# Exploring Econometric Model Selection Using Sensitivity Analysis

William Becker
Paolo Paruolo
Andrea Saltelli

Nice, 2nd July 2013

# Outline

- What is the problem we are addressing?
- Past approaches – Hoover and Perez
- Our method – overview and preliminary study
- Sensitivity analysis
- Adaption to model selection
- Algorithm and extensions
- Results
- Conclusions

# A short abstract

- A foray into the problem of model selection or “data mining” in econometrics, i.e. finding which variables are driving a dependent variable of interest
- We use the “total effect index” *on variable triggers* as a tool for determining the importance of variables in a regression
- We compare it to that of a well-known data mining paper on a series of test problems using Monte Carlo simulations.
- The results indicate that our algorithm identifies correct variables more often and is more robust (on the test cases investigated).

# Econometrics

*For those who don't know (like me last year…)*

- Looking for, and quantifying, relationships between economic variables.
- Then can be used for forecasting/policy evaluation/etc.
- Relationships are usually assumed to be linear (in the parameters)
- Examples of econometric variables (which are a function of time):
  - GDP
  - Unemployment rate
  - Disposable income
  - Unemployment rate
  - Bank reserves
  - etc….

# Model selection

- Problem: economic variables are a tangled web of interactions, correlations, and lags.
- Typically an econometrician has to identify which variables are driving a dependent variable of interest, from a large pool of candidate variables. This is the model selection problem, or "data mining".
- E.g. if the dependent variable of interest is GDP, there may be hundreds of candidate variables (stock prices, price indices, employment rates) to consider.
- Occam's Razor says that we cannot simply include *all* candidate variables in a regression model because this will lead to over-fitting and a poor predictive model. We also would like to know which variables are driving the output for policy information.

# Hoover and Perez Study

- A paper was published in 1999 by Hoover and Perez investigating how well the contemporary approach to data mining performed. This is the basis for our work.
- They assess how well the "LSE" data mining approach works by examining artificially-generated data based on real macroeconomic measurements.

*Econometrics Journal* (1999), volume **2**, pp. 167–191.

**Data mining reconsidered: encompassing and the general-to-specific approach to specification search**

KEVIN D. HOOVER, STEPHEN J. PEREZ

*Department of Economics, University of California, Davis, California 95616-8578, USA*
E-mail: kdhoover@ucdavis.edu; Homepage: www.ucdavis.edu/~kdhoover/
*Department of Economics, Washington State University, Pullman, Washington 99164-4741, USA*
E-mail: sjperez@wsu.edu; Homepage: www.cbe.wsu.edu/~sjperez/

**Test data:** 18 macroeconomic variables, recorded quarterly from 1959-1995, with lags. Artificial dependent variable (PCE) is created as a function of some of these variables. All variables lagged including PCE.

| Variable | Variable number | | | | | Times differenced for stationarity[a] | CITIBASE identifier[b] |
|---|---|---|---|---|---|---|---|
| | Current | Lag | | | | | |
| | | 1 | 2 | 3 | 4 | | |
| Index of four coincident indicators | 1 | 19 | | | | 1 | DCOINC |
| GNP price deflator | 2 | 20 | | | | 2 | GD |
| Government purchases of goods and services | 3 | 21 | | | | 2 | GGEQ |
| Federal purchases of goods and services | 4 | 22 | | | | 1 | GGFEQ |
| Federal government receipts | 5 | 23 | | | | 2 | GGFR |
| GNP | 6 | 24 | | | | 1 | GNPQ |
| Disposable personal income | 7 | 25 | | | | 1 | GYDQ |
| Gross private domestic investment | 8 | 26 | | | | 1 | GPIQ |
| Total member bank reserves | 9 | 27 | | | | 2 | FMRRA |
| Monetary base (federal reserve bank of St. Louis) | 10 | 28 | | | | 2 | FMBASE |
| M1 | 11 | 29 | | | | 1 | FM1DQ |
| M2 | 12 | 30 | | | | 1 | FM2DQ |
| Dow Jones stock price | 13 | 31 | | | | 1 | FSDJ |
| Moody's AAA corporate bond yield | 14 | 32 | | | | 1 | FYAAAC |
| Labor force (16 years+, civilian) | 15 | 33 | | | | 1 | LHC |
| Unemployment rate | 16 | 34 | | | | 1 | LHUR |
| Unfilled orders (manufacturing, all industries) | 17 | 35 | | | | 1 | MU |
| New orders (manufacturing, all industries) | 18 | 36 | | | | 2 | MO |
| Personal consumption expenditure[c] | N/A | 37 | 38 | 39 | 40 | 1 | GCQ |

Note: Data run 1959.1–1995.1. All data from CITIBASE: Citibank economic database (Floppy disk version), July 1995 release. All data converted to quarterly by averaging or summing as appropriate. All dollar denominated data in billions of constant 1987 dollars. Series FMRRA, FMBASE, GGFR, FSDJ, MU, and MO are deflated using the GNP price deflator (Series GD). [a] Indicates the number of times the series had to be differenced before a Phillips–Perron test could reject the null hypothesis of non-stationarity at a 5% significance level (Phillips and Perron 1988). [b] Indicates the identifier code for this series in the CITIBASE economic database. [c] For calibrating models in Table 4 actual personal consumption expenditure data is used as the dependent variables; for specification searches, actual data is replaced by artificial data generating according to models in Table 3. Variable numbers refer to these artificial data, which vary from context to context.

11 data-generating processes (DGPs) are created from the data.
Each has added noise term.

| | | |
|---|---|---|
| Random errors | | |
| $u_t \sim N(0, 1)$ | | |
| $u_t^* = 0.75u_{t-1}^* + u_t\sqrt{7/4}$ | | |
| Models | | |
| Model 1: | $y1_t = 130.0u_t$ | |
| Model 2: | $y2_t = 130.0u_t^*$ | |
| Model 2′: | $y2_t = 0.75y2_{t-1} + 85.99u_t$ | |
| Model 3: | $\ln(y3)_t = 0.395\ln(y3)_{t-1} + 0.3995\ln(y3)_{t-2} + 0.00172u_t$ | s.e.r. = 0.00172, $R^2$ =0.99 |
| Model 4: | $y4_t = 1.33x11_t + 9.73u_t$ | s.e.r. = 9.73, $R^2$ = 0.58 |
| Model 5: | $y5_t = -0.046x3_t + 0.11u_t$ | s.e.r. = 0.11, $R^2$ = 0.93 |
| Model 6: | $y6_t = 0.67x11_t – 0.023x3t + 4.92u_t$ | s.e.r. = 4.92, $R^2$ = 0.58 |
| Model 6A: | $y6_t = 0.67x11_t – 0.32x3t + 4.92u_t$ | s.e.r. = 4.92, $R^2$ = 0.64 |
| Model 6B: | $y6_t = 0.67x11_t – 0.65x3t + 4.92u_t$ | s.e.r. = 4.92, $R^2$ = 0.74 |
| Model 7: | $y7_t = 1.33x11_t + 9.73u_t^*$ | s.e.r. = 9.73, $R^2$ = 0.58 |
| Model 7′: | $y7_t = 0.75y7_{t-1} + 1.33x11_t – 0.9975x29_t + 6.73u_t$ | |
| Model 8: | $y8_t = -0.046x3_t + 0.11u_t^*$ | s.e.r. = 0.11, $R^2$ = 0.93 |
| Model 8′: | $y8_t = 0.75y8_{t-1} - 0.046x3_t + 0.00345x21_t + 0.073u_t$ | |
| Model 9: | $y9_t = 0.67x11_t – 0.023x3_t + 4.92u_t^*$ | s.e.r. = 4.92, $R^2$ = 0.58 |
| Model 9′: | $y9_t = 0.75y9_{t-1} – 0.023x3_t + 0.01725x21_t + 0.67x11_t – 0.5025x29_t + 3.25u_t$ | |

