

Extending the Boundaries of PcGets: Non-linear Models.

Jennifer Castle and David F. Hendry*
Department of Economics, Oxford University.

June 23, 2005

Abstract

Econometric modelling using automatic algorithms such as PcGets has become prevalent in recent years. This paper considers automatic model selection when there is non-linearity inherent in the data. Non-linearity poses a number of problems. We address issues of collinearity generated by non-linear transformations, extreme observations leading to fat tailed distributions and a failure of normality, and how to approximate non-linearity in the data by sufficiently general expansions whilst avoiding excess retention of irrelevant variables. Solutions to all these problems are proposed. Not only is a non-linear capability for PcGets feasible, but it is an essential extension that would broaden the scope of PcGets in keeping with its general-to-specific philosophy.

1 Introduction

Some method of model selection is required for empirical econometric modelling. One systematic approach to model selection is the general-to-specific (*Gets*) methodology, automated in the PcGets program developed by Hendry and Krolzig (2001). An extensive assessment of Monte Carlo simulation studies has revealed that the operational characteristics of the PcGets algorithm are excellent across a wide range of states of nature. However, the program is currently designed to select linear models. Non-linearity is inherent in economic data, but we often make a simplifying assumption to reduce the model to a linear representation. If the underlying process is indeed non-linear, the model will be misspecified. This paper examines the feasibility of selecting non-linear models within PcGets, with a view to extending the algorithm to handle a wide class of non-linear functions.

The proposed strategy for selecting non-linear models is to commence with a general model that includes all potentially relevant variables. As the number of possible non-linear functions is enormous, we shall use a class of general non-linear approximations to capture the non-linearity in the data. The class we focus on is polynomials. Once a general unrestricted model (GUM) is specified and is congruent, we shall use the PcGets selection algorithm to select an undominated congruent final model.

There are problems with selecting under non-linearity that need to be addressed before an operational algorithm can be implemented. First, undertaking a non-linear transformation of a variable can generate substantial collinearity between the original linear and the transformed non-linear function. Collinearity is problematic for any selection procedure because the information content of an extra highly collinear variable is very small, yet disrupts existing information attribution. A model selection algorithm will struggle to determine the relevant variables, and will therefore select poorly between the relevant and irrelevant variables depending on sampling error. Orthogonality is highly beneficial for model selection in general, both for that reason, and because deleting small, insignificant coefficients leaves the retained

*Preliminary and incomplete, prepared for the first ESF-EMM conference: please do not cite without the author's permission. Financial support from the ESRC under grant RES051270035 is gratefully acknowledged.

estimates almost unaltered. We provide operational rules to transform the non-linear models to a more orthogonal form prior to undertaking model selection, and this results in greatly improved properties of the selection procedure.

We also address the consequential problem of extreme observations which some classes of non-linearity can generate. The resulting fat-tailed distributions may cause problems, because the assumption of normality is often inbuilt for the critical values used by model selection procedures. Non-linear functions can ‘line up’ with outliers causing the functions to be retained too often. We propose a solution by removing the extreme observations using indicator saturation techniques developed by Hendry, Johansen and Santos (2004) to ensure near normality for inference. This also avoids the problem of undetectable outliers.

Finally, we consider the types of non-linear functions that might capture non-linearity inherent in economic data. A sufficiently broad class of functions is needed, and we focus on polynomials. This class can be easily orthogonalized, maintains linearity in the parameters and approximates a wide range of non-linear models such as Smooth Transition Regression models (STR). We seek to control the problem of excess retention of non-linear functions due to an overparameterized GUM by proposing a ‘super-conservative’ strategy for selecting non-linear functions.

The structure of the paper is as follows. Section 2 addresses the problems of collinearity between linear and non-linear functions, proposing a solution of orthogonalizing; operational rules are provided for the static, stationary dynamic, and unit-root cases. Section 3 outlines the issue of non-normality, demonstrating the problems of fat tails. A Monte Carlo study highlights the problem of extreme observations for model selection, and explains the solution of using indicator saturation techniques prior to model selection. Section 4 considers the polynomial class of functions and investigates their ability to approximate a STR model. The super-conservative strategy is outlined to ensure non-linear functions are retained only when there is definite evidence of non-linearity in the data. Finally, section 5 concludes.

2 Collinearity

Multicollinearity was first outlined by Frisch (1934) within the context of static general equilibrium linear relations. Confluence analysis was developed to address the problem, although this method is not in common practice now (see Hendry and Morgan, 1989). The definition of collinearity has shifted over the years, and we can define perfect collinearity as $|\mathbf{X}'\mathbf{X}| = 0$, and perfect orthogonality as a diagonal $(\mathbf{X}'\mathbf{X})$ matrix. Since collinearity is not invariant under linear transformations, it is difficult to identify the degree of collinearity. A linear model is invariant under linear transformations, and so a model could be defined by various isomorphic representations which nevertheless deliver very different inter-correlations. Hence, collinearity is a property of the parameterization of the model and not the variables *per se*.

Non-linear transformations can generate substantial collinearity between the linear and non-linear functions. We initially consider a simple case in which we add the transformation $f(x) = x^2$. This polynomial transform is common in economics, for example, age and the square of age often enter in labour force participation models. If we consider a white-noise process given in equation (2), the collinearity between x and x^2 is 0 when x has a mean of zero, but a non-zero mean dramatically changes these results. To show this, section 2.1 derives analytical results for this simple case.

2.1 Analytical results for the correlation of x and x^2

We formulate the data generating process (DGP) as the linear conditional relation:

$$y_i = x_i + e_i = 0 + x_i + 0x_i^2 + e_i \tag{1}$$

with:

$$x_i \sim \text{IN}[0, 1] \quad (2)$$

$$e_i \sim \text{IN}[0, 1]. \quad (3)$$

Since (1) is invariant under linear transformations, it can also be written as for $z_i = x_i + \mu$:

$$\begin{aligned} y_i &= -\mu + (x_i + \mu) + 0(x_i + \mu)^2 + e_i \\ &= -\bar{z} + z_i + 0z_i^2 + e_i \\ &= 0 + (z_i - \bar{z}) + 0(z_i - \bar{z})^2 + e_i \end{aligned} \quad (4)$$

We consider 3 models, including the zero-mean case:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i, \quad (5)$$

the complete zero-mean case:

$$y_i = \lambda_0 + \lambda_1 x_i + \lambda_2 (x_i^2 - \bar{x}^2) + u_i = \lambda_0 + \lambda_1 x_i + \lambda_2 w_i + u_i, \quad (6)$$

and the non-zero mean case:

$$y_i = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2 + u_i. \quad (7)$$

First, examining the general case, equation (7), in which there is a non-zero mean:

$$\mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(\mu)}] = \mathbb{E} \left[\begin{pmatrix} 1.0 & \bar{z} & \bar{z}^2 \\ \bar{z} & T^{-1} \sum z_i^2 & T^{-1} \sum z_i^3 \\ \bar{z}^2 & T^{-1} \sum z_i^3 & T^{-1} \sum z_i^4 \end{pmatrix} \right] = \begin{pmatrix} 1.0 & \mu & \mu^2 + 1 \\ \mu & \mu^2 + 1 & \mu^3 + 3\mu \\ \mu^2 + 1 & \mu^3 + 3\mu & 3 + \mu^4 + 6\mu^2 \end{pmatrix} \quad (8)$$

with the inverse computed as:

$$(\mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(\mu)}])^{-1} = 0.5 \begin{pmatrix} \mu^4 + 3 & -2\mu^3 & \mu^2 - 1 \\ -2\mu^3 & 2 + 4\mu^2 & -2\mu \\ \mu^2 - 1 & -2\mu & 1 \end{pmatrix}. \quad (9)$$

There is substantial collinearity between the variables, except for the squared term which is irrelevant in the DGP.

Next, consider the zero-mean model in equation (5):

$$\mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(0)}] = \mathbb{E} \left[\begin{pmatrix} 1.0 & \bar{x} & \bar{x}^2 \\ \bar{x} & T^{-1} \sum x_i^2 & T^{-1} \sum x_i^3 \\ \bar{x}^2 & T^{-1} \sum x_i^3 & T^{-1} \sum x_i^4 \end{pmatrix} \right] = \begin{pmatrix} 1.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 3.0 \end{pmatrix} \quad (10)$$

so the inverse is:

$$(\mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(0)}])^{-1} = \begin{pmatrix} 1.5 & 0 & -0.5 \\ 0 & 1.0 & 0 \\ -0.5 & 0 & 0.5 \end{pmatrix}. \quad (11)$$

There is no collinearity between x and x^2 . There is an effect on the intercept, as the ‘correlation’ between the intercept and x^2 is -0.577 , but this does not cause a problem for the PcGets selection algorithm, as demonstrated by the Monte Carlo evidence in section 2.2.

