$

where $\boldsymbol{V}_{\boldsymbol{\gamma}} = \boldsymbol{H}_{\boldsymbol{\gamma},\perp}$ contains the columns in $I_p$ not contained in $H_{\boldsymbol{\gamma}}$, and the last equality is a special case of a non-orthogonal projection identity, see e.g. eq. (2.13) in Paruolo and Rahbek (1999) and references therein. Here $\boldsymbol{H}_\perp$ indicates a basis of the orthogonal complement of the space spanned by the columns in $\boldsymbol{H}$. Observe that the $(p - k_{\boldsymbol{\gamma}}) \times r_0$ matrix $\boldsymbol{C}_{\boldsymbol{\gamma}} \coloneqq \boldsymbol{V}_{\boldsymbol{\gamma}}'\,\boldsymbol{H}_0$ contains the columns of $\boldsymbol{I}_{p-r_{\boldsymbol{\gamma}}}$ corresponding to the index set of regressors in $v_{\boldsymbol{\gamma}} \coloneqq \mathbb{T} \setminus a_{\boldsymbol{\gamma}}$. Hence, using e.g. eq. (A.4) in Paruolo and Rahbek (1999), one finds $\left(\boldsymbol{V}_{\boldsymbol{\gamma}}'\,\Sigma_{\boldsymbol{x}\boldsymbol{x}}^{-1}\,\boldsymbol{V}_{\boldsymbol{\gamma}}\right)^{-1} = \Sigma_{v_{\boldsymbol{\gamma}}v_{\boldsymbol{\gamma}}.b_{\boldsymbol{\gamma}}}$. Substituting one finds

$$n^{-1}\boldsymbol{\phi}_0'\,\boldsymbol{H}_0'\,\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{\gamma}}\boldsymbol{X}\boldsymbol{H}_0\boldsymbol{\phi}_0 \xrightarrow{p} \boldsymbol{\phi}_0'\,\boldsymbol{C}_{\boldsymbol{\gamma}}'\,\Sigma_{v_{\boldsymbol{\gamma}}v_{\boldsymbol{\gamma}}.b_{\boldsymbol{\gamma}}}\boldsymbol{C}_{\boldsymbol{\gamma}}\boldsymbol{\phi}_0.$$

Simplifying one obtains (10). □

**Lemma 5** (Large sample behavior of $\sigma^2_{Ti}$). *As* $n \to \infty$ *one has*

$$\sigma^2_{Ti} \xrightarrow{p} c_i \coloneqq \frac{1}{4 \cdot 2^{p-1}} \sum_{\boldsymbol{\gamma}_{-i} \in \Gamma_{-i}} \log\left(\frac{\sigma^2 + \sum\limits_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}(i,1)}} \beta_{0,h}\Sigma_{hj.b_{\boldsymbol{\gamma}(i,1)}}\beta_{0,j}}{\sigma^2 + \sum\limits_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}(i,0)}} \beta_{0,h}\Sigma_{hj.b_{\boldsymbol{\gamma}(i,0)}}\beta_{0,j}}\right)$$

*where* $a_{\boldsymbol{\gamma}}$ *is the set of indices of the regressors in the* $\boldsymbol{\gamma}$ *specification, and* $b_{\boldsymbol{\gamma}} \coloneqq a_{\boldsymbol{\gamma}} \setminus \mathbb{T}$ *includes the indices of regressors included in the* $\boldsymbol{\gamma}$ *specification except the ones that belong to the DGP.*

*Proof*. Apply Lemma 3 and 4. □

Lemma 5 shows that the limit behavior of $\sigma^2_{Ti}$ depends on the covariance structure $\Sigma$. Some covariance structures imply that, in the limit, the value of $S_T$ for true regressors is greater than the value of $S_T$ for irrelevant regressors. There also exist other covariance structures which can imply a reverse ordering.[21] In the special case when true and irrelevant regressors are uncorrelated, the next Lemma 6 shows that $S_T$ converges to 0 for irrelevant regressors, while $S_T$ converges to a positive constant for true regressors. This result proves Theorem 2 that shows that the ordering based on $S_T$ separates true and irrelevant regressors in this special case.

**Lemma 6** (Orthogonal regressors in $\mathbb{M}$ and $\mathbb{T}$). *Assume that* $\Sigma_{\ell j} = 0$ *for all* $j \in \mathbb{T}$ *and* $\ell \in \mathbb{M}$. *Then when* $i \in \mathbb{M}$ *one has, as* $n \to \infty$, $\sigma^2_{Ti} \xrightarrow{p} 0$, *whereas otherwise when* $i \in \mathbb{T}$ *one finds*

$$\sigma^2_{Ti} \xrightarrow{p} c_i > 0. \tag{11}$$

*Proof*. From Lemma 4, one finds

$$q_{-i}(1) - q_{-i}(0) = \log\left(\frac{\sigma^2 + \sum\limits_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}(i,1)}} \beta_{0,h}\Sigma_{hj_{\boldsymbol{\gamma}(i,1)}.b}\beta_{0,j}}{\sigma^2 + \sum\limits_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}(i,0)}} \beta_{0,h}\Sigma_{hj_{\boldsymbol{\gamma}(i,0)}.b}\beta_{0,j}}\right) + o_p(1), \tag{12}$$