Note: The variables $y\#_t$ are the artificial variables created by each model. The variables $x\#_t$ correspond to the variables with the same number in Table 1. The coefficients for models 3, 4, and 5 come from the regression of personal consumption expenditures (Dep. in Table 1) on independent variables as indicated by the models. The standard error of the regression for models 3, 4, and 5 is scaled to set $R^2$ equal to that for the analogous regressions run on non-stationary data to mirror Lovell. Model 6 is the average of models 4 and 5. Models 7, 8, and 9 have same coefficients as models 4, 5, and 6 with autoregressive errors. Models 2′, 7′, 8′, and 9′ are exactly equivalent expressions for models 2, 7, 8, 9 in which lags of the variables are used to eliminate the autoregressive parameter in the error process.

# Variables are highly correlated in some instances.

**Table 2.** Correlation matrix for search variables.

| Variable name and number | Variable number 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | Dep.* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Four coincident indicators | **0.67** | | | | | | | | | | | | | | | | | | |
| *2. GNP price deflator* | *0.21* | ***0.24*** | | | | | | | | | | | | | | | | | |
| 3. Government purchases of goods and services | 0.04 | −0.09 | **8.81** | | | | | | | | | | | | | | | | |
| *4. Federal purchases of goods and services* | *−0.07* | *−0.08* | *0.54* | ***6.22*** | | | | | | | | | | | | | | | |
| 5. Federal government receipts | 0.21 | 0.28 | 0.03 | 0.01 | **22.16** | | | | | | | | | | | | | | |
| *6. GNP* | *0.83* | *0.16* | *0.13* | *0.03* | *0.20* | ***30.71*** | | | | | | | | | | | | | |
| 7. Disposable personal income | 0.57 | 0.07 | 0.07 | −0.09 | 0.06 | 0.49 | **25.09** | | | | | | | | | | | | |
| *8. Gross private domestic investment* | *0.76* | *0.19* | *0.03* | *−0.18* | *0.13* | *0.83* | *0.40* | ***25.91*** | | | | | | | | | | | |
| 9. Total member bank reserves | −0.02 | 0.24 | 0.07 | 0.14 | 0.40 | −0.03 | 0.24 | −0.16 | **514.26** | | | | | | | | | | |
| *10. Monetary base (federal reserve bank of St. Louis)* | *−0.02* | *0.49* | *−0.02* | *0.07* | *0.25* | *−0.06* | *0.10* | *−0.06* | *0.54* | ***1.38*** | | | | | | | | | |
| 11. M1 | 0.24 | −0.04 | −0.04 | 0.00 | 0.16 | 0.27 | 0.17 | 0.17 | 0.25 | 0.20 | **8.49** | | | | | | | | |
| *12. M2* | *0.20* | *−0.06* | *−0.08* | *0.07* | *0.11* | *0.20* | *0.17* | *0.08* | *0.21* | *0.14* | *0.60* | ***25.08*** | | | | | | | |
| 13. Dow Jones stock price | −0.04 | −0.06 | −0.06 | −0.06 | −0.12 | 0.03 | −0.03 | −0.02 | −0.08 | 0.01 | 0.27 | 0.04 | **95.40** | | | | | | |
| *14. Moody's AAA corporate bond yield* | *0.23* | *0.11* | *−0.04* | *−0.05* | *0.07* | *0.11* | *0.07* | *0.20* | *−0.16* | *−0.06* | *−0.33* | *−0.33* | *−0.26* | ***0.42*** | | | | | |
| 15. Labor force (16 years+, civilian) | 0.17 | 0.04 | 0.03 | −0.04 | −0.03 | 0.11 | 0.09 | 0.07 | −0.17 | 0.01 | −0.04 | −0.07 | 0.13 | 0.11 | **321.15** | | | | |
| 16. *Unemployment rate* | *−0.85* | *−0.13* | *−0.01* | *−0.02* | *−0.09* | *−0.73* | *−0.31* | *−0.66* | *0.08* | *0.07* | *−0.23* | *−0.22* | *0.02* | *−0.22* | *0.02* | ***0.35*** | | | |
| 17. Unfilled orders (manufacturing, all industries) | 0.21 | 0.24 | −0.08 | 0.04 | 0.03 | 0.16 | 0.05 | 0.10 | −0.10 | 0.09 | −0.39 | −0.21 | 0.06 | 0.27 | 0.14 | −0.23 | **6248.9** | | |
| *18. New orders (manufacturing, all industries)* | *0.23* | *0.12* | *−0.29* | *−0.15* | *0.25* | *0.22* | *0.15* | *0.10* | *0.21* | *0.01* | *0.28* | *0.19* | *0.06* | *0.12* | *0.01* | *−0.12* | *−0.04* | ***4114.8*** | |
| *Dep. personal consumption expenditure | 0.60 | −0.02 | −0.02 | −0.02 | 0.15 | 0.65 | 0.40 | 0.30 | 0.07 | −0.03 | 0.47 | 0.41 | 0.18 | −0.05 | 0.13 | −0.50 | −0.01 | 0.39 | *15.85* |

Note: Variables are differenced as indicated in Table 1. Elements in bold type on the main diagonals are the standard deviations of each variable for the period beginning 1959.2 or 1959.3, depending on the number of differences. Off-diagonal elements correlations are calculated for the variables in Table 1 for the period 1959.3 to 1995.1. *Dep. indicates that personal consumption expenditure is the dependent variable used in calibrating the models in Table 3. It is not a search variable. The dependent variables and its lags used in the simulations below are constructed according to those models.

**Some notation**

Let $\mathbf{x}=(x_1, x_2, \ldots, x_D)^T$ be the vector of D candidate variables, and $y$ be the dependent variable.

- We suspect that $y$ is a function of some subset of **x**. i.e. perhaps it is only a function of $x_2$. Or $x_3$ and $x_7$.

We have access to quarterly measurements of **x** and $y$ over a number of years, i.e. we have a data set consisting of $n$ measurements of each variable:

$$\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

$$\mathbf{y}=(y_1, y_2, \ldots, y_n)^T$$

We want to use this data to select the variables that are driving $y$.

We therefore consider models, indexed by k, of the form,

$$\hat{y} = \hat{\beta}^{\mathsf{T}}_{\mathbf{z}_k} \mathbf{x}_{\mathbf{z}_k}$$

i.e. linear regressions on subsets of the **x**.