Then, examining the complete zero-mean model in equation (6):

$$\mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(0,0)}] = \mathbb{E} \left[\begin{pmatrix} 1.0 & \bar{x} & \bar{w} \\ \bar{x} & T^{-1} \sum x_i^2 & T^{-1} \sum x_i w_i \\ \bar{w} & T^{-1} \sum x_i w_i & T^{-1} \sum w_i^2 \end{pmatrix} \right] = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 3.0 \end{pmatrix} \quad (12)$$

as:

$$T^{-1} \sum x_i w_i = T^{-1} \sum x_i (x_i^2 - \bar{x}^2) = T^{-1} \sum x_i^3 - \bar{x}^2 T^{-1} \sum x_i = T^{-1} \sum x_i^3 - \bar{x} \bar{x}^2,$$

so the inverse is:

$$(\mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(0,0)}])^{-1} = \begin{pmatrix} 0.33 & 0.0 & 0.0 \\ 0.0 & 0.33 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}. \quad (13)$$

We noted that collinearity is a property of the parameterization of the model and so we seek a near orthogonal representation of the general model. This can be achieved simply by taking deviations from means, which re-creates the specification in terms of the original variables x and x^2 . Observe that $z_i = x_i + \mu$ where $\bar{z} = \mu$ and $\bar{z}^2 = \mu^2 + 1$. Hence, we can calculate:

$$\begin{aligned} \mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(\bar{\mu})}] &= \mathbb{E} \left[\begin{pmatrix} 1.0 & z_i - \bar{z} & z_i^2 - \bar{z}^2 \\ z_i - \bar{z} & T^{-1} \sum (z_i - \bar{z})^2 & T^{-1} \sum (z_i - \bar{z}) (z_i^2 - \bar{z}^2) \\ z_i^2 - \bar{z}^2 & T^{-1} \sum (z_i - \bar{z}) (z_i^2 - \bar{z}^2) & T^{-1} \sum (z_i^2 - \bar{z}^2)^2 \end{pmatrix} \right] \\ &= \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 2\mu \\ 0.0 & 2\mu & 4\mu^2 + 2 \end{pmatrix} \end{aligned} \quad (14)$$

with the inverse:

$$(\mathbb{E} [T^{-1} \mathbf{X}' \mathbf{X}_{(\bar{\mu})}])^{-1} = 0.5 \begin{pmatrix} 2.0 & 0.0 & 0.0 \\ 0.0 & 4\mu^2 + 2 & -2\mu \\ 0.0 & -2\mu & 1.0 \end{pmatrix}. \quad (15)$$

Taking deviations from means delivers some reduction in collinearity, which is particularly marked for the intercept, but worse for the linear term $(z_i - \bar{z})$. Again the irrelevant squared term ‘benefits’.

If we assume $\mu = 10$, the correlation between x and x^2 is -0.9775 , the ‘correlation’ of x with the intercept is -0.99735 and the ‘correlation’ of x^2 with the intercept is 0.98985 . Hence, the general case can generate near perfect collinearity. Taking deviations from means, as in equation (14) results in a correlation of 0.99875 for x and x^2 , and so this correlation is higher after undertaking an ‘orthogonalizing’ transform with the intercept. To remove the collinearity between x and x^2 , we also need to de-mean x^2 . Hence, the linear term remains $(z_i - \bar{z})$, but the squared term becomes $(z_i - \bar{z})^2 - [\mathbb{E} (z_i - \bar{z})^2]$ which will result in a model that is identical to equation (6) under the zero-mean case, with the inverse given in equation (13). Double de-meaning has removed the collinearity generated by the non-zero mean in the white-noise process.

2.2 Monte Carlo evidence on collinearity

Given these simple analytical results, we consider Monte Carlo evidence assessing the probability of retaining relevant and irrelevant variables when selecting a specific model from the GUM. We consider 3 models in which the DGP is a static process, a dynamic process, and a unit-root process. Monte Carlo evidence for linear models is given in Hendry and Krolzig (1999, 2003, 2005) which shows that PcGets

retains relevant variables close to the theoretical maximum given by an independent t-test at significance level α . PcGets also eliminates variables at the chosen significance level, and so the ‘size’ and ‘power’ of the test battery is controlled.

PcGets offers 2 pre-programmed strategies, denoted the liberal and conservative strategies, which are calibrated to deliver a 5% and 1% probability of retaining nuisance parameters respectively. The liberal strategy is looser in that it focuses on minimizing the non-selection probability of relevant variables whereas the conservative strategy is tighter, focusing on minimizing the non-deletion probability of nuisance variables. Results for both strategies are reported.

2.2.1 A static model

The DGP is given by:

$$y_t = \beta x_t + u_t \quad u_t \sim \text{IN} [0, \sigma_u^2] \quad (16)$$

$$x_t = \mu + \nu_t \quad \nu_t \sim \text{IN} [0, \sigma_v^2] \quad (17)$$

for $t = 1, \dots, T$. We examine both $\mu = 0$ and $\mu = 10$. We set $\sigma_u^2 = \sigma_v^2 = 1$ and assess 2 sample sizes, $T = 100$ and 1000. The number of replications, M , is 10,000. To ensure the probability of retaining the relevant variable is near unity we set $E[t_\beta] = \psi = 5$, which corresponds to $\beta = 0.5$ for $T = 100$ and $\beta = 0.15811$ for $T = 1000$ in the orthogonal case. The theoretical maximum power is 0.9987 at the 5% level and 0.9912 at the 1% significance level. The GUM contains 3 variables, two of which are nuisance, and is given by:

$$y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + \epsilon_t, \quad \epsilon_t \sim \text{IN} [0, \sigma_\epsilon^2]. \quad (18)$$

To overcome the collinearity problem induced by the non-zero mean, we seek a near orthogonal representation of the model. As $E[x_t] = \mu$ and $E[x_t^2] = \mu^2 + \sigma_v^2$, the rules used to de-mean are given by:

$$\bar{x} = x_t - \mu, \quad (19)$$

$$\overline{x^2} = (x_t - \mu)^2 - \sigma_v^2. \quad (20)$$

Results Figure 1 records the probability of retaining variables for $\mu = 0$, $\mu = 10$ and the de-meaned case using rules (19) and (20). The $\mu = 0$ case is analogous to a linear GUM where the probability of retaining x_t is almost 1 and is higher for the liberal strategy than the conservative strategy. The probability of retaining the irrelevant variables is marginally higher than 5% and 1%, but it is essentially controlled at the selected significance levels and reliability weighting would reduce this size further. However, for the $\mu = 10$ case, there is a dramatic fall in the probability of retaining x_t to below 60% and a corresponding increase in the probability of retaining the intercept and x_t^2 to over 40%. The conservative strategy can deliver higher power than the liberal strategy (for the $T = 100$ case), implying that it deletes the irrelevant collinear variable more often and so finds the DGP variable slightly more often. The transformation to an orthogonal representation results in correct retention probabilities for the liberal and conservative strategies, matching the $\mu = 0$ results. Hence, by undertaking these simple operational rules the collinearity between x_t and x_t^2 is removed.

2.2.2 Varying non-centralities for the DGP variables

We next consider the properties of PcGets when the non-centralities of the DGP variables vary. The DGP is formulated in equation (21), in which two linear variables now enter the DGP:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + u_t \quad u_t \sim \text{IN} [0, \sigma_u^2] \quad (21)$$

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\nu}_t \quad \boldsymbol{\nu}_t \sim \text{IN}_2 [0, \boldsymbol{\Omega}_\nu] \quad (22)$$

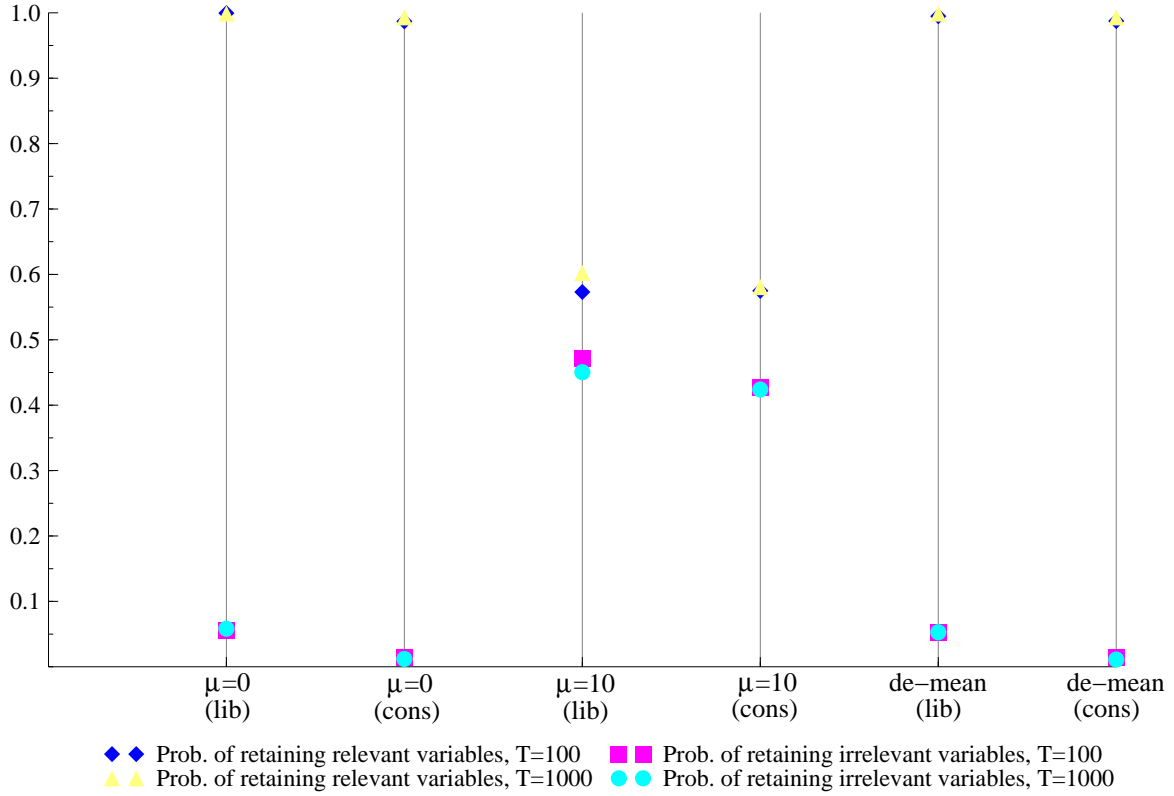


Figure 1: The probability of PcGets retaining variables after selection for the static non-linear case, comparing orthogonal and collinear models.

for $t = 1, \dots, T$, where $\Omega_{\nu,ij} = 0$ for $i \neq j$ and $\Omega_{\nu,ii} = \sigma_u^2 = 1$ for $i = 1, 2$. The GUM contains 3 nuisance parameters and is given by:

$$y_t = \alpha_0 + \alpha_1 x_{1,t} + \alpha_2 x_{2,t} + \alpha_3 x_{1,t}^2 + \alpha_4 x_{2,t}^2 + \epsilon_t, \quad \epsilon_t \sim \text{IN}[0, \sigma_\epsilon^2]. \quad (23)$$

Four cases are considered in which the non-centralities of the DGP variables are varied. If we define $E[t_{\beta_i}] = \psi_{\beta_i}$, under orthogonality the cases include:

$$\begin{aligned} \psi_{\beta_1} &= \psi_{\beta_2} = 2, \\ \psi_{\beta_1} &= \psi_{\beta_2} = 3, \\ \psi_{\beta_1} &= \psi_{\beta_2} = 4, \\ \psi_{\beta_1} &= 3, \quad \psi_{\beta_2} = 6. \end{aligned}$$

Results are reported for $T=100$ and $M = 10,000$ replications are undertaken.¹

Results Figure 2 records the retention probabilities of the liberal and conservative strategies for the 4 cases outlined when $\mu = 10$ and after de-meaning using rules (19) and (20). Transforming to an orthogonal representation increases the probability of retaining the relevant variables (other than for the conservative strategy with $\psi_{\beta_i} = 2$) and tightens the probability of retaining the irrelevant variables to 5% and 1% for the liberal and conservative strategies respectively. Figure 3 records the power for a t-test of a single null hypothesis, H_0 , where $\psi_{\beta_i} = 0$ under the null using a 2-sided test at critical value c_α . To calculate the power to reject the null when $E[t_{\beta_i}] = \psi_{\beta_i} > 0$ we use:

$$P(t \geq c_\alpha \mid E[t] = \psi) \simeq P(t - \psi \geq c_\alpha - \psi \mid H_0).$$

¹Results for $T = 1000$ were analogous to those for $T = 100$ and are available on request.