Assume that $\Sigma_{\ell j} \neq 0$ for some $j \in \mathbb{T}$ and $\ell \in \mathbb{M}$; then for some $\boldsymbol{\gamma}_{-i} \in \Gamma_{-i}$ one has $\Sigma_{hj.b_{\boldsymbol{\gamma}(i,1)}} \neq \Sigma_{hj.b_{\boldsymbol{\gamma}(i,0)}}$ in the numerator and denominator on the r.h.s. of (12); let $c \neq 1$ indicate the corresponding ratio. Hence $(q_{-i}(1) - q_{-i}(0))^2$ converges in probability to $\log^2 c > 0$, and because the terms in $\mathbb{E}(\mathbb{V}(q|\boldsymbol{\gamma}_{-i})) = \frac{1}{4 \cdot 2^{p-1}} \sum\limits_{\boldsymbol{\gamma}_{-i} \in \Gamma_{-i}} (q_{-i}(1) - q_{-i}(0))^2$, see Lemma 5, are non-negative, one concludes that $\sigma^2_{Ti} \xrightarrow{p} c_i > 0$.

Assume instead that $\Sigma_{\ell j} = 0$ for all $j \in \mathbb{T}$ and $\ell \in \mathbb{M}$ and $i \in \mathbb{M}$. Then $\mathbb{T} \setminus a_{\boldsymbol{\gamma}^{(i)}} = \mathbb{T} \setminus a_{\boldsymbol{\gamma}_{-i}}$ and, because $\Sigma_{\ell j} = 0$ for all $j \in \mathbb{T}$ and $\ell \in \mathbb{M}$, one has $\Sigma_{jb_{\boldsymbol{\gamma}(i,\cdot)}} = 0$. This implies $\Sigma_{hj.b_{\boldsymbol{\gamma}(i,\cdot)}} := \Sigma_{hj} - \Sigma_{hb_{\boldsymbol{\gamma}(i,\cdot)}}\Sigma^{-1}_{b_{\boldsymbol{\gamma}(i,\cdot)}b_{\boldsymbol{\gamma}(i,\cdot)}}\Sigma_{b_{\boldsymbol{\gamma}(i,\cdot)}j} = \Sigma_{hj}$. Hence

$$q_{-i}(1) - q_{-i}(0) = \log\left(\frac{\sigma^2 + \sum\limits_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}_{-i}}} \beta_{0,h}\Sigma_{hj}\beta_{0,j}}{\sigma^2 + \sum\limits_{h,j \in \mathbb{T} \setminus a_{\boldsymbol{\gamma}_{-i}}} \beta_{0,h}\Sigma_{hj}\beta_{0,j}}\right) + o_p(1) = o_p(1),$$

for all $\boldsymbol{\gamma}_{-i} \in \Gamma_{-i}$ because the numerator and denominator are identical. Thus $(q_{-i}(1) - q_{-i}(0))^2$ converges in probability to $\log^2 1 = 0$ for all $\boldsymbol{\gamma}_{-i} \in \Gamma_{-i}$, and this implies $\sigma^2_{Ti} \xrightarrow{p} 0$. □

**21** Worked out examples illustrating both situations are available from the authors upon request.

# Appendix B: HP design and algorithm

**B.1 HP DGPs**. HP's experiments are constructed as follows. Following Lovell (1983), HP chose a set of 18 major US quarterly macroeconomic variables. Only two variables considered in Lovell (1983) were discarded in HP, namely the linear trend and the 'potential level of GNP in $1958', because they were no longer relevant or available. Unlike in Lovell (1983), HP applied 0, 1 or 2 differences to the data; the order of differencing was selected by HP in order to obtain stationary variables according to standard unit root tests, see their Table 1.

The values of these (differenced) 18 major US quarterly macroeconomic series are then fixed in HP's experiments; they are here indicated as $x^{\star}_{it}$, where $t = 1, \ldots, n$ indicates quarters and $i = 1, \ldots, k$, with $k = 18$ indexes variables. The values of $y_t$ were then generated by the following scheme

$$y_t = \sum_{i=1}^{k} \beta_i^{\star} x_{it}^{\star} + u_t, \ \rho(L) u_t = \epsilon_t, \tag{13}$$

where $\varepsilon_t$ are i.i.d. $N(0, \sigma_{\epsilon}^2)$, and $\rho(z) = 1 - \rho z$ for all DGPs except for DGP 3, for which $\rho(z) = 1 - \rho_1 z - \rho_2 z^2$. Here $\beta_i^{\star}$ for $i = 1, \ldots, k$ and $\sigma_{\epsilon}^2$ are known constants, which define the DGP. In practice $\varepsilon_t s$ are simulated using a computer random number generator, $u_t$ is then calculated as an autoregressive series of order 1, AR(1), with coefficient $\rho$. $u_t$ is then fed into the equation for $y_t$, where $x_{it}^{\star}$ are kept fixed and do not change across replications.