$\mathbf{z}_k$ is a binary vector which denotes which variables to include in model $k$, e.g.

$$\mathbf{z}_k = (0,1,1,0,1,0,0,1,\ldots) \Rightarrow \mathbf{x}_{\mathbf{z}_k} = (x_2, x_3, x_5, x_8, \ldots)$$

Note that there will be $2^D$ possible models (subsets of **x**)

The aim is to choose the best one.

## Hoover and Perez (HP) method (based on LSE practice)

(this is a conceptual outline only, not the full algorithm)

- Fit a linear regression with all 40 candidate regressors
- Use a batch of diagnostic tests for each regression to check suitability for linear regression. Discard data upon failure.
- Calculate t-statistics
- Remove variables with insignificant t-statistics
- Remove regressors one by one starting with the lowest significant t-statistic
- After removing each regressor, perform an F-test comparing the restricted model to the general one (with all 40 regressors)
- The algorithm is allowed to skip variables in the ordering and to try a number of search paths.

The main tools are therefore the t-statistics for ordering regressors, and the F-test for comparing regression models.

## Their Results

Using 1000 data sets from each DGP, Monte Carlo expected values.

**Table 7.** Specification search at 1% nominal size.[a]

| | True model[b] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1[g] | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Means |
| Percentage of searches for which the true and final specifications are related in categories:[c] | | | | | | | | | | |
| 1. True = Final | 79.9 | 0.8 | 70.2 | 80.2 | 79.7 | 0.7 | 24.6 | 78.0 | 0.8 | 46.1 |
| 2. True $\subset$ Final, $SER_F < SER_T$ | 20.1 | 99.2 | 19.0 | 19.6 | 20.2 | 0.1 | 57.4 | 21.7 | 1.3 | 28.7 |
| 3. True $\subset$ Final, $SER_F > SER_T$ | 0.0 | 0.0 | 0.2 | 0.1 | 0.1 | 0.0 | 0.0 | 0.2 | 0.6 | 0.1 |
| 4. True $\not\subset$ Final, $SER_F < SER_T$ | 0.0 | 0.0 | 3.7 | 0.1 | 0.0 | 56.3 | 13.0 | 0.1 | 77.0 | 16.7 |
| 5. True $\not\subset$ Final, $SER_F > SER_T$ | 0.0 | 0.0 | 6.9 | 0.0 | 0.0 | 42.9 | 5.0 | 0.0 | 20.3 | 8.3 |
| True variable number[d] | Null set | 37 | 37/38 | 11 | 3 | 3/11 | 11/29/37 | 3/21/37 | 3/11/21/29/37 | |
| Frequency variables included (percent) | | 100.0 | 95.7/93.6 | 99.9 | 100.0 | 0.8/99.8 | 100.0/82.0/ 100.0 | 100.0/99.9/ 99.9 | 1.5/100.0/ 1.4/83.5/99.9 | |

- DGP is correctly identified around half the time
- Tends to include unwanted variables quite often, also discards true variables.
- At other significance levels (5% and 10%), gave worse results.

## Ranking Regressors

- The HP approach ranks variables according to t-statistics, then uses this a basis for model selection.
- Ranking should ideally have true regressors with the highest t-statistics.

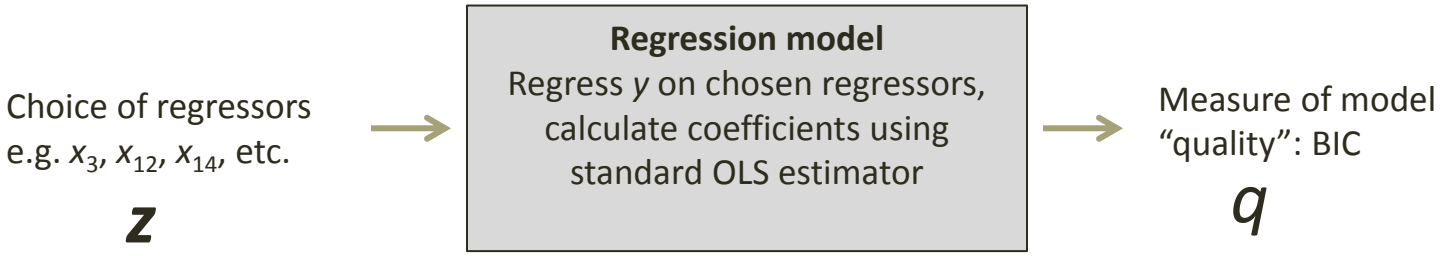
Our proposed way of ranking regressors is by using an application of the *total effect index,* $S_T$

# Adaption to Model Selection

## *How to frame the problem?*

Going back to the model selection problem, we have a fixed data set of $n$ data points (in our test cases $n$=139) in 40 variables.

We want to know which variables give the best fit to the data. We cannot change the *values* of a given variable, we only decide whether to include it in our regression model or not. This gives a system:




i.e. $q_k = BIC(\boldsymbol{z}_k)$, where $\boldsymbol{z}_k$ is a $n$-length binary vector with $z_{ki} = 0$ denotes the exclusion of $z_{ki}$, and $z_{ki} = 1$ includes it. E.g.

$$\boldsymbol{z}_k = (0,1,1,0,1,0,0,1,\ldots), \text{ then } \boldsymbol{x}_k = (x_2, x_3, x_5, x_6,\ldots)$$

We perform a sensitivity analysis on the system $q = BIC(\boldsymbol{z})$, i.e. how sensitive is model quality to the inclusion or exclusion of each regressor?

- Treat the $z_i$ as discrete independent random variables $Z_i$ such that $p(Z_i = 0) = p(Z_i = 1) = ½$.
- Use BIC (Bayesian Information Criterion) as measure of model quality
- Use $S_T$ to judge importance of including each variable in the regression model

## Adaption to Model Selection

We can now define the population of models, Γ, by the following characteristics,

$$\Gamma = \left\{(\mathbf{z}_1, q_1), (\mathbf{z}_2, q_2), \ldots, (\mathbf{z}_{2^{40}}, q_{2^{40}})\right\}$$

i.e. the set of all regressor combinations and their corresponding BIC values.

We represent this set of values as a finite (but big) population, which we sample from. Each model is assigned equal probability.

We define $\mathbb{V}$ and $\mathbb{E}$ as the variance and expectation operators in this population (just to be extra clear).

Now we use the total effect index in the following way,

$$S_{Ti} = \frac{\mathbb{E}_{\mathbf{z}_{\sim i}}(\mathbb{V}_{\mathbf{z}_i}(q|\mathbf{z}_{\sim i}))}{\mathbb{V}(q)}$$

i.e. we are taking the total effect index of *q* (BIC), with respect to each regressor.

We estimate $S_{Ti}$ by sampling from Γ, using the Monte Carlo estimator.

We expect $(\mathbb{V}_{\mathbf{z}_i}(q|\mathbf{z}_{\sim i}))$ (and hence $S_{Ti}$) to be small if $x_i$ is not included in the model, and vice versa.

**Estimating Sensitivity Indices**

We use the Monte Carlo Method.