There is a 50% chance of retaining a single variable with a $|t| = 2$ when $\alpha = 0.05$ but this falls to 27% when $\alpha = 0.01$. The power to detect relevant variables increases with the non-centrality, and $|t|$ s of 4 are retained above 90% of the time. We also consider the theoretical probability of retaining 2 relevant variables, matching the 4 cases examined. PcGets retention rates for the de-meanned variables are excellent compared to the theoretical single t-test results, matching the linear studies by Hendry and Krolzig (1999, 2003).

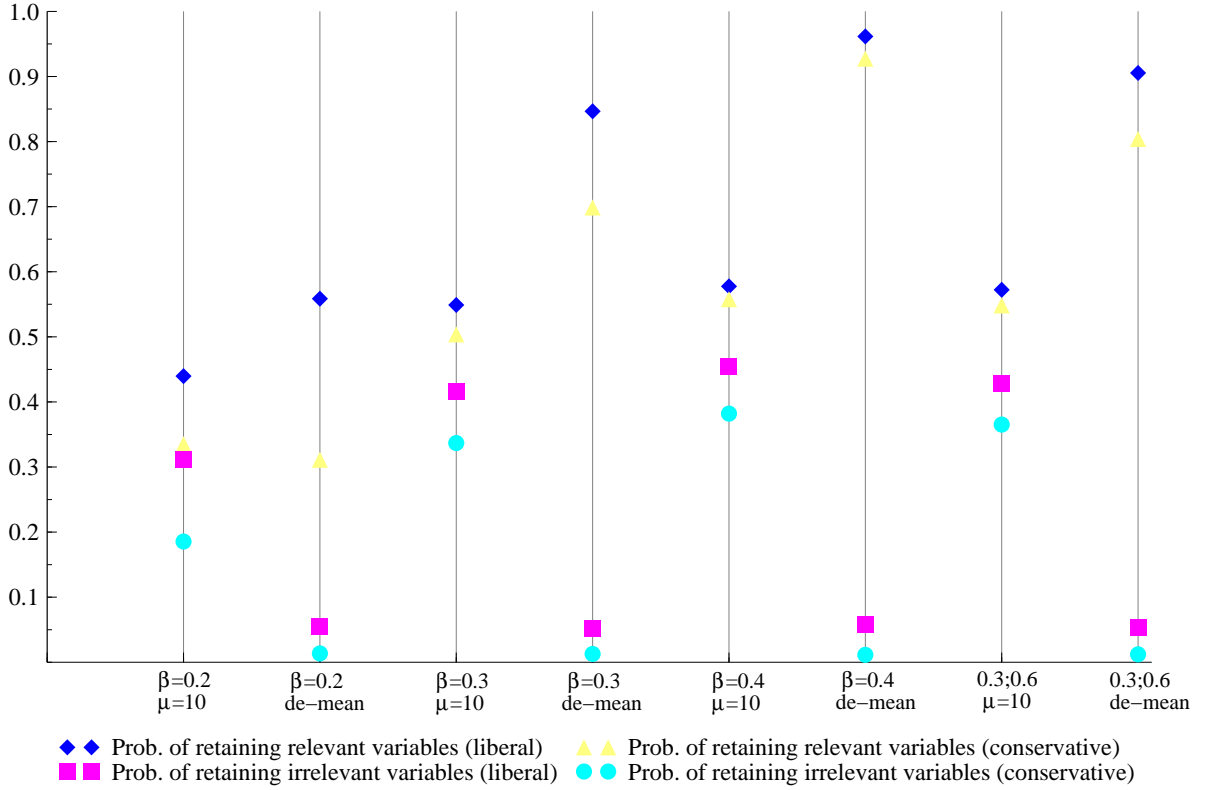


Figure 2: Probability of retaining relevant and irrelevant variables for the static non-linear case in which there are 2 dependent variables.

2.2.3 A stationary AR(1) process

We next consider a stationary AR(1) process for the regressor in which the DGP is given by:

$$y_t = \beta x_t + u_t \quad u_t \sim \text{IN} [0, \sigma_u^2] \quad (24)$$

$$x_t = \mu + \rho x_{t-1} + \nu_t \quad \nu_t \sim \text{IN} [0, \sigma_v^2] \quad (25)$$

for $t = 1, \dots, T$. We set $\sigma_u^2 = \sigma_v^2 = 1$ and $\rho = 0.8$. We examine the zero mean case ($\mu = 0$) and the case where $E[x_t] = \mu / (1 - \rho) = 10$, setting $\mu = 2$. We follow the static case by de-meaning and we use two rules. The first removes the population means and the second removes the sample means:

$$1) \quad \bar{x} = x_t - \frac{\mu}{1 - \rho}, \quad \overline{x^2} = \left(x_t - \frac{\mu}{1 - \rho} \right)^2 - \frac{\sigma_v^2}{1 - \rho^2}. \quad (26)$$

$$2) \quad \bar{x} = x_t - \frac{\hat{\mu}}{1 - \hat{\rho}}, \quad \overline{x^2} = \left(x_t - \frac{\hat{\mu}}{1 - \hat{\rho}} \right)^2 - \frac{\hat{\sigma}_v^2}{1 - \hat{\rho}^2}. \quad (27)$$

The first 50 observations are discarded for each replication. We set $\beta = 0.5$ for $T = 100$ and $\beta = 0.15811$ for $T = 1000$, with $M = 10,000$ replications. We consider two GUMs, equation (28) in which

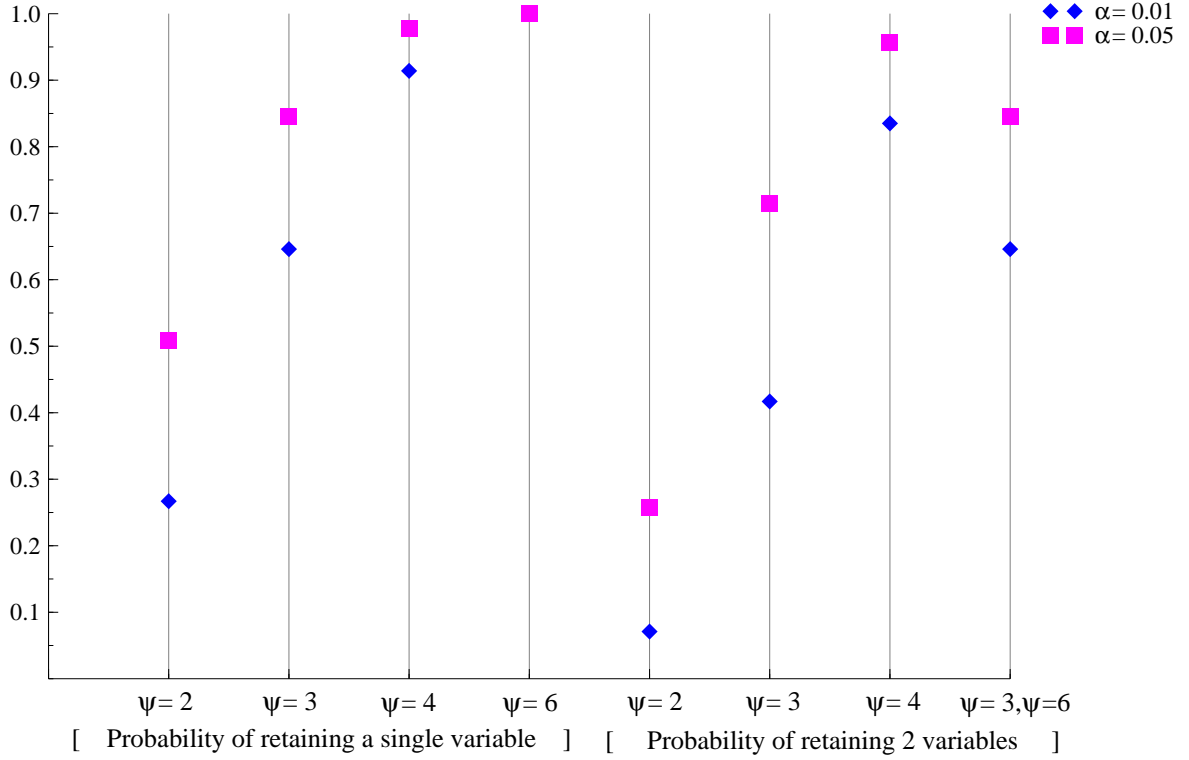


Figure 3: t-test powers for a single null hypothesis test and for 2 null hypothesis tests

there are 5 nuisance parameters and equation (28) with no dynamics where $\alpha_1 = \alpha_3 = \alpha_5 = 0$.