It is useful to express (13) as a special case of (1). To this end one can substitute $(y_t - \sum_{i=1}^{k} \beta_i^{\star} x_{it}^{\star})$ in place of $u_t$ in the dynamic equation of $u_t$; one hence finds the following equivalent representation of the DGP

$$y_t = \rho y_{t-1} + \sum_{i=1}^{2k} \beta_i x_{it} + \epsilon_t \tag{14}$$

for all DGPs except for DGP 3, where $\beta_i = \beta_i^{\star}$ and $x_{it} = x_{it}^{\star}$ for $i = 1, \ldots, k$ while $\beta_i = -\rho \beta_i^{\star}$ and $x_{it} = x_{it-1}^{\star}$ for $i = k+1, \ldots, 2k$. For DGP 3, one has $y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t$. Both these representations are of the form (1), and the parameters can be estimated as in (3).

Regressions in HP were performed setting the elements $x_{i,t}$ in column $\boldsymbol{X}_i$ equal to variable $x_{it}$ from (14), for $i = 1, \ldots, 2k$ with $2k = 36$, and setting the elements $x_{i,t}$ of the remaining columns $\boldsymbol{X}_i$ for $i = 2k + 1, \ldots, p$, i.e. from 37 to 40, equal to the first, second, third and fourth lag of $y_t$. Therefore, four lags were always considered in estimation regardless of how many lags are in the DGP, and the only part of the $\boldsymbol{X}$ that changes across replications is block of the last four columns.

HP defined 11 DGPs by choosing values for the parameters $\rho$, $\beta_i^\star$ and $\sigma_\epsilon^2$. Table 9 summarizes the chosen parameter values. The choice of these values was made to reflect the coefficient estimates obtained on US data, using personal consumption expenditure as dependent variable, following the rationale in Lovell (1983). Because they were chosen as explanatory variables for a consumption equation, not all the macroeconomic time series were included in the DGP; in particular only (the second differences of the) Government purchases on goods and services $G$ and the (first differences of the) $M1$ monetary aggregate, and their respective first lags, were included in the experiments.

**B.2 HP algorithm**. This subsection gives an overview of the algorithm proposed by HP, following Hansen (1999). The algorithm proposed in HP aimed to provide a close approximation to a subset of what practitioners of the LSE approach actually do; further details can be found in the original reference.

The HP algorithm can be described by a choice of a triplet $(R, f, \Gamma_s)$ composed of (i) a test procedure $R$, (ii) a measure of fit $f$ and (iii) a subset $\Gamma_s$ of all models $\Gamma$, $\Gamma_s \subseteq \Gamma$. For any model $\boldsymbol{\gamma}$, the test procedure $R$ is defined as

$$R(\boldsymbol{\gamma}) = 1\Big(\min_{1\le \ell \le v} p_\ell \le \alpha\Big) \tag{15}$$

where $p_\ell$ are the $p$-values of $v$ specification tests and $\alpha$ is the chosen significance level. Note that $R(\boldsymbol{\gamma}) = 0$ when all $v$ tests do not reject the null, which corresponds to the hypothesis of correct specification and/or constant parameters.[22]

HP's measure of fit $f$ is based on the least-square estimate of $\sigma^2$, the regression variance, which equals $\tilde{\sigma}^2_{\boldsymbol{\gamma}} \coloneqq \frac{1}{n-k_{\boldsymbol{\gamma}}} \widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{\gamma}} \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{\gamma}}$, where $k_{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{\gamma}}$ are the number of regressors and the residuals in model $\boldsymbol{\gamma}$. HP's measure of fit is $f(\boldsymbol{\gamma}) = \tilde{\sigma}_{\boldsymbol{\gamma}}$, which should be minimized. Finally the subset $\Gamma_s$ is selected recursively, going from general to specific models, starting from the GUM, $\boldsymbol{\gamma} = \boldsymbol{\iota}_p$; the recursion continues as long as $R(\boldsymbol{\gamma}) = 0$. Details on HP's choice of $\Gamma_s$ are given in the next subsection.

Overall the HP algorithm selects a model $\widehat{\boldsymbol{\gamma}}$ as the preferred model using the rule

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\Gamma_s : R(\boldsymbol{\gamma})=0}{\arg\min} f(\boldsymbol{\gamma}).$$

The above description shows that the HP algorithm depends on $\alpha$, which is a tuning parameter, as well as on the choice of specific path $\Gamma_s$. For large $n$, Hansen (1999)

**22** The tests are the following: (1) Jarque Bera test for normality of residuals; (2) Breusch Pagan residual autocorrelation tests; (3) Engle's ARCH test on residuals; (4) Chow sample-split parameter stability tests; (5) Chow out-of-sample stability test using the first 90% of observations versus the last 10%; (6) $F$ test of the restrictions imposed by model $\gamma$ versus the GUM. The tests are performed on the first 90% of observations during the search.