Standard estimator:

$$S_{Ti} = \frac{E_{\mathbf{z}_{\sim i}}(V_{\mathbf{z}_i}(q|\mathbf{z}_{\sim i}))}{V(q)} = \frac{\sigma_{Ti}^2}{\sigma^2}$$

$$\hat{\sigma}_{Ti}^2 = \frac{1}{4N}\sum_{k=1}^{N}\left[q(\mathbf{z}_k) - q(\mathbf{z}_{k,i'})\right]^2$$

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left[q(\mathbf{z}_k) - E(q(\mathbf{z}))\right]^2$$

This involves choosing a random *z*, then for each regressor, turning it on if it is off, and off it is on, i.e.

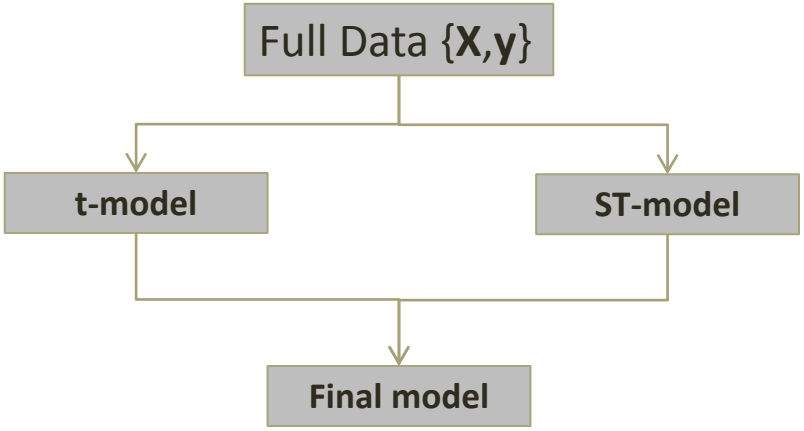
$$\begin{array}{lcllcllcl}
\mathbf{z}_k & = & (0,1,1,0,1,0,0,1,...) & \Rightarrow & \mathbf{x}_{\mathbf{z}_k} & = & (x_2,x_3,x_5,x_8,...) & \Rightarrow & q_k \\
\mathbf{z}_{k,1'} & = & (1,1,1,0,1,0,0,1,...) & \Rightarrow & \mathbf{x}_{\mathbf{z}_{k,1'}} & = & (x_1,x_2,x_3,x_5,x_8,...) & \Rightarrow & q_{k,1'} \\
\mathbf{z}_{k,2'} & = & (0,0,1,0,1,0,0,1,...) & \Rightarrow & \mathbf{x}_{\mathbf{z}_{k,2'}} & = & (x_3,x_5,x_8,...) & \Rightarrow & q_{k,2'} \\
\mathbf{z}_{k,3'} & = & (0,1,0,0,1,0,0,1,...) & \Rightarrow & \mathbf{x}_{\mathbf{z}_{k,3'}} & = & (x_2,,x_5,x_8,...) & \Rightarrow & q_{k,3'}
\end{array}$$

## Building an Algorithm

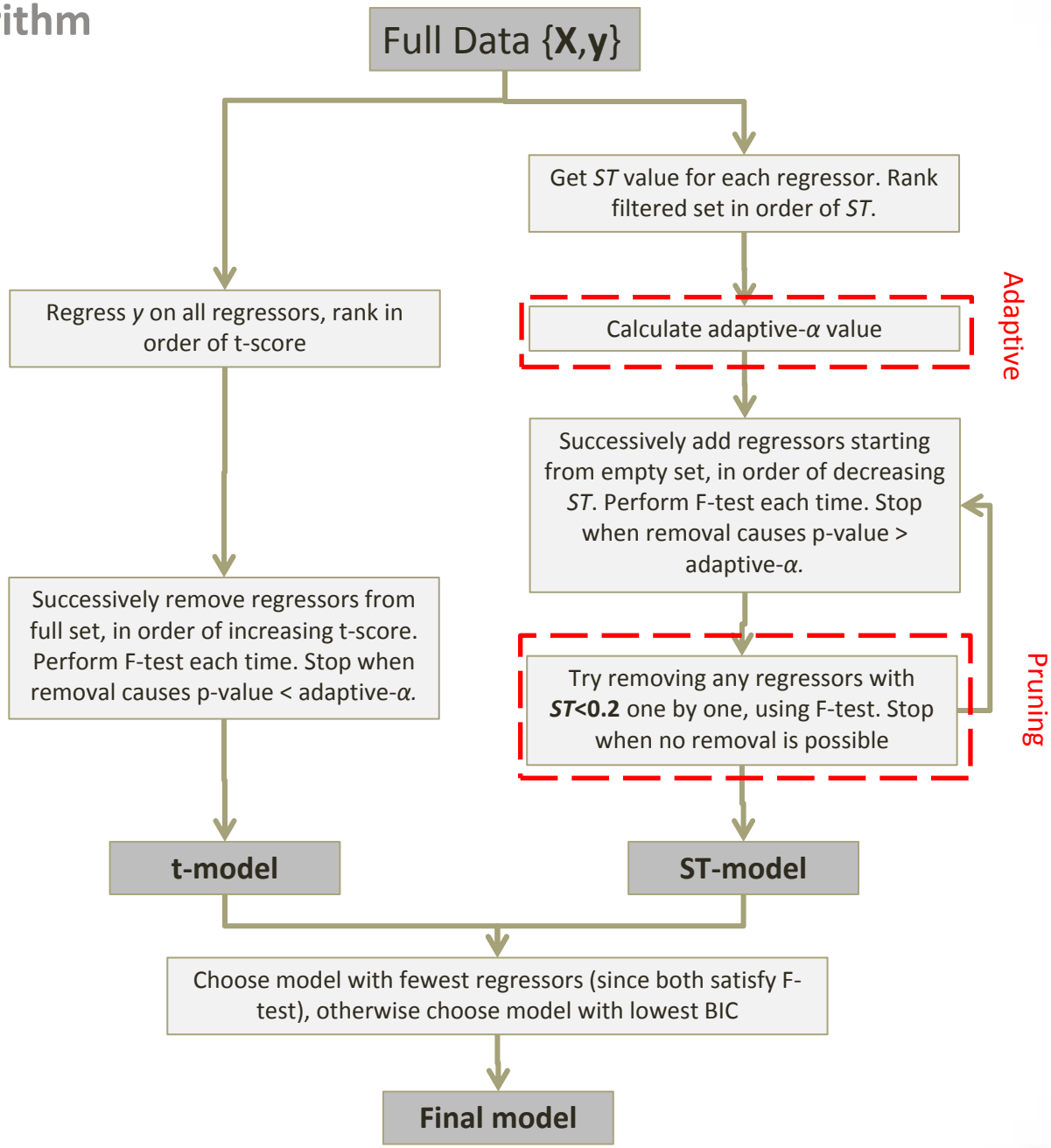
We discovered that $S_T$ cannot be used alone to select regressors – sometimes gets confused with correlations and lagged variables. But does have significant advantages in certain DGPs.

From the preliminary study it was found that $S_T$ and t-statistics can perhaps be used in a complementary fashion to rank regressors.

Our algorithm consists of finding a regression model using the t-ranking, and a regression model using the $S_T$-ranking, then a final comparison step to find the final model (variable selection).