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 x_t + \alpha_3 x_{t-1} + \alpha_4 x_t^2 + \alpha_5 x_{t-1}^2 + \epsilon_t, \quad \epsilon_t \sim \text{IN} [0, \sigma_\epsilon^2]. \quad (28)$$

Results Figure 4 records the retention probabilities of PcGets for the AR(1) process for the regressor when $T = 100$.² The retention probabilities for the GUM containing 3 variables and 6 variables are almost identical, indicating that exogenous dynamics do not affect the selection probabilities to the extent that the non-linear functions do. With a zero mean process, the probability of retaining x_t is near unity for both strategies, and the probability of retaining irrelevant variables corresponds to 5% and 1% for the liberal and conservative strategies respectively. As in the static case, a mean of 10 results in a decrease in the probability of retaining x_t to below 70% and the probability of retaining irrelevant variables is much higher for the 3-variable GUM than the 6-variable GUM. This is because the collinearity between x_t and x_t^2 is causing the problem rather than the collinearity between the variables dated t and $t - 1$. Therefore averaging retention probabilities across two variables results in a higher size than averaging across 5 variables. There is a correlation between variables dated t and $t - 1$ with retention probabilities of about 14% for the liberal strategy and about 3% for the conservative strategy, but the average retention probabilities for the square and intercept are about 35% and 33% respectively.

If we knew the underlying process in equation (25), we could de-mean using rule (26) and this removes the collinearity problem. However, it is unlikely that we would know the true μ , ρ and σ_v^2 parameters. Rule (27) is based on the sample estimates and this has no impact on retention probabilities compared to population values and so we can use estimated values in our operational rules.

²Results for $T = 1000$ are similar and are therefore not reported, but are available on request.

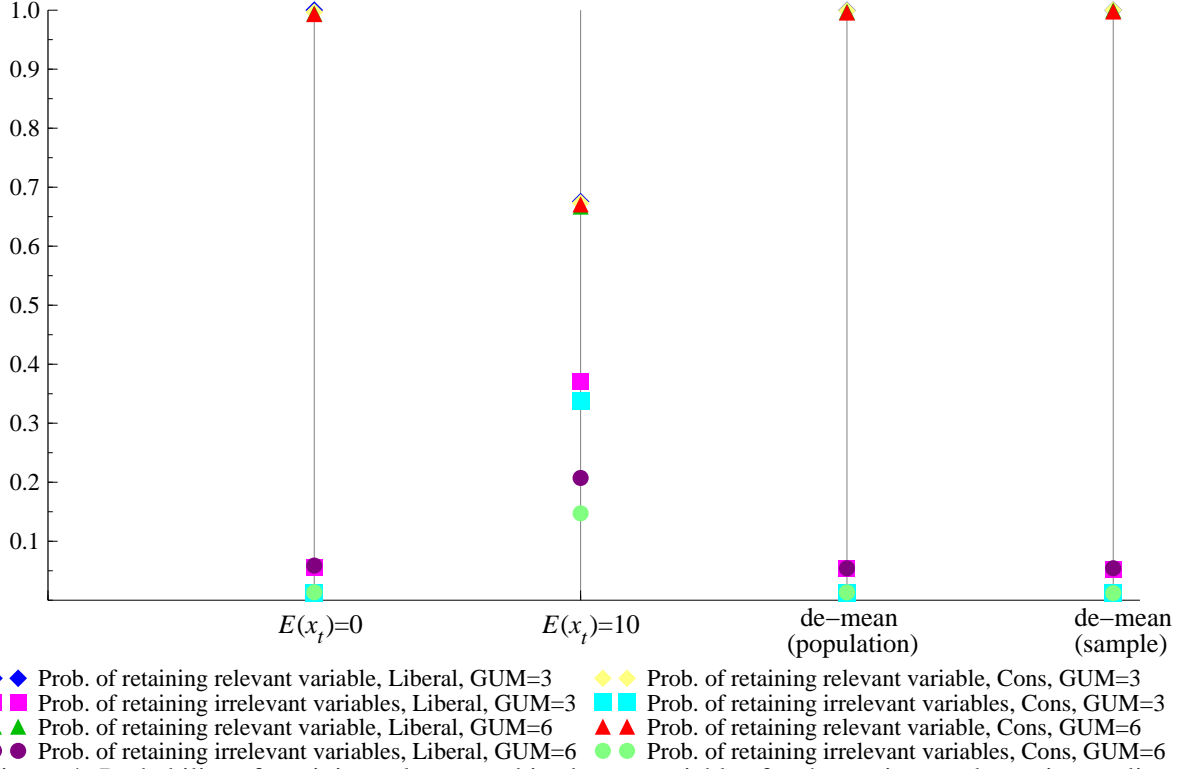


Figure 4: Probability of retaining relevant and irrelevant variables for the stationary dynamic non-linear case. $T = 100$.

2.2.4 A unit-root process

Finally we consider a DGP that consists of a unit-root process outlined in equation (30).

$$y_t = \beta x_t + u_t \quad u_t \sim \text{IN} [0, \sigma_u^2] \quad (29)$$

$$x_t = x_{t-1} + \nu_t \quad \nu_t \sim \text{IN} [0, \sigma_\nu^2] \quad (30)$$

for $t = 1, \dots, T$. We consider 2 initial conditions, $x_0 = 0$ and 10 and set $\sigma_u^2 = \sigma_\nu^2 = 1$. A coefficient of $\beta = 0.5$ is used for $T = 100$ and $\beta = 0.15811$ for $T = 1000$ with $M = 10,000$. The GUM is given by equation (28), both the full equation, and setting $\alpha_1 = \alpha_3 = \alpha_5 = 0$.

In order to de-mean the data, we consider removing sample averages, given in rule (31).

$$\bar{x} = x_t - \frac{1}{T} \sum_{t=1}^T x_t, \quad \bar{x}^2 = \left(x_t - \frac{1}{T} \sum_{t=1}^T x_t \right)^2 - \frac{1}{T} \sum_{t=1}^T \left(x_t - \frac{1}{T} \sum_{t=1}^T x_t \right)^2. \quad (31)$$

Results Figure 5 records retention probabilities for the exogenous unit-root process for $T = 100$. For an initial condition of zero, the probability of retaining x_t for both the liberal and conservative strategy is unity and the probability of retaining irrelevant variables is approximately 5% and 1% respectively. Hence, l(1)ness has no impact on selection when there are non-linear functions. However, imposing an initial condition of 10 dramatically reduces the probability of retaining relevant variables and increases the probability of retaining irrelevant variables. For a GUM of 6, the probability of retaining x_t is marginally higher for the conservative strategy compared to the liberal strategy. Rule (31) which removes the sample means from the level and square results in a retention probability of 1 for x_t with corresponding correct probabilities for the retention of the irrelevant variables. Thus, rule (31) ensures near orthogonal

non-linear regressors for I(1) variables. Observe that removing the initial condition would deliver results analogous to those in which there is a zero initial condition.³ Removing the initial condition would result in a retention probability of approximately 5% and 1% for irrelevant variables and a retention probability near to the theoretical upper bound for relevant variables.

Figure 6 records retention probabilities for the unit root process for $T = 1000$ and it is clear that the adverse retention probabilities are mitigated as the sample size increases. The random walk has deviated from the initial condition substantially and whilst the correlations do still depend on x_0 the impact is rapidly declining with T . Surprisingly, the probability of retaining relevant variables is marginally higher for the conservative strategy than that of the liberal strategy for $x_0 = 10$ and 6 variables, with a probability of 0.982 compared to 0.977. As de-meaning using the sample average yields no cost, it would be recommended regardless of the sample size.

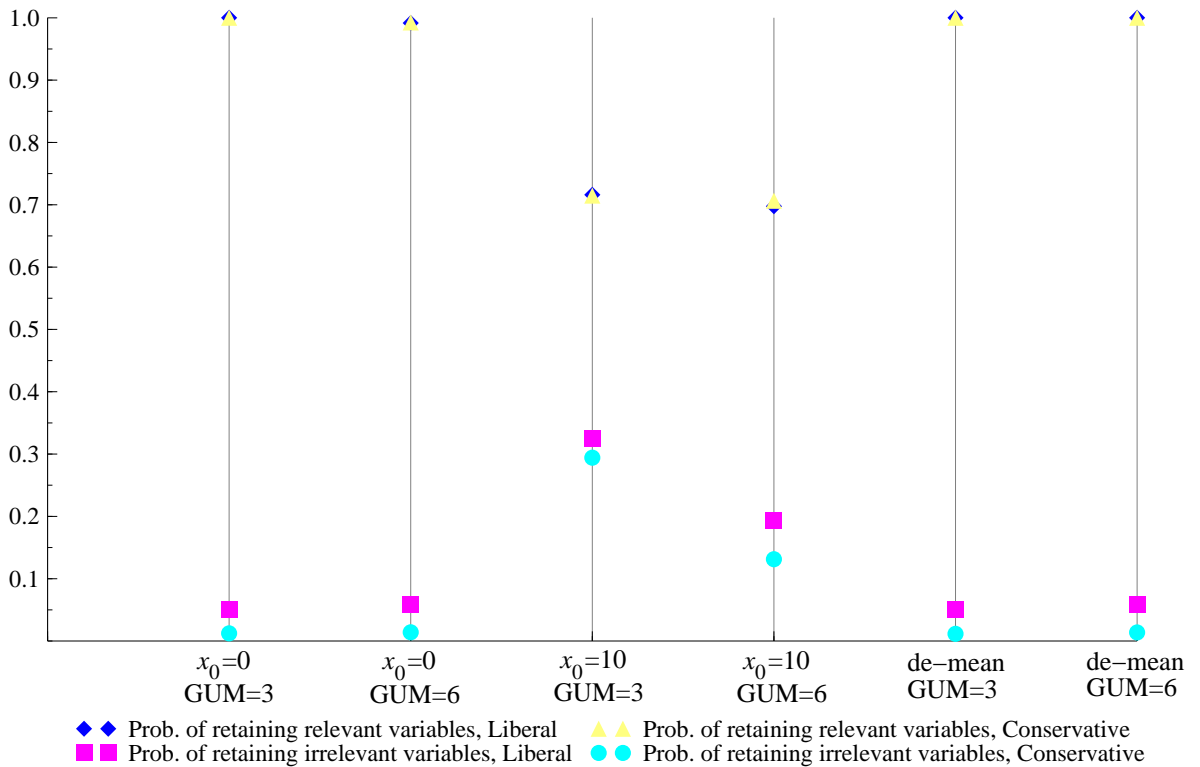


Figure 5: Probability of retaining relevant and irrelevant variables for the unit-root case for a sample size of 100

2.2.5 A differenced I(1) process

Whilst we have identified the primary difficulty in this aspect of model selection to be collinearity between variables and their corresponding non-linear transformations, the correlation between variables and their lags is also significant. Retention of x_{t-1} and x_{t-1}^2 is approximately 3 times higher than the 5% and 1% size of the liberal and conservative strategies respectively for both the stationary process when $E[x_t] = 10$ and the non-stationary process when $x_0 = 10$ for $T = 100$. Hence, we next consider the impact of differencing to remove the collinearity between the variables and their lags.