**Table 9:** DGPs design. $y_{t-j}$ indicates lags of the dependent variable, $G_{t-j}$ denotes (lags of) second differences of government purchases of goods and services and $M1_{t-j}$ indicates (lags of) first differences of M1.

| DGP | 1 | 2 | 3[♦] | 4 | 5 | 6 | 6A | 6B | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Coefficients in DGP | | | | | |
| $y_{t-1}$ | | 0.75 | 0.395 | | | | | | 0.75 | 0.75 | 0.75 |
| $y_{t-2}$ | | | 0.3995 | | | | | | | | |
| $G_t$ | | | | | −0.046 | −0.023 | −0.32 | −0.65 | | −0.046 | −0.023 |
| $G_{t-1}$ | | | | | | | | | | 0.00345 | 0.01725 |
| $M1_t$ | | | | 1.33 | | 0.67 | 0.67 | 0.67 | 1.33 | | 0.67 |
| $M1_{t-1}$ | | | | | | | | | −0.9975 | | −0.5025 |
| $\sigma_\varepsilon$ | 130 | 85.99 | 0.00172 | 9.73 | 0.11 | 4.92 | 4.92 | 4.92 | 6.73 | 0.073 | 3.25 |

[♦]: in DGP 3 the regression analysis is performed on $y_t^* = \exp(y_t)$, where $y_t$ is simulated as in (14).

noted that $\widehat{\boldsymbol{y}}$ corresponds approximately to minimizing the information criterion $HP(\boldsymbol{y}) = \log\widehat{\sigma}^2_{\boldsymbol{y}} + k_{\boldsymbol{y}}/n$, where $\widehat{\sigma}^2_{\boldsymbol{y}} \coloneqq \frac{1}{n}\widehat{\boldsymbol{\varepsilon}}'_{\boldsymbol{y}}\,\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{y}}$ is the ML estimator of $\sigma^2$. This differs from Akaike's information criterion $AIC(\boldsymbol{y}) = \log\widehat{\sigma}^2_{\boldsymbol{y}} + 2k_{\boldsymbol{y}}/n$ and from the Bayesian information criterion of Schwarz $BIC(\boldsymbol{y}) = \log\widehat{\sigma}^2_{\boldsymbol{y}} + k_{\boldsymbol{y}}\log(n)/n$ by the different choice of penalty term.[23]

**B.3 HP's choice of search paths**. The choice of subset $\Gamma_s$ of $\Gamma$ is a critical aspect of the HP algorithm, as well as of any selection method based e.g. on information criteria, see Section 5.2. in Hansen (1999) and Burnham and Anderson (2002).

In particular, HP select a subset $\Gamma_s$ as follows. All paths start from the GUM regression, and the regressors are ranked in ascending order according their $t$-ratios. The 10 lowest variables in this list are then candidates for elimination; this starts an iterative elimination path. Each candidate model $\boldsymbol{y}_\star$ then becomes the current specification provided $R(\boldsymbol{y}_\star) = 0$. In this stage, the first 90% of the observations are used in the specification tests. Each search is terminated when for any choice of regressor the test $R$ rejects.

At this final stage, the HP algorithm reconsiders all the observations in a 'block search'; this consists in considering the joint elimination of all the regressors with an insignificant $t$-ratios. If the $R$ tests for the block search does not reject, the resulting model becomes the terminal specification. Otherwise, the specification that entered the final stage becomes the terminal specification. Once all 10 search paths have ended in a terminal specification, the final specification is the one among these with lowest $f(\boldsymbol{y}) = \tilde{\sigma}_{\boldsymbol{y}}$.

# Appendix C: Effective DGP

This appendix describes how the notion of 'weak regressors' was made operational in the present context. The 'Parametricness Index' (PI), Liu and Yang (2011), is used here to identify the 'effective DGP' (EDGP). Parametricness, in the sense of Liu and Yang, is a measure dependent both on sample size *and* the proposed model; a model is parametric if omission of any of its variables implies a marked change in its fit, and nonparametric otherwise.[24] Here parametricness is taken as a sign of detectability, i.e. of a sufficiently high signal-noise ratio. This concept is applied both to complete specifications as well as to single regressors; in particular the

**23** Remark that information criteria are equivalent to LR testing with a tunable significance level; see for instance Poetscher (1991).

**24** For example, consider a data set generated by a sine function, with added noise. If it is proposed to model this with a quadratic equation, the data/model should be considered nonparametric. However, if the proposed model included sinusoidal terms, it should be considered parametric.

EDGP is defined as the subset of DGP regressors which the PI would classify as parametric.