# Full Algorithm



## General Steps – "Testing Up"

1. Rank all regressors by t-score or $S_T$ .
2. Define the initial candidate model as the empty set of regressors.
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 probability of the candidate model is below a given significance level , go to step 3 (continue adding regressors), otherwise, go to step 6.
6. Since the F-test has invalidated the addition of the last regressor, remove the last regressor added – this is the final model.

This results in two models, a "t-model" and an "$S_T$-model". The final model is chosen by,

- The model with the fewest regressors (since they are already both validated by F-test)

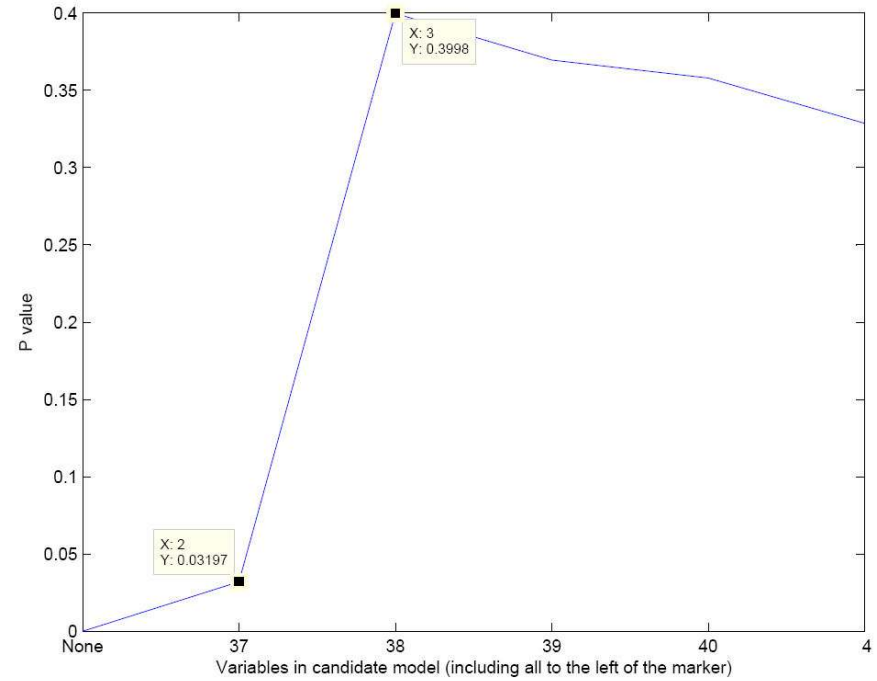
Or, in the case where the regressors are equal in number but not the same,

- The model with the lowest BIC

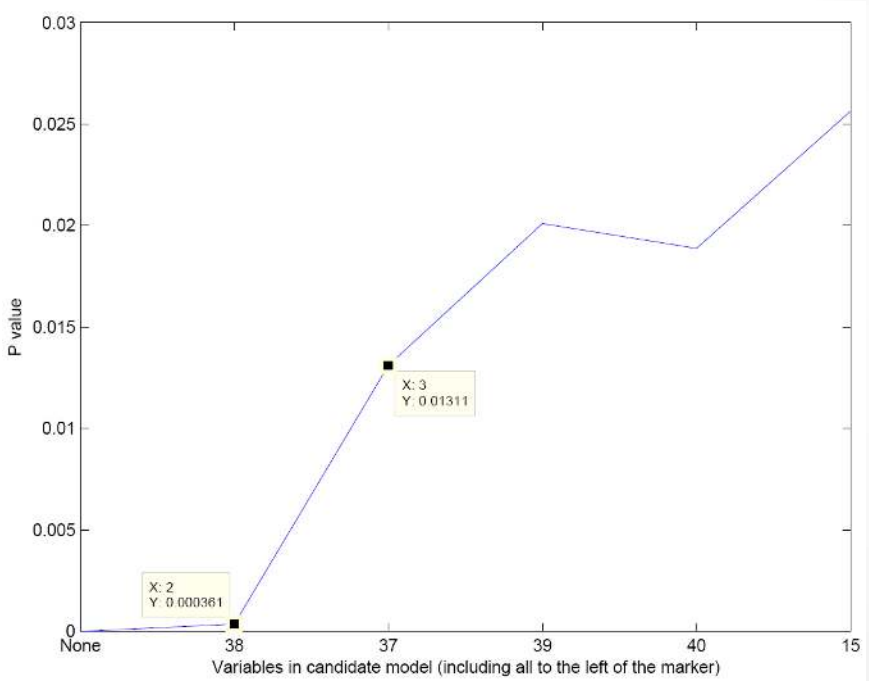
## Extension: “Adaptive alpha”

- Algorithm relies on the F-test, which requires a significance level $\alpha$ to define what is a valid reduction of the GUM.
- Problem: some DGPs require low $\alpha$ to accurately identify DGP, while others require higher value. These best $\alpha$ values will not be known in real problems.

E.g. DGP3




Requires $\alpha$ above 0.032




Requires $\alpha$ below 0.013

## Extension: “Adaptive alpha”

Solution is to allow $\alpha$ to vary with the data.

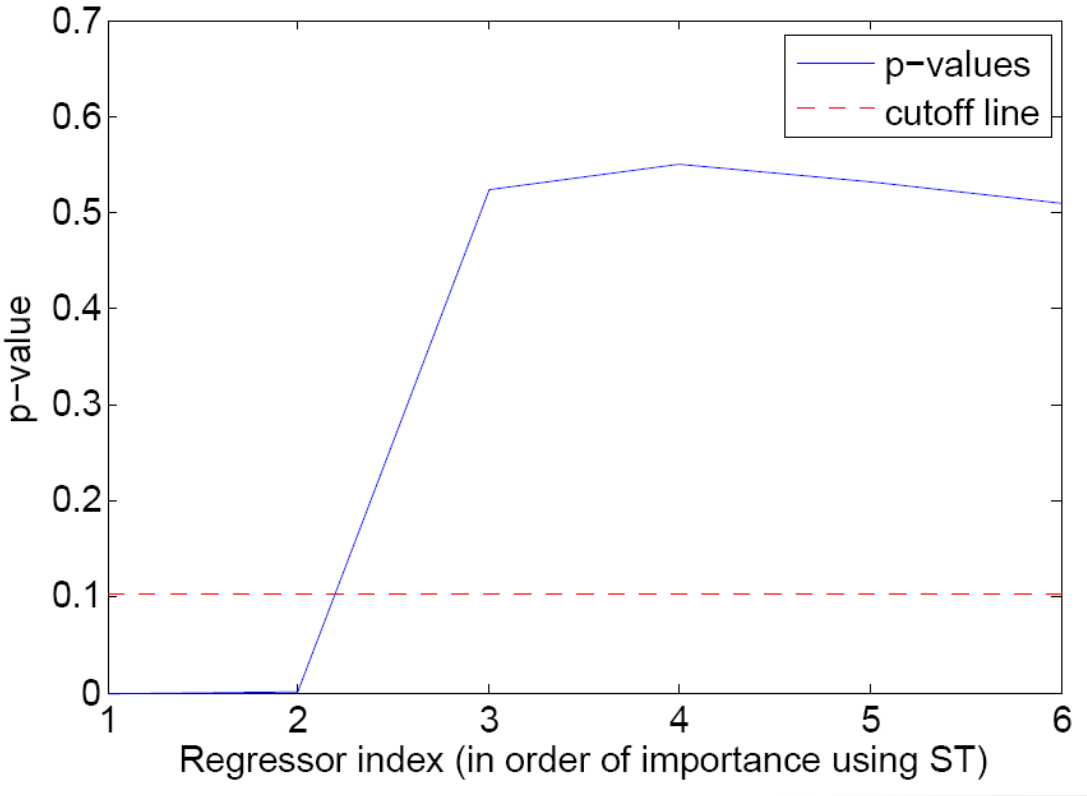
- Define $p_L$ as the value returned from the F-test by comparing the GUM against the empty set of regressors (i.e. assuming that $y$ is dependent on at least one regressor, this will usually be a “false” model)
- Define $p_H$ as the value returned from the F-test by comparing all regressors with $S_T$ >0.01 (this will very likely contain the DGP regressors – could also define in other ways).