The DGP is given as:

$$y_t = \beta x_{t-1} + u_t \quad u_t \sim \text{IN} [0, \sigma_u^2] \quad (32)$$

³In practice the initial condition will be unknown and the first observation, x_1 , could be used as an estimate of the initial condition x_0 .

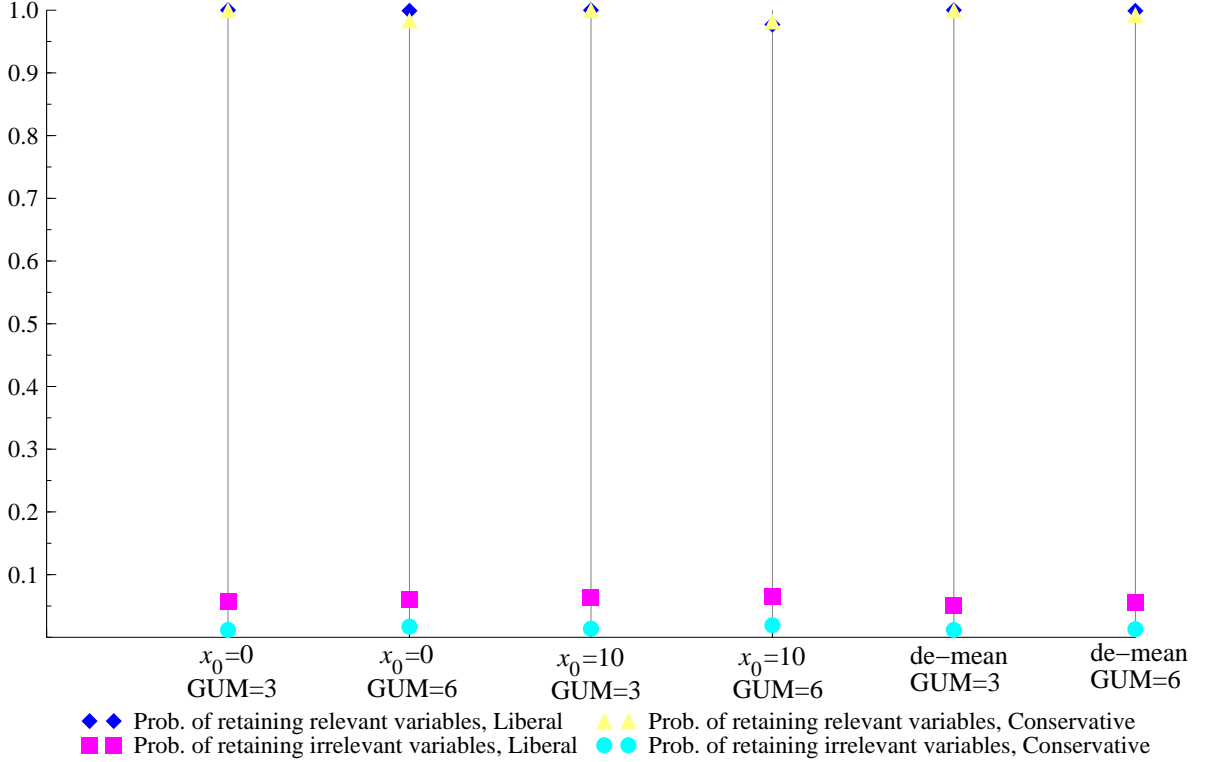


Figure 6: Probability of retaining relevant and irrelevant variables for the unit-root case for a sample size of 1000

for $t = 1, \dots, T$ with x_t given in equation (30). Again we set $\sigma_u^2 = \sigma_v^2 = 1$, $\beta = 0.5$ for $T = 100$ and $\beta = 0.15811$ for $T = 1000$ and $M = 10,000$. The GUM is given by:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 \Delta x_t + \alpha_3 x_{t-1} + \alpha_4 \Delta x_t^2 + \alpha_5 x_{t-1}^2 + \epsilon_t, \quad \epsilon_t \sim \text{IN}[0, \sigma_\epsilon^2], \quad (33)$$

where $\text{cov}[\Delta x_t, x_{t-1}] = \text{cov}[\Delta x_t^2, x_{t-1}^2] = 0$. We consider the case in which $x_0 = 10$ and we use rule (31) to de-mean the data.

Results Figure 7 records the retention probabilities of the unit-root process for $T = 100$. The first column is in levels, where the DGP is given in equation (32) but the GUM is given by equation (28). PcGets struggles to identify the DGP variable and the retention probability is substantially larger than 5% and 1% for the irrelevant variables for the two strategies. The second column uses the GUM in which the lagged variables are orthogonalized, given in equation (33), and the retention probabilities are analogous to the level's model. Thus, orthogonalizing the lagged variables alone without ensuring orthogonality of the contemporaneous variables yields no improvement in selection. The final column removes sample means from the data and orthogonalizes by differencing and this results in a retention probability of near 1 for the relevant variable (x_t) and a retention probability of approximately 5% and 1% for the irrelevant variables for the liberal and conservative strategies respectively. Hence, all forms of collinearity need to be removed when undertaking model selection.

Figure 8 provides more detail to figure 7, recording the retention probabilities of all GUM variables for the three cases outlined. Retention of the lagged dependent variable has the appropriate probabilities, indicating that LDVs are not problematic for model selection when they do not enter the DGP. In levels, x_t is retained approximately 3 times too often at 15% and 4% respectively. Taking differences halves the retention probabilities to 7% and 2% but de-meaning as well as taking differences reduces the retention probabilities further to 5% and 1%. The same pattern is evident with x_t^2 . The DGP variable is retained

about 70% of the time for the levels GUM, with the conservative strategy having a higher power than the liberal strategy. Taking differences does not alter this probability, but de-meaning does increase the probability of retaining the DGP variable to unity. Both the lagged squared term and the intercept are retained far too often at about 30% which is analogous to the above results. Orthogonalizing by differencing does not alter these probabilities as the collinearity between x_{t-1} and x_{t-1}^2 is unaffected but de-meaning reduces the retention probability to appropriate levels.

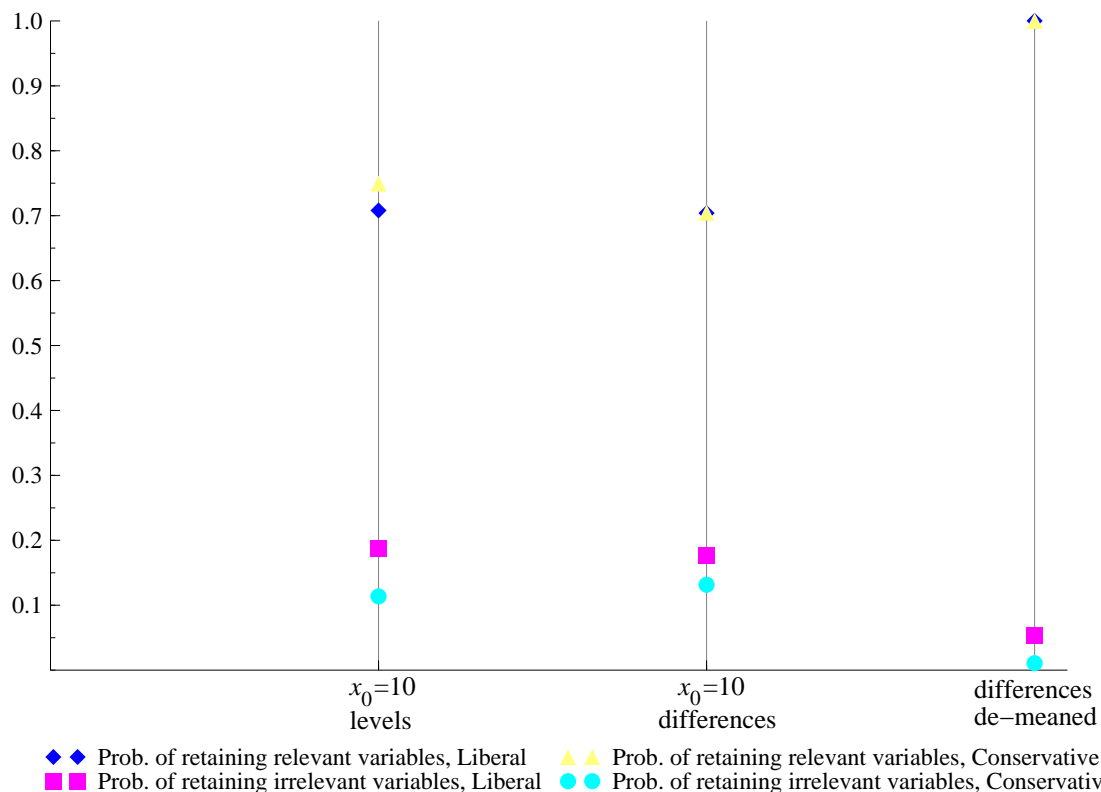


Figure 7: Retention probabilities for the relevant and irrelevant variables for the unit root case with a differenced GUM for a sample size of 100.