Considering a model $\boldsymbol{\gamma}_k \in \Gamma$, one can express the regression fit as $\hat{\boldsymbol{y}}_k = \boldsymbol{P}_k \boldsymbol{y}$, where $\boldsymbol{P}_k$ is the projection matrix on col $(\boldsymbol{X}\boldsymbol{H}_{\boldsymbol{\gamma}_k})$, and col indicates the column space; let $r_{\boldsymbol{\gamma}_k}$ be the dimension of col $(\boldsymbol{X}\boldsymbol{H}_{\boldsymbol{\gamma}_k})$. The index PI is defined in terms of an information criterion *IC*, which depends on $\lambda_n$, $d$ and $\hat{\sigma}^2$. Here $\lambda_n$ is a nonnegative sequence that satisfies $\lambda_n \geq (\log n)^{-1}$, $d$ is a nonnegative constant and $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2$ such as $\|\boldsymbol{y} - \hat{\boldsymbol{y}}_k\|^2/(n - r_{\boldsymbol{\gamma}_k})$ with $\gamma_k$ consistent for $\boldsymbol{\gamma}_{\mathbb{T}}$. In the application $\boldsymbol{\gamma}_k = \boldsymbol{\gamma}_{\mathbb{T}}$ was used. The information criterion IC is defined by

$$IC_{\lambda_n, d}(\boldsymbol{\gamma}_k, \hat{\sigma}^2) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}_k\|^2 + \lambda_n \log(n) r_k \hat{\sigma}^2 - n\hat{\sigma}^2 + d n^{1/2} \log(n) \hat{\sigma}^2 \tag{16}$$

where $\|\cdot\|$ represents Euclidean distance; here the values $\lambda_n = 1$ and $d = 0$ are used, as suggested in Liu and Yang (2011).

Let now $\boldsymbol{\gamma}_{\mathbb{T}}$ be the DGP; PI is defined in the present context as,

$$PI = \begin{cases} \inf\limits_{\gamma_k \in \Gamma_1(\boldsymbol{\gamma}_{\mathbb{T}})} \dfrac{IC_{\lambda_n, d}(\boldsymbol{\gamma}_k, \hat{\sigma}^2)}{IC_{\lambda_n, d}(\boldsymbol{\gamma}_{\mathbb{T}}, \hat{\sigma}^2)} & \text{if } r_{\boldsymbol{\gamma}_0} > 1 \\ n & \text{if } r_{\boldsymbol{\gamma}_{\mathbb{T}}} = 1 \end{cases} \tag{17}$$

where $\Gamma_1(\boldsymbol{\gamma}_{\mathbb{T}})$ is the set of submodels $\gamma_k$ of the DGP $\boldsymbol{\gamma}_{\mathbb{T}}$ such that $r_{\boldsymbol{\gamma}_k} = r_{\boldsymbol{\gamma}_{\mathbb{T}}} - 1$, i.e. all submodels obtained by removing one regressor at a time (with replacement).[25]

The reasoning is that if the model is parametric (and correctly specified for the data), removing any of the regressors will have a marked impact on *IC*. In contrast, if (some of the) regressors are just incremental terms in a nonparametric approximation, removing one of these regressors will have little effect on IC. Liu and Yang (2011) show that PI converges to 1 for a nonparametric scenario, and goes to infinity in a parametric scenario. The authors suggest to take $PI = 1.2$ as a cutoff point between parametric and nonparametric scenarios; this threshold was adopted in the present paper.

PI is applied in this paper at the level of each DGP; if PI indicates that the DGP is nonparametric, it is also investigated which of the submodels is responsible for this and label the corresponding omitted variables as 'weak'. As in the rest of the paper, a MC approach is used. Five thousand datasets are generated from each DGP and PI is calculated for each sample, hence obtaining a distribution of PI values. Table 10 summarizes the MC distribution of PI values through the empirical distribution function $F_m(x) = m^{-1} \sum_{j=1}^{m} 1(PI_j \leq x)$, where $m = N_R$ and $PI_j$ is the PI value in replication $j = 1, \ldots, N_R$. Quantiles of PI are indicated as $PI_\alpha$, with $\alpha = 0.01, 0.1, 0.9,$

**25** In the original paper $\boldsymbol{\gamma}_{\mathbb{T}}$ is replaced by the model $\hat{\gamma}_k$, selected by a weakly consistent information criterion, such as BIC.

**Table 10:** Distribution of PI values for DGPs 1–9. $F_n(\cdot)$ is the MC cumulative distribution function of PI and $PI_\alpha$ is the $\alpha$-quantile of $F_m(\cdot)$. DGPs where EDGP ≠ DGP are in boldface.