Now define adaptive alpha, $\alpha_a$ as,

$$\alpha_a = p_{\mathrm{L}} + \phi(p_{\mathrm{H}} - p_{\mathrm{L}})$$

Will be a cutoff some fraction $\varphi$ between the high and low values.

$\varphi$ = 0.2 approx. gives good results



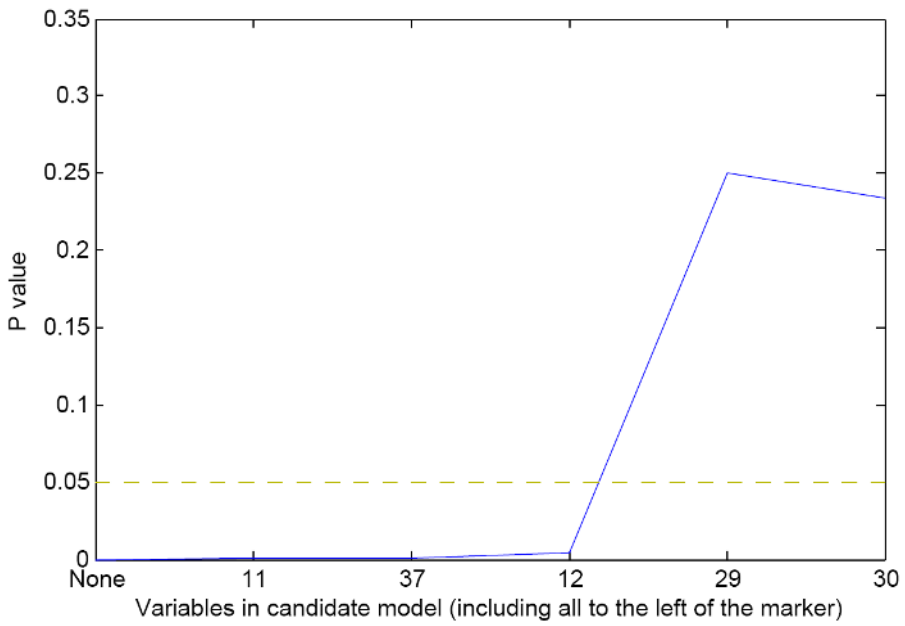
# Extension: “Pruning”

We also try a “pruning” stage in the algorithm, i.e. after regressors are chosen using the testing-up steps, try removing any of the regressors in the final model, one by one, using the F-test as a criterion of success. Allows for occasional ranking errors due to “bad data” or errors in calculation of $S_T$

E.g. DGP9

$S_T$ ordering (highest to lowest)

| $S_T$ | 0.45 | 0.32 | 0.16 | 0.07 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Regressor index** | 11 | 37 | 12 | 29 | 30 | 32 | 38 | 2 | 10 | 9 |




**Candidate models**

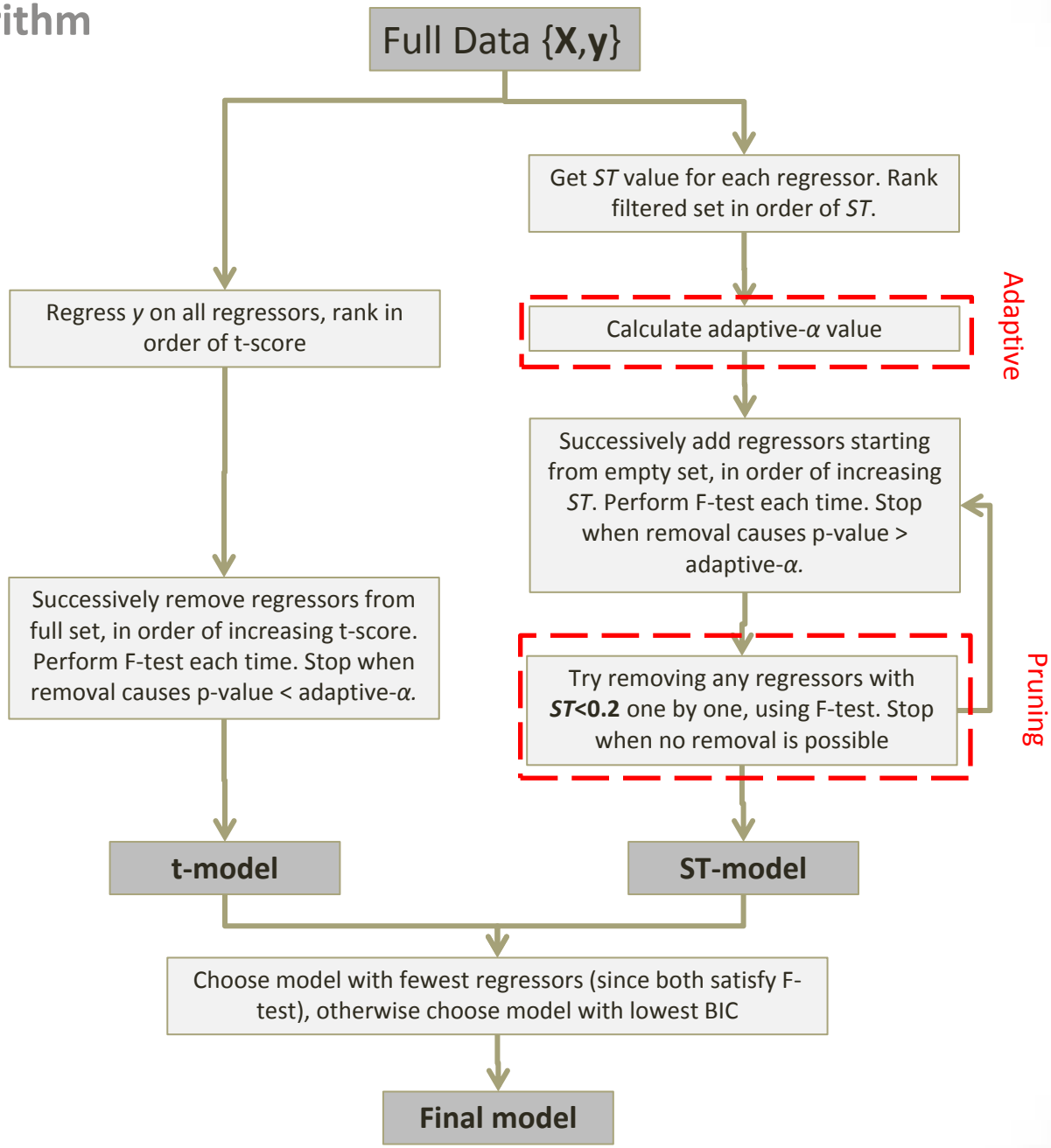
| | |
|---|---|
| $\{\}$ | $p < \alpha_a$ |
| $\{x_{11}\}$ | $p < \alpha_a$ |
| $\{x_{11}, x_{37}\}$ | $p < \alpha_a$ |
| $\{x_{11}, x_{37,}, x_{12}\}$ | $p < \alpha_a$ |
| $\{x_{11}, x_{37,}, x_{12}, x_{29}\}$ | $p > \alpha_a$ |