3 Non-normality

Normality is a basic assumption in PcGets. A test for normality based on Doornik and Hansen (1994) is performed on the GUM, and the diagnostics are checked at every subsequent reduction stage. If a reduction brings about a rejection of a diagnostic test, the search is terminated at the preceding level. When we consider non-linear models, normality becomes an essential requirement. Problems arise when extreme observations result in fat-tailed distributions. There is an increased probability that non-linear functions will align with extreme observations, effectively acting as indicators and therefore being retained too often. This can be demonstrated by considering a simple case in which an outlier is modelled as an indicator variable, $I_{\{t=s\}}$, which takes the value 1 in period s and 0 otherwise. Consider a regression between 2 unconnected variables:

$$y_t = \beta x_t + \delta I_{\{t=s\}} + u_t \quad (34)$$

$$x_t = \gamma I_{\{t=s\}} + v_t \quad (35)$$

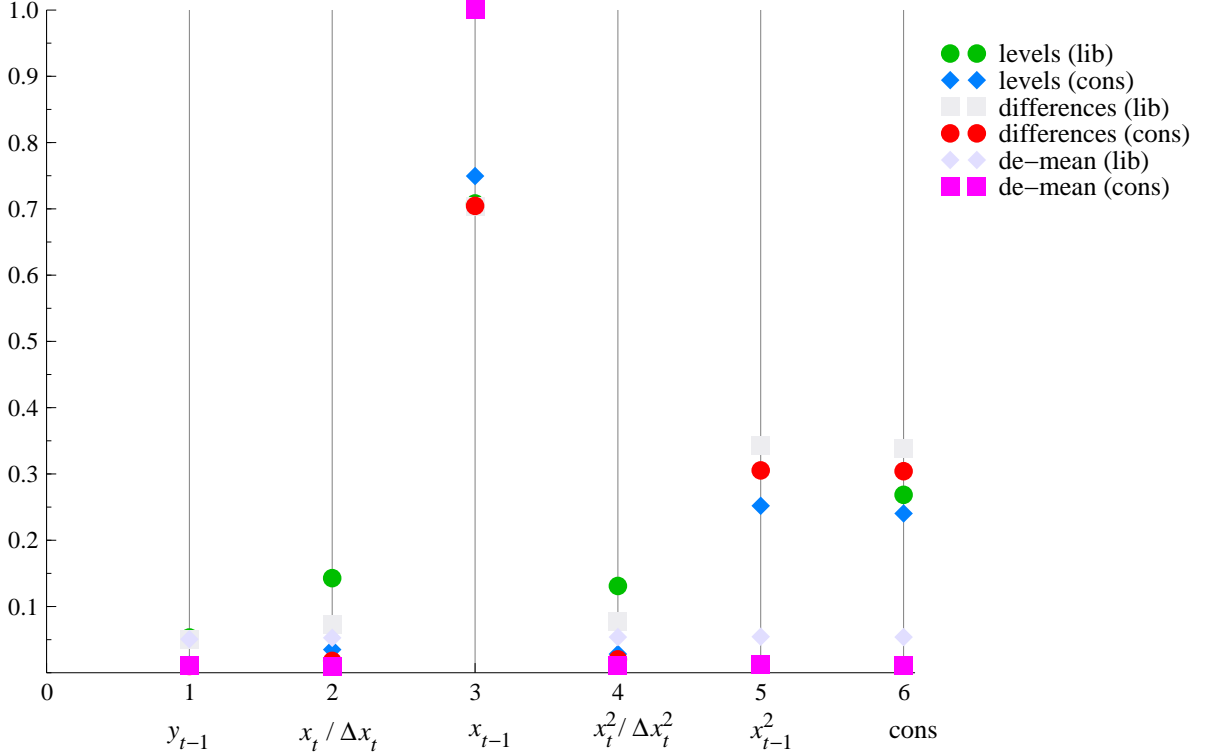


Figure 8: Retention probabilities of the GUM variables for the unit root case with $T=100$

where $\beta = 0$. We can calculate the coefficient $\hat{\beta}$ as:

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\delta\gamma \sum I_{\{t=s\}}^2 + \sum (\delta v_t + \gamma u_t) I_{\{t=s\}} + \sum v_t u_t}{\gamma^2 \sum I_{\{t=s\}}^2 + 2\gamma \sum v_t I_{\{t=s\}} + \sum v_t^2} = \frac{\delta\gamma + (\delta v_s + \gamma u_s) + \sum v_t u_t}{\gamma^2 + 2\gamma v_s + \sum v_t^2} \quad (36)$$

as $\sum I_{\{t=s\}}^2 = 1$. Also:

$$\text{V}[\hat{\beta}] = \frac{\hat{\sigma}_u^2}{\sum x_t^2} \quad (37)$$

and:

$$t_{\hat{\beta}} = \frac{\hat{\beta} \sqrt{\sum x_t^2}}{\hat{\sigma}_u} = \frac{\sum x_t y_t}{\hat{\sigma}_u \sqrt{\sum x_t^2}}. \quad (38)$$

Hence, if we approximate by $v_s = u_s = 0$, $\hat{\sigma}_u = 1$, $\gamma/\hat{\sigma}_v = 1$ and $\sum v_t u_t \simeq 0$ we can calculate:

$$t_{\hat{\beta}}^2 = \frac{\delta^2 \gamma^2}{\gamma^2 + T}. \quad (39)$$

To illustrate this phenomenon suppose $\delta = 6$, $\gamma = 5$, and $T = 100$. Then:

$$t_{\hat{\beta}}^2 = \frac{6^2 \times 5^2}{5^2 + 100} = 7.2. \quad (40)$$

Thus, outliers need to be quite large for this effect. This is entirely plausible when considering non-linear transformations. For example, in one draw of an $\text{IN}[0, 1]$ process with $T = 100$, the standard deviation of the inverse transformation is $\sigma = 25.608$ and the largest outlier is -169.2 .

Non-existence of moments is a concern. Non-linear transformations such as the inverse or squared inverse can explode with a zero-mean process. Observe that in Monte Carlo experiments, the existence

of moments can be crucial. By increasing the number of replications, the probability that a draw will take a value very near zero is increased. For a small number of replications, the probability that a draw is zero is negligible but this increases with M . Hence, there is a dichotomy between standard Monte Carlo theory which says increase the number of replications to determine asymptotic results and the problem of increasing the probability of a zero draw: see Sargan (1982).

3.1 Extreme observations: Monte Carlo example

Monte Carlo evidence illustrates the outlier problem. Consider a DGP given by:

$$x_{i,t} = \nu_{i,t} \quad \nu_{i,t} \sim \text{IN}[0, 1] \quad \text{for } i = 1, \dots, 4. \quad (41)$$

We shall generate non-linear functions given by the inverses of these normal distributions:

$$x_{i,t}^{-1} = \frac{1}{x_{i,t}}. \quad (42)$$

The GUM contains 20 irrelevant variables given by:

$$x_{i,t}^{-1} = \alpha_0 + \sum_{k=1}^4 \alpha_{i,t-k} x_{i,t-k}^{-1} + \sum_{j=1}^4 \sum_{m=0}^4 \alpha_{j,t-m} x_{j,t-m}^{-1} + \epsilon_t. \quad (43)$$

Equation (43) led to $|t|$ -values as large as 19 for variables with zero non-centralities. The variable would unequivocally, but incorrectly, be retained as a DGP variable. On average, two of the 20 irrelevant regressors were retained at the 1% significance level. This implies that a fat-tailed distribution would have a ‘size’ of 10% at the 1% significance level. If the dependent variable was $x_{i,t}$ rather than $x_{i,t}^{-1}$, the retention probabilities are correct. Non-normal errors can also pose a similar problem. Hence, the problem of model selection when there are extreme observations is exacerbated by the inclusion of non-linear functions such as inverses.

3.2 Solution: Indicator saturation techniques

To overcome the problem of fat tails, we draw on a technique proposed by Hendry *et al.* (2004) in which the data is saturated with as many indicators as observations and the indicators are selected at a chosen significance level to identify outliers. Once we have identified outliers, we can remove them and the selection process will not be biased in favour of non-linear functions that are proxying indicators for the outliers.

The key to the saturation with indicator variables technique lies in the fact that PcGets can handle more variables than observations. When $n > T$, PcGets uses subsets of $n_1 \leq T/2$ variables as the initial GUMs and a joint model is formulated from the terminal models of these subsets of variables, from which the standard PcGets procedure is applied. To add T indicators $I_t = 1_{\{t=t_i\}}$, a regression of all indicators on the dependent variable, y_t , would result in a perfect fit. Instead, (say) half the indicators are included in the GUM and the two resulting terminal models are stored.⁴ The joint model is formulated from these two terminal models and PcGets re-selects the indicators. Under the null that there are no outliers, αT indicators will be retained on average for a significance level α .

This technique also overcomes the problem of undetectable outliers. One concern with non-linearity is that it is difficult to distinguish between extreme observations that are outliers and extreme observations that are due to the non-linearity in the data. Methods that remove extreme observations could be in danger

⁴An intercept is also included in the GUMs.

of removing the underlying non-linearity that should be modelled. Indicator saturation techniques can avoid this problem by including all potentially relevant variables as well as indicators for all observations in the initial GUM. If there are n potentially relevant variables including non-linear transformations and an intercept, and T indicators, then $n + T \gg T$. Using the technique in which the variables are divided into subsets, n_j for $j = 1, \dots, J$ and T_k for $k = 1, \dots, K$, we can formulate JK GUMs in which all cross pairings of $n_j + T_k$ ($\forall j, k$) are included. The union of the resulting terminal models is formulated and PcGets re-selects. The process can be performed iteratively if the union after the first reduction stage is still larger than T . By removing the extreme observations in conjunction with selecting the non-linear functions we avoid the problem of removing observations that generate the non-linearity.

For examples of more variables than observations techniques and indicator saturation techniques see Hendry and Krolzig (2004) for an application to growth regressions and Castle (2005) for an application to telephone services demand.

4 Non-linear functions

Econometric modelling of non-linear processes presents many problems over and above those encountered when developing linear econometric models. Identifying a unique non-linear representation of an economic process can be formidable given the complexity of possible local data generating processes (LDGPs). As there are a substantial number of potential functional forms that the DGP may take, specifying a GUM that nests the unknown DGP is problematic. In the non-linear world we are not only concerned with specifying the potentially relevant variables but we also need to consider the nature of the non-linearity.