| DGP | DGP indices | $F_N(1.2)$ | $PI_{0.01}$ | $PI_{0.1}$ | $E_N(PI)$ | $PI_{0.9}$ | $PI_{0.99}$ | EDGP indices |
|---|---|---|---|---|---|---|---|---|
| 1 | {} | – | – | – | – | – | – | {} |
| 2 | {37} | 0.00 | 16.55 | 25.59 | 41.80 | 60.83 | 84.61 | {37} |
| 3 | {37, 38} | 0.04 | 0.88 | 1.53 | 2.54 | 3.52 | 4.17 | {37, 38} |
| 4 | {11} | 0.00 | 30.50 | 37.82 | 49.19 | 61.78 | 74.88 | {11} |
| 5 | {3} | 0.00 | 365.84 | 415.39 | 493.63 | 578.17 | 668.84 | {3} |
| 6 | {3, 11} | 0.98 | 0.37 | 0.38 | 0.53 | 0.79 | 1.40 | **{11}** |
| 6A | {3, 11} | 0.00 | 2.77 | 4.15 | 6.44 | 8.95 | 11.69 | {3, 11} |
| 6B | {3, 11} | 0.00 | 15.11 | 18.10 | 23.04 | 28.38 | 33.72 | {3, 11} |
| 7 | {11, 29, 37} | 0.00 | 2.84 | 4.16 | 6.46 | 8.96 | 11.76 | {11, 29, 37} |
| 8 | {3, 21, 37} | 0.00 | 5.77 | 8.40 | 13.49 | 19.22 | 26.41 | {3, 21, 37} |
| 9 | {3, 11, 21, 29, 37} | 1.00 | 0.75 | 0.75 | 0.77 | 0.81 | 0.93 | **{11, 29, 37}** |

**Table 11:** Distribution of ICRs for DGPs 6 and 9. Notation as in Table 10. Variables that are excluded from the EDGP are in boldface.

| DGP | Variable | $F_N(1.2)$ | $ICR_{0.01}$ | $ICR_{0.1}$ | $E$(ICR) | $ICR_{0.9}$ | $ICR_{0.99}$ |
|---|---|---|---|---|---|---|---|
| 6 | $x_3$ | **0.98** | **0.37** | **0.38** | **0.53** | **0.79** | **1.40** |
| | $x_{11}$ | 0.00 | 15.79 | 19.27 | 25.03 | 31.33 | 37.31 |
| 9 | $x_3$ | **0.99** | **0.75** | **0.75** | **0.82** | **0.94** | **1.20** |
| | $x_{11}$ | 0.00 | 8.44 | 9.95 | 12.56 | 15.41 | 18.36 |
| | $x_{21}$ | **0.99** | **0.75** | **0.75** | **0.81** | **0.92** | **1.18** |
| | $x_{29}$ | 0.00 | 2.15 | 2.88 | 4.24 | 5.73 | 7.38 |
| | $x_{37}$ | 0.00 | 3.87 | 5.46 | 8.82 | 12.61 | 17.25 |

0.99, and the MC mean PI is indicated as $E_N$(PI), where for simplicity the subscript $R$ is omitted from $N_R$.

The reference threshold is $PI = 1.2$, and $F_N(1.2)$ shows the frequency of PI being below this limit; in other words this gives an estimate for the DGP to be classified as nonparametric. There is a very clear distinction: DGPs 6 and 9 are regarded as nonparametric 98 and 100% of the time respectively. In contrast, all other DGPs are always regarded as parametric, with the slight exception of DGP 3, which is a little less clear cut.

Examining the quantiles, DGP 3 has a mean PI value of 2.54 and $PI_{0.1} = 1.53$, which puts it in the parametric class in the large majority of cases. DGP 6 has a mean PI of 0.53, and $PI_{0.9} = 0.79$, making it almost always nonparametric. DGP 9 has $PI_{0.99} = 0.93$, making it the most obviously nonparametric DGP. Of the

remaining DGPs, all are well above the threshold and can be considered parametric.

Next it was further investigated which regressors were causing the nonparametricness, i.e. which regressors are 'weak', examining the individual IC ratios for each regressor of a given DGP, see (17). Let *ICR*(*i*) indicate the IC ratio between the DGP and the submodel of the DGP where variable *i* is removed. Table 11 reports the distribution of *ICR*(*i*) for DGPs 6 and 9, which are the nonparametric DGPs. One can clearly see that in DGP 6, it is $x_3$ that is causing the nonparametricness, since it has a mean ICR(3) of 0.53. Removing this regressor improves the information criterion given the data. The same is true for $x_3$ and $x_{21}$ in DGP 9, which both have ICRs with a mean of around 0.8. In contrast, removing any of the other regressors has a significant impact on the quality of the model fit. In practice, therefore, one could consider these as the weak regressors.

Therefore, in DGPs 6 and 9, the variables in boldface inTable 11 are excluded from the EDGP. The EDGP are defined as the remaining regressors in each case, see the last column in Table 10.

# References

Abadir, K., and J. R. Magnus. 2002. "Notation in Econometrics: A Proposal for a Standard." *The Econometrics Journal* 5: 76–90.

Anderson, T. W. 1971. *The Statistical Analysis of Time Series*. New York: Wiley.

Archer, G., A. Saltelli, and I. Sobol. 1997. "Sensitivity Measures, Anova-like Techniques and the Use of Bootstrap." *Journal of Statistical Computation and Simulation* 58 (2): 99–120.

Becker, W., M. Saisana, P. Paruolo, and I. Vandecasteele. 2017. "Weights and Importance in Composite Indicators: Closing the Gap." *Ecological Indicators* 80: 12–22.