… would stop here. Pruning tries removing regressors from the accepted model. We find that we can remove $x_{12}$

$\{x_{11}, x_{37,}, x_{29}\}$ $p = 0.1002 > \alpha_a$

This is the true DGP.

# Full Algorithm



## Measuring Performance

We take *R* = 500 datasets from each DGP (to average over noise) and caculate the following measures

1. $C_1 = \frac{1}{R}\sum_{r=1}^{R} 1(\mathbf{x}_{\mathbf{z}_0} = \mathbf{x}_{\mathbf{z}_f})$ - Exact ID (best)
2. $C_2 = \frac{1}{R}\sum_{r=1}^{R} 1(\mathbf{x}_{\mathbf{z}_0} \subset \mathbf{x}_{\mathbf{z}_f})$ - DGP nested in final model
3. $C_3 = \frac{1}{R}\sum_{r=1}^{R} 1(\mathbf{x}_{\mathbf{z}_0} \not\subseteq \mathbf{x}_{\mathbf{z}_f})$ - DGP not nested in final model (worst)

Additionally, we use measures of "potency" and "gauge"

**Potency:** A measure of the frequency of inclusion of correct variables (ideal = 1)

**Gauge:** A measure of the frequency of inclusion of incorrect variables (ideal = 0)

Note that these measure will vary depending on the values of the tuning parameters in the HP algorithm and the one presented here (particularly $\alpha$ and $\varphi$).

We measure performance at optimised parameter values and (qualitatively) when parameter values are unknown.

# Results – Optimum Performance

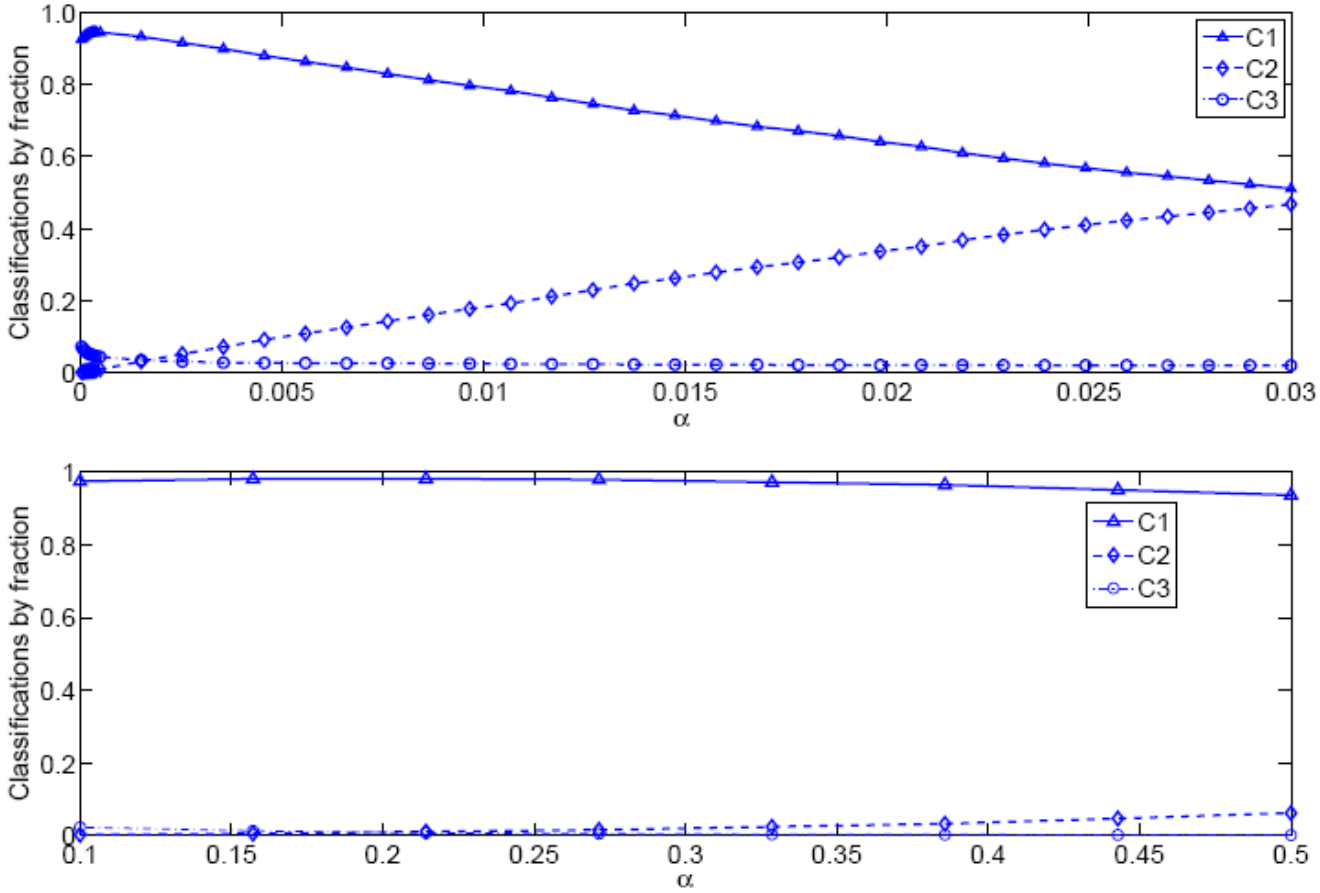
We measure performance ($C_1$), averaged over all DGPs, at a range of values of $\alpha$ and $\varphi$ and use the optimum values, i.e. the best possible performance.