The methodology suggested by the *Gets* framework would be to explore, conditional on theory, sufficiently general models to nest a class of DGPs. If the model does not match with the conjectured DGP, the non-linear GUM is revised to consider another class of non-linear models and the procedure continues iteratively. As the selected non-linear model should nest its linear counterpart, the reduction to a linear model can be tested. Hence, there is no loss of information by commencing with a more general non-linear model and testing downwards, in keeping with the *Gets* philosophy.

We focus on polynomials as they provide a good local approximation for a wide range of non-linear models. One such class of models are smooth transition regression models. Other advantages of the polynomial class are that simple operational rules for orthogonalizing can be applied and the class will retain linearity in the parameters. PcGets uses standard OLS or IV estimation, and we aim to incorporate a non-linear capability within this framework. In principle, a general non-linear, likelihood based, system approach is feasible, but no software yet exists to implement such an approach. However, models that are non-linear in the parameters are just restrictions imposed on the parameters, and the gains achieved from these complex models are often limited compared to the costs of estimating such models within a general-to-specific framework.

Other potential approximations that were considered include orthogonal series such as Hermite polynomials but these perform poorly in the tails, Fourier series but these require a large number of terms to obtain a close approximation and asymptotic series but these tend to be intractable. Confluent hypergeometric functions provide a very general functional form that can capture a wide range of non-linearity, see Abadir (1999), and this warrants further investigation.

It is important that the expansion can capture the non-linearity with a small number of terms. One concern with model selection for non-linear models is that if there are many potential non-linear functions the number of variables in the GUM will be large which could result in excess adventitious retention of irrelevant variables. This problem is addressed in section 4.3.

4.1 Polynomial series

If the functional form of the non-linear DGP is unknown we can postulate a general model given by:

$$y_t = f(\mathbf{W}_t, \mathbf{W}_{t-1}, \dots, \mathbf{W}_{t-q}) + v_t \quad (44)$$

where \mathbf{W}_t is a distinct vector of n variables. A Taylor series expansion around 0 will result in the dual of the Volterra series, see Priestley (1981), given by:

$$\begin{aligned} \psi(\mathbf{W}_t, \dots, \mathbf{W}_{t-q}; \theta) = & \psi_0 + \sum_{s=0}^q \sum_{i=1}^n \psi_{1,is} w_{i,t-s} + \sum_{r=0}^q \sum_{s=0}^q \sum_{i=1}^n \sum_{j=1}^i \psi_{2,ijsr} w_{i,t-s} w_{j,t-r} \\ & + \sum_{p=0}^q \sum_{r=0}^q \sum_{s=0}^q \sum_{i=1}^n \sum_{j=1}^i \sum_{k=1}^j \psi_{3,ijk srp} w_{i,t-s} w_{j,t-r} w_{k,t-p} + \dots \end{aligned} \quad (45)$$

This motivates the use of polynomial functions, although equation (45) shows how quickly the number of parameters will increase. Hence, consideration of the relevant polynomial functions and strategies for selection when the GUM is large is needed.⁵ Polynomials provide a good local approximation because we can take a Taylor expansion around any postulated function.

4.2 Application: Approximation to a STR model

Smooth transition regression (STR) models are a form of regime-switching model developed by Maddala (1977), Granger and Teräsvirta (1993), Teräsvirta (1994). Restrictions on the STR model result in various regime switching models including smooth-transition autoregression models (STAR), see Chan and Tong (1986), Luukkonen, Saikkonen and Teräsvirta (1988); threshold autoregression models (TAR), see Tong (1990); switching regression models, see Quandt (1983); and exponential autoregression models (EAR), see Priestley (1981). The aim of this section is to see how well we can approximate a logistic STR model with a polynomial approximation.

The STR model is given by:

$$y_t = \beta' \mathbf{X}_t + (\theta' \mathbf{X}_t) G(\gamma, c, s_t) + u_t, \quad u_t \sim \text{IN}[0, \sigma_u^2] \quad (46)$$

for $t = 1, \dots, T$ where $G(\cdot)$ is the transition function. Various distributional assumptions can be made on the transition function and we investigate the properties of the logistic STR(1) given by:

$$G(\gamma, c, s_t) = \left[1 + \exp \left\{ -\gamma \left(\frac{s_t - c}{\hat{\sigma}_s} \right) \right\} \right]^{-1}. \quad (47)$$

In this monotonic transition function, γ determines how rapid the transition is from 0 to 1 as a function of the transition variable, s_t , and c determines where the transition occurs. As $\gamma \rightarrow \infty$ the model becomes a 2 regime-switching regression model and $\gamma > 0$ is the identifying restriction. Estimation of γ is particularly difficult as the likelihood function is not well behaved. An upper bound on $\hat{\gamma}$ of approximately 5 can be deduced from Chebyshev's inequality. For $\hat{\gamma} \geq 5$ the transition function acts as 2 regime-switching process and a simplification to a switching regression model can be made. For values of $\hat{\gamma}$ close to 0, the increased uncertainty regarding the regime increases the uncertainty of the $\hat{\theta}$ parameter estimates and consequently, the estimates of $\hat{\beta}$ that correspond to the variables for $\hat{\theta}$.

⁵Polynomial functions are commonly used in economics and are useful because of Weierstrass's approximation theorem which states that any continuous function on a closed and bounded interval can be approximated by polynomials, i.e. if $x \in [a, b]$, for any $\epsilon > 0$ there exists a polynomial $p(x) \in [a, b]$ such that $|f(x) - P(x)| < \epsilon \forall x \in [a, b]$.

The model can be approximated by replacing the logistic transition function with a 3rd order Taylor expansion:⁶

$$y_t \simeq \beta' \mathbf{X}_t + (\theta' \mathbf{X}_t) \left[\frac{1}{2} + \frac{z_t}{4} - \frac{z_t^3}{48} \right] + v_t, \quad v_t \sim \text{IN} [0, \sigma_v^2] \quad (48)$$

where:

$$z_t = \gamma \left(\frac{s_t - c}{\hat{\sigma}_s} \right). \quad (49)$$

This approximation results in a linearized model given by:

$$y_t \simeq \alpha'_1 \mathbf{X}_t + \alpha'_2 \mathbf{X}_t s_t + \alpha'_3 \mathbf{X}_t s_t^2 + \alpha'_4 \mathbf{X}_t s_t^3 + v_t, \quad v_t \sim \text{IN} [0, \sigma_v^2] \quad (50)$$

which can be estimated in PcGets.⁷ In practice, we would wish to start with a more general approximation.

4.2.1 Selection

The DGP will have 2 unknown components, the functional form and the relevant variables. The polynomial approximation overcomes the former and general-to-specific modelling solves the latter. We need to formulate a sufficiently general unrestricted model (GUM) that will include all potentially relevant variables and transition variables for all possible lag lengths. The general linearized GUM based on the LSTR(1) model, for a set of n potential regressors, \mathbf{W}_t , (including an intercept) and m potential transition variables, \mathbf{S}_t , is given by:

$$\begin{aligned} y_t = & \sum_{i=1}^n \alpha_i W_{i,t} + \sum_{i=1}^n \sum_{j=1}^m \delta_{ij} W_{i,t} S_{j,t} + \sum_{i=1}^n \sum_{j=1}^m \lambda_{ij} W_{i,t} S_{j,t}^2 \\ & + \sum_{i=1}^n \sum_{j=1}^m \phi_{ij} W_{i,t} S_{j,t}^3 + \epsilon_t, \quad \epsilon_t \sim \text{IN} [0, \sigma_\epsilon^2] \end{aligned} \quad (51)$$

The number of variables in the GUM is $n + 3nm$. The set of potential regressors will include variables that enter in either the linear or non-linear multiplicative function or both. If a regressor only enters 1 of the components, the parameter corresponding to the other component will be 0. For variables that enter both components (including a possible intercept), the Taylor approximation will give a parameter estimate that combines both components.

⁶Observe that $\frac{\partial^2 G(z)}{\partial z^2} \Big|_{z=0} = 0$ where $G(z) = [1 + e^{-z}]^{-1}$ and so the z_t^2 term drops out of the Taylor expansion. There is still a quadratic component in s_t as z_t^3 is included, where $z = \gamma \left(\frac{s_t - c}{\hat{\sigma}_s} \right)$.

⁷The transition variable is scaled by $\hat{\sigma}_s$. When estimating the polynomial approximation, we can pull the scale factor into the coefficient estimates. Assuming \mathbf{X}_t is a scalar for tractability, the mappings from the coefficients in equation (46) to equation (50) are given by:

$$\begin{aligned} \alpha_1 &= \beta + \frac{\theta}{2} - \frac{\theta\gamma c}{4\hat{\sigma}_s} + \frac{\theta\gamma^3 c^3}{48\hat{\sigma}_s^3} \\ \alpha_2 &= \frac{\theta\gamma}{4\hat{\sigma}_s} - \frac{3\theta\gamma^3 c^2}{48\hat{\sigma}_s^3} \\ \alpha_3 &= \frac{3\theta\gamma^3 c}{48\hat{\sigma}_s^3} \\ \alpha_4 &= -\frac{\theta\gamma^3}{48\hat{\sigma}_s^3}. \end{aligned}$$

A direct test of linearity is whether the PcGets selection results in the non-linear functions being retained:

$$H_0 : \boldsymbol{\delta} = \boldsymbol{\lambda} = \boldsymbol{\phi} = \mathbf{0}. \quad (52)$$

The model selected by PcGets should capture the non-linearity inherent in an LSTR model whilst enabling a much more general specification to be tested. The approximation to the LSTR model can be tested by estimating the corresponding LSTR model to the specific model and then augmenting the specific model selected by PcGets with the non-linear component of the LSTR model. Suppose k relevant variables were retained ($k \leq n$) and 1 transition variable was selected given by s_1 , then the test of the approximation to the LSTR model would be:

$$H_0 : \boldsymbol{\kappa} = \boldsymbol{\mu} = \boldsymbol{\psi} = \mathbf{0}, \quad (53)$$

for the regression:

$$y_t = \sum_{i=1}^k \tau_i W_{i,t} + \sum_{i=1}^k \kappa_i W_{i,t} s_{1,t} + \sum_{i=1}^k \mu_i W_{i,t} s_{1,t}^2 + \sum_{i=1}^k \psi_i W_{i,t} s_{1,t}^3 + \sum_{i=1}^k (\theta_i W_{i,t}) \left[1 + \exp \left\{ -\gamma \left(\frac{s_t - c}{\hat{\sigma}_s} \right) \right\} \right]^{-1} + \eta_t. \quad (54)$$