Becker, W., and A. Saltelli. 2015. "Design for Sensitivity Analysis." In *Handbook of Design and Analysis of Experiments, Chapter 18*, edited by A. Dean, M. Morris, J. Stufken, and D. Bingham, 631–78. Boca Raton, New York: CRC Press.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B* 57 (1): 289–300.

Billingsley, P. 1995. *Probability and Measure*. NY: John Wiley & Sons.

Bittman, R., J. P. Romano, C. Vallarino, and M. Wolf. 2009. "Testing Multiple Hypotheses with Common Effect." *Biometrika* 96: 399–410.

Bonferroni, C. E. 1936. *Teoria statistica delle classi e calcolo delle probabilita*. Florence, Italy: Libreria internazionale Seeber.

Brell, G., G. Li, and H. Rabitz. 2010. "An Efficient Algorithm to Accelerate the Discovery of Complex Material Formulations." *Journal of Chemical Physics* 132 (17): 174103-1–10.

Brunea, F. 2008. "Consistent Selection via the Lasso for High Dimensional Approcimating Regression Models." In *Pushing the Limits of Contemporary Statistics: Essays in Honor of J. K. Gosh*, edited by B. Clarke, and S. Ghosal, 122–37. Dordrecht: IMS.

Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference – A Practical Information-Theoretic Approach*, 2nd ed. New York: Springer.

Castle, J. L., J. A. Doornik, and D. F. Hendry. 2011. "Evaluating Automatic Model Selection." *Journal of Time Series Econometrics* 3: 1–33.

Claeskens, G., and N. L. Hjort. 2003. "The Focused Information Criterion, with Discussion." *Journal of the American Statistical Association* 98: 900–845.

Danilov, D., and J. Magnus. 2004. "On the Harm that Ignoring Pretesting Can Cause." *Journal of Econometrics* 122: 27–46.

Deckers, T., and C. Hanck. 2014. "Variable Selection in Cross-Section Regressions: Comparisons and Extensions." *Oxford Bulletin of Economics and Statistics* 76 (6): 841–73.

Doornik, J. A. 2009. "Autometrics." In *The Methodology and Practice of Econometrics: Festschrift in Honour of David F. Hendry*. Oxford, UK: Oxford University Press.

Fernandez, C., E. Ley, and M. F. Steel. 2001. "Model Uncertainty in Cross-Country Growth Regressions." *Journal of Applied Econometrics* 16 (5): 563–76.

Foster, D. P., and E. I. George. 1994. "The Risk Inflation Criterion for Multiple Regression." *Annals of Statistics* 22: 1947–75.

Freedman, D. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37: 152–5.

Freedman, D., and P. Humphreys. 1999. "Are There Algorithms that Discover Causal Structure?" *Synthese* 121: 29–54.

Freedman, L. E., and D. Pee. 1989. "Return to a Note on Screening Regression Equations." *The American Statistician* 43: 279–82.

Freedman, L. E., D. Pee, and N. Midthune. 1992. "The Problem of Underestimating the Residual Error Variance in Forward Stepwise Regression." *Journal of the Royal Statistical Society, Series D (The Statistician)* 41: 405–12.

Hansen, B. 1999. "Discussion of 'Data Mining Reconsidered' by K.D. Hoover and S.J. Perez." *Econometrics Journal* 2: 192–201.

Harrell, F. 2001. *Regression Modeling Strategies*. New York: Springer.

Hendry, D., and H.-M. Krolzig. 1999. "Improving on 'Data Mining Reconsidered' by K.D. Hoover and S.J. Perez." *Econometrics Journal* 2: 41–58.

Hendry, D. F., and A. Krolzig. 2005. "The Properties of Automatic Gets Modelling." *Economic Journal* 115: C32–61.

Hjort, N. L., and G. Claeskens. 2003. "Frequentist Model Average Estimators." *Journal of the American Statistical Association* 98: 879–99.

Homma, T., and A. Saltelli. 1996. "Importance Measures in Global Sensitivity Analysis of Nonlinear Models." *Reliability Engineering & System Safety* 52 (1): 1–17.

Hoover, K., and S. Perez. 1999. "Data Mining Reconsidered: Encompassing and the General-to-specific Approach to Specification Search." *The Econometrics Journal* 2 (2): 167–91.

Imbens, G., and J. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47: 5–86.

Jansen, M. J. W. 1999. "Analysis of Variance Designs for Model Output." *Computer Physics Communications* 117 (1–2): 35–43.

Johansen, S. 1996. *Likelihood-based Inference in Cointegrated Vector Auto-Regressive Models*. Oxford, UK: Oxford University Press.

Krolzig, H.-M., and D. F. Hendry. 2001. "Computer Automation of General-to-specific Model Selection Procedures." *Journal of Economic Dynamics and Control* 25 (6): 831–66.

Leamer, E. E. 1983. "Let's take the con out of econometrics." *The American Economic Review* 73 (1): 31–43.

Leeb, H., and B. M. Poetscher. 2006. "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *Annals of Statistics* 34: 2554–91.