| DGP | ST (no skip, no adapt) $\alpha = 0.0371$ | | | ST (no skip) $\varphi = 0.3$ | | | ST (full) $\varphi = 0.3$ | | | HP $\alpha = 4 \cdot 10^{-4}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | Gauge | Pot. | C1 | Gauge | Pot. | C1 | Gauge | Pot. | C1 | Gauge | Pot. |
| 1 | 98.70 | 0.00 | 1.00 | 99.83 | 0.00 | 1.00 | 99.83 | 0.00 | 1.00 | 99.22 | 0.00 | 1.00 |
| 2 | 98.52 | 0.00 | 1.00 | 99.37 | 0.00 | 1.00 | 99.37 | 0.00 | 1.00 | 98.94 | 0.00 | 1.00 |
| 3 | 79.36 | 0.01 | 0.95 | 95.23 | 0.00 | 0.99 | 95.97 | 0.00 | 0.98 | 62.01 | 0.00 | 0.81 |
| 4 | 98.59 | 0.00 | 1.00 | 99.16 | 0.00 | 1.00 | 99.16 | 0.00 | 1.00 | 99.29 | 0.00 | 1.00 |
| 5 | 98.79 | 0.00 | 1.00 | 99.86 | 0.00 | 1.00 | 99.86 | 0.00 | 1.00 | 99.26 | 0.00 | 1.00 |
| 6 | 98.70 | 0.00 | 1.00 | 99.22 | 0.00 | 1.00 | 99.22 | 0.00 | 1.00 | 99.19 | 0.00 | 1.00 |
| 6A | 65.31 | 0.00 | 0.88 | 78.37 | 0.01 | 0.98 | 96.24 | 0.00 | 0.99 | 85.30 | 0.01 | 0.93 |
| 6B | 97.61 | 0.00 | 1.00 | 98.57 | 0.00 | 1.00 | 99.37 | 0.00 | 1.00 | 98.38 | 0.00 | 1.00 |
| 7 | 92.66 | 0.00 | 0.99 | 97.09 | 0.00 | 1.00 | 99.50 | 0.00 | 1.00 | 98.76 | 0.00 | 1.00 |
| 8 | 98.44 | 0.00 | 1.00 | 99.91 | 0.00 | 1.00 | 99.92 | 0.00 | 1.00 | 99.05 | 0.00 | 1.00 |
| 9 | 91.38 | 0.00 | 0.99 | 96.53 | 0.00 | 1.00 | 99.61 | 0.00 | 1.00 | 98.18 | 0.00 | 1.00 |
| Mean | 92.55 | 0.00 | 0.98 | 96.65 | 0.00 | 1.00 | 98.91 | 0.00 | 1.00 | 94.33 | 0.00 | 0.98 |

- HP performance is increased significantly due to optimisation (46.1% to 94.6% exact DGP)
- Our full algorithm improves performance, though requires extensions for best performance. Good improvement on DGPs 3 and 6A.
- Small percentage point increase, but 5-fold decrease in incorrect identification (depending on how you look at it!). These are the hardest data sets.

# Results – General Performance

- In real situations, best tuning parameter values would not be known.
- We can compare qualitatively what happens when tuning parameters are varied (acknowledging that scales are not strictly comparable)




- HP algorithm (left) has a sharp peak at about $\alpha$=10e-4. Much higher or lower values result in large drop in performance
- Our algorithm shows robust performance over all $\varphi$. $\alpha$ is in a sense optimised for automatically.

## Conclusions

- $S_T$ can be used to rank regressors using the “triggers” problem-framing
- Appears to have advantages over t-scores in certain DGPs
- Our algorithm can outperform that of Hoover and Perez, both in the optimised (and perhaps more notably) in the unoptimised case.
- We hope that $S_T$ could be a useful tool to econometricians.

## Open Questions

- Can we prove any theoretical properties of the procedure (e.g. is it consistent)?
- Performance against other model selection procedures
- Generalising to other test functions

# Extra

## Ranking Regressors

$$\delta = \frac{\text{Size of smallest ranked set including DGP}}{\text{Size of DGP set}}$$

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

- $S_T$ outperforms the t-statistics on average
- Performance is however dependent on the DGP
- $S_T$ better for DGPs 3 and 6A in particular
- t-test better for DGP 8

Suggests that $S_T$ might be better overall, but a hybrid between the two measures could be best?

# Effective DGP (EDGP)

Hoover and Perez say in their work: “searches for both DGPs 6 and 9 most frequently end in failure. This suggests, not a failure of the algorithm, but unavoidable properties of the data.”

In these DGPs the signal-to-noise ratio is too low for certain variables to be identified.

We propose the concept of an “effective DGP” (EDGP), as the (sub)set of DGP variables which can be reasonably identified by an algorithm given the signal-to-noise ratio.

- We use a tool called the “parametricness index” (PI):

$$IC_{\lambda_n,d}(\mathbf{z}_k, \hat{\sigma}^2) = ||\mathbf{y} - \hat{\mathbf{y}}_k||^2 + \lambda_n \log(n) r_k \hat{\sigma}_y^2 - n\hat{\sigma}_y^2 + dn^{1/2} \log(n)\hat{\sigma}_y^2$$

$$PI = \begin{cases} \inf_{\mathbf{z}_k \in \zeta_1(\hat{\mathbf{z}}_k)} \left( \frac{IC_{\lambda_n,d}(\mathbf{z}_k, \hat{\sigma}_y^2)}{IC_{\lambda_n,d}(\hat{\mathbf{z}}_k, \hat{\sigma}_y^2)} \right) & \text{if } r_{\hat{k}} > 1 \\ n & \text{if } r_{\hat{k}} = 1 \end{cases}$$

- Measures ratio of information criteria when removing regressors one by one with replacement.
- Essentially a measure of suitability of a model for given data set.
- PI < 1.2 indicates non-parametric (i.e. one or more regressors can be removed without significantly affecting model quality)

We examine PI values of DGPs and IC ratios of removing specific regressors…

# Effective DGP (EDGP)

| DGP | DGP Indices | $F_m(1.2)$ | $PI_{0.01}$ | $PI_{0.1}$ | $PI$ (mean) | $PI_{0.9}$ | $PI_{0.99}$ | 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} |

DGPs 6 and 9 have significantly lower mean PI (av. over 5000 samples). These correspond to those identified by HP. Now examining IC ratios…

| DGP | Variable | $F_m(1.2)$ | $ICR_{0.01}$ | $ICR_{0.1}$ | $ICR$ (mean) | $ICR_{0.9}$ | $ICR_{0.99}$ |
|---|---|---|---|---|---|---|---|
| 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 |

Weak regressors are defined as mean ICR < 1.2.

We now redefine EDGPs by removing these regressors (this changes DGPs 6 and 9)

# Results – Optimum Performance

Check results compared to true DGPs:

| DGP | ST (no skip, no adapt) $\alpha = 0.0371$ | | | ST (no skip) $\varphi = 0.3$ | | | ST (full) $\varphi = 0.3$ | | | HP $\alpha = 4 \cdot 10^{-4}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | Gauge | Pot. | C1 | Gauge | Pot. | C1 | Gauge | Pot. | C1 | Gauge | Pot. |
| 6 | 0.010 | 0.001 | 0.500 | 0.010 | 0.000 | 0.500 | 0.020 | 0.000 | 0.500 | 0.030 | 0.000 | 0.499 |
| 9 | 0.000 | 0.002 | 0.592 | 0.000 | 0.000 | 0.600 | 0.000 | 0.000 | 0.600 | 0.000 | 0.000 | 0.599 |
| Mean | 0.010 | 0.001 | 0.546 | 0.010 | 0.001 | 0.550 | 0.010 | 0.000 | 0.550 | 0.020 | 0.000 | 0.549 |

- Results are only different for DGPs 6 and 9
- No correct IDs of the DGP for any algorithms due to signal to noise ratio too low.