Ley, E., and M. F. Steel. 2009. "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Journal of Applied Econometrics* 24 (4): 651–74.

Liu, W., and Y. Yang. 2011. "Parametric or Nonparametric? A Parametricness Index for Model Selection." *The Annals of Statistics* 39 (4): 2074–102.

Lovell, M. C. 1983. "Data Mining." *The Review of Economics and Statistics* 65: 1–12.

Magnus, J. 2007. "Local Sensitivity in Econometrics." In *Measurement in Economics*, edited by M. Boumans, 295–319. San Diego: Academic Press.

Magnus, J., and J. Durbin. 1999. "Estimation of Regression Coefficients of Interest when Other Regression Coefficients are of No Interest." *Econometrica* 67: 639–43.

Magnus, J., and A. Vasnev. 2007. "Local Sensitivity and Diagnostic Tests." *The Econometrics Journal* 10: 166–92.

Magnus, J., O. Powell, and P. Prufer. 2010. "A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics." *Journal of Econometrics* 154: 139–53.

Miller, A. 2002. *Subset Selection in Regression*, 2nd ed. Boca Raton, USA: Chapman and Hall, CRC Press.

Norton, J. 2015. "An Introduction to Sensitivity Assessment of Simulation Models." *Environmental Modelling & Software* 69: 166–74.

Pantula, S. G. 1989. "Testing for Unit Roots in Time Series Data." *Econometric Theory* 5 (2): 256–71.

Paruolo, P. 2001. "The Power of Lambda Max." *Oxford Bulletin of Economics and Statistics* 63: 395–403.

Paruolo, P., and A. Rahbek. 1999. "Weak Exogeneity in I(2) VAR Systems." *Journal of Econometrics* 93: 281–308.

Paruolo, P., A. Saltelli, and M. Saisana. 2013. "Ratings and Rankings: Voodoo or Science?" *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 176: 609–34.

Paulsen, J. 1984. "Order Determination of Multivariate Autoregressive Time Series with Unit Roots." *Journal of Time Series Analysis* 5: 115–27.

Pearson, K. 1905. *On the General Theory of Skew Correlation and Non-linear Regression, Volume XIV of Mathematical Contributions to the Theory of Evolution, Drapers' Company Research Memoirs*. London: Dulau & Co. Reprinted in: Early Statistical Papers, Cambridge University Press, Cambridge, UK, 1948.

Phillips, P. C. B. 1997. "Econometric Model Determination." *Econometrica* 64: 763–812.

Phillips, P. C. B. 2003. "Laws and Limits of Econometrics." *Economic Journal* 113: C26–52.

Poetscher, B. M. 1991. "Effects of Model Selection on Inference." *Econometric Theory* 7: 163–85.

Pretis, F., 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, Articles* 86 (3): 1–44.

Romano, J. P., and A. M. Shaikh. 2006. "On Stepdown Control of the False Discovery Proportion." In *Optimality*, 33–50. Institute of Mathematical Statistics.

Romano, J. P., A. M. Shaikh, and M. Wolf. 2008. "Control of the False Discovery Rate under Dependence Using the Bootstrap and Subsampling." *Test* 17 (3): 417–42.

Romano, J. P., and M. Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73: 1237–82.

Sala-i-Martin, X. 1997. "I Just Ran Two Million Regressions." *The American Economic Review, Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association*, Vol. 87, 178–83.

Saltelli, A., and P. Annoni. 2010. "How to Avoid a Perfunctory Sensitivity Analysis." *Environmental Modelling & Software* 25 (12): 1508–17.

Saltelli, A., and S. Tarantola. 2002. "On the Relative Importance of Input Factors in Mathematical Models." *Journal of the American Statistical Association* 97 (459): 702–9.

Saltelli, A., M. Ratto, S. Tarantola, and F. Campolongo. 2012. "Sensitivity Analysis for Chemical Models." *Chemical Reviews* 112: PR1–PR21.

Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. 2010. "Variance Based Sensitivity Analysis of Model Output. Design and Estimator for the Total Sensitivity Index." *Computer Physics Communications* 181 (2): 259–70.

Saltelli, A., S. Tarantola, and F. Campolongo. 2000. "Sensitivity Analysis as an Ingredient of Modelling." *Statistical Science* 15 (4): 377–95.

Saltelli, A., T. Andres, and T. Homma. 1993. "Sensitivity Analysis of Model Output: An Investigation of New Techniques." *Computational Statistics & Data Analysis* 15: 211–38.

Sobol', I. M. 1967. "On the Distribution of Points in a Cube and the Approximate Evaluation of Integrals." *USSR Computational Mathematics and Mathematical Physics* 7 (4): 86–112.

Sobol', I. M. 1993. "Sensitivity Estimates for Nonlinear Mathematical Models." *Mathematical Modeling and Computational Experiment* 1 (4): 407–14.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B* 58 (1): 267–88.

Wei, P., Z. Lu, and J. Song. 2015. "Variable Importance Analysis: A Comprehensive Review." *Reliability Engineering & System Safety* 142: 399–432.