4.2.2 Monte Carlo results

To examine the ability of the Taylor expansion to approximate an LSTR model, we undertake a simple Monte Carlo experiment based on Granger and Teräsvirta (1993), ch.7. The DGP is given by:

$$y_t = 1 + 2x_t + x_{t-1} + 0.5x_{t-1} - (2x_t + x_{t-1} + 0.5x_{t-2}) [1 + \exp \{-4(x_{t-1} + 3)\}]^{-1} + u_t \quad (55)$$

where the data is generated as a stationary AR(1) process:

$$x_t = \alpha x_{t-1} + v_t, \quad v_t \sim \text{IN} [0, \sigma_v^2]. \quad (56)$$

$\alpha = 0, 0.8$ and $\text{var}[x_t] = 2.78$. Hence, $\sigma_v^2 = 1$ if $\alpha = 0.8$. $u_t \sim \text{IN} [0, \sigma_u^2]$ in which we set $\sigma_u^2 = 0.0625, 0.25, 1$ and $\text{cov}(u_t, v_s) = 0 \forall t, s$. $T = 100$ and the first 100 observations were discarded to avoid initialization effects. 1000 replications were undertaken. The models estimated include the LSTR model, the linear model and the polynomial approximation, given in equations (57), (58) and (59). Also, selection using the PcGets liberal and conservative strategies was conducted on equation (59).

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} - (\beta_4 x_t + \beta_5 x_{t-1} + \beta_6 x_{t-2}) [1 + \exp \{-\gamma(x_{t-1} - c)\}]^{-1} + \epsilon_t, \quad (57)$$

$$y_t = \delta_0 + \delta_1 x_t + \delta_2 x_{t-1} + \delta_3 x_{t-2} + \eta_t, \quad (58)$$

$$y_t = \phi_0 + \phi_1 x_t + \phi_2 x_{t-1} + \phi_3 x_{t-2} + \phi_4 x_t x_{t-1} + \phi_5 x_t x_{t-1}^2 + \phi_6 x_t x_{t-1}^3 + \phi_7 x_{t-1}^2 + \phi_8 x_{t-1}^3 + \phi_9 x_{t-1}^4 + \phi_{10} x_{t-2} x_{t-1} + \phi_{11} x_{t-2} x_{t-1}^2 + \phi_{12} x_{t-2} x_{t-1}^3 + \epsilon_t. \quad (59)$$

Table 1 records the equation standard errors (with standard deviations of the errors in parentheses) for the models outlined above. The LSTR model is an excellent fit, with the equation standard error equal to the true error in all cases. This is because the functional form of equation (57) was known in the experiments and so the only error comes through estimation uncertainty. In practice, it is unlikely that the exact specification of the DGP is known, but it provides a good baseline.

σ_u^2	LSTR(1)	Linear	Polynomial	Liberal	Conservative
$\alpha = 0.8$					
0.0625	0.251 (0.018)	1.261 (0.784)	0.414 (0.186)	0.451 (0.204)	0.457 (0.209)
0.25	0.506 (0.067)	1.365 (0.725)	0.609 (0.147)	0.625 (0.209)	0.634 (0.217)
1	0.995 (0.074)	1.660 (0.613)	1.060 (0.120)	1.066 (0.267)	1.083 (0.277)
$\alpha = 0$					
0.0625	0.250 (0.018)	0.734 (0.281)	0.310 (0.053)	0.318 (0.037)	0.322 (0.039)
0.25	0.501 (0.036)	0.864 (0.245)	0.533 (0.050)	0.536 (0.054)	0.543 (0.056)
1	1.01 (0.084)	1.234 (0.194)	1.015 (0.079)	1.015 (0.158)	1.033 (0.166)

Table 1: Equation standard errors for the models approximating an LSTR(1) DGP.

The linear model is poor approximation in all cases. The polynomial approximation performs extremely well when σ_u^2 is large, with an equation standard error of just 6% and 2% larger than the true DGP for the $\alpha = 0.8$ and $\alpha = 0$ cases respectively. The approximation is much poorer when the error variance is small, with an equation standard error that is more than 60% larger than the true DGP for $\sigma_u^2 = 0.0625$ and $\alpha = 0.8$. The residual is a composite of the squared approximation error and the DGP shock, so as the latter falls, the former dominates. The squared approximation error is approximately 0.1 for $\alpha = 0.8$, and 0.04 for $\alpha = 0$. This would be large empirically in a log model, although is dependent on the scaling of σ . Undertaking selection on the polynomial DGP slightly increases the equation standard error but it delivers a more parsimonious model. There is very little cost to selection and the non-deletion probabilities for PcGets are close to the theoretical upper bounds. A further extension would be to see how well PcGets performs when commencing from a more general model.

The ability of the polynomial class of functions to approximate non-linear models will clearly depend on the specific data generating process that is being modelled. However, for DGPs in which there is more uncertainty, the polynomial model performs extremely well. Further work is needed to see how well the polynomial expansion can approximate other classes of non-linear models, but it seems feasible that for a certain range of models, the polynomial class will perform well.

4.3 Super-conservative strategy

Hendry and Krolzig (2003) consider the deletion probabilities of PcGets. If there are many variables in the DGP, but few are relevant, one may expect the search costs to be high. In a pure t-testing strategy at significance level α , the probability distribution of one or more null coefficients being significant is given by the $k + 1$ terms of the binomial expansion of $1 = (\alpha + (1 - \alpha))^k$ and hence the average number of irrelevant variables retained is $k\alpha$ where k is the number of irrelevant variables in the GUM. If k is large due to the inclusion of many non-linear terms, a greater number of irrelevant terms will be retained. For example, if a linear GUM contains 20 irrelevant variables and $\alpha = 0.05$, one variable will be retained on average. If non-linear functions are included to test for the presence of non-linearity, perhaps 100 extra irrelevant functions will enter the GUM. This increases the average number of irrelevant variables retained to 6. Furthermore, irrelevant non-linear functions are likely to be detrimental to both modelling

and forecasting. Non-linear functions should only be retained if there is definite evidence of non-linearity, because these models are much less robust than linear models, both to changes in collinearity between regressors and location shifts within the equation or in any retained but irrelevant variable.

Given the preference for linear models unless strong evidence for non-linearity is presented and the possible excess retention of irrelevant functions due to the number of non-linear functions tested, we propose a ‘super-conservative’ strategy for PcGets. This strategy would use more stringent critical values for the non-linear functions compared to the linear functions which would all be tested within the same procedure. Hence, diagnostic tests would apply to the full GUM, but pre-search tests and multi-path search tests would be conducted at more stringent critical values for the non-linear functions. The critical value would depend on the number of functions included in the model, but research examining RETINA, another automatic model selection algorithm developed by Perez-Amaral, Gallo and White (2003, 2005), would suggest that non-linear functions should only be retained if approximately $|t| > 5$. As with all significance levels, the choice will depend on the preferences of the econometrician.

Block F-tests on classes of non-linear functions could be incorporated into the pre-search stage in PcGets. Tight significance levels would again be used and a sequential testing procedure on classes of non-linear functions entering the GUM would be undertaken until just those classes that are significant are retained to formulate the GUM. This would narrow down the number of non-linear functions in the multi-path search stage.

5 Conclusion

This paper addresses issues associated with model selection in which there is non-linearity, providing solutions to three potential difficulties with the inclusion of a non-linear capability into PcGets. Collinearity is one of the most fundamental problems with model selection. We show the correlation magnitudes between linear and non-linear functions can be extremely high, causing selection algorithms to struggle to identify the relevant variables. This is due to non-zero means in the data. We propose a solution of de-meaning both the linear term prior to the transformation and after the non-linear function has been generated. This removes the collinearity, and PcGets is shown to have good selection properties with this orthogonalization. We emphasize the importance of normality for model selection, and show that the ‘size’ can be greatly increased relative to the significance level if non-linear functions capture extreme observations. The solution of indicator saturation is proposed. Finally we look at approximating non-linear data with polynomials. A polynomial approximation is shown to capture the non-linearity generated by an LSTR model well over certain specifications. The problem of excess retention of irrelevant variables is addressed, and a super-conservative strategy is proposed.

The paper has gone some way to demonstrating the feasibility of a non-linear strategy in PcGets. Further work is needed to assess the ability of the polynomial approximation to capture various forms of non-linearity in the data. Also, other classes of approximating functions such as hypergeometric functions need to be considered in more detail. Nevertheless, this paper has demonstrated that PcGets is applicable to a broader range of models than previously imagined, in keeping with the general-to-specific philosophy of the program. Numerous applications are now feasible in which non-linearity is suspected, including applications such as stock returns, Phillip’s curves, unemployment and business cycles, and production functions, all of which warrant investigation.

References

- Abadir, K. M. (1999). An Introduction to Hypergeometric Functions for Economists. *Econometric Reviews*, **18**, 287–330.

- Castle, J. (2005). Evaluating PcGets and RETINA as automatic model selection algorithms. *Oxford Bulletin of Economics and Statistics*, forthcoming.
- Chan, K. S., and Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, **7**, 179–194.
- Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.
- Frisch, R. (1934). *Statistical Confluence Analysis by means of Complete Regression Systems*. Oslo: University Institute of Economics.
- Granger, C. W. J., and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
- Hendry, D. F., Johansen, S., and Santos, C. (2004). Selecting a regression saturated by indicators. Unpublished paper, Economics Department, University of Oxford.
- Hendry, D. F., and Krolzig, H.-M. (1999). Improving on ‘Data mining reconsidered’ by K.D. Hoover and S.J. Perez. *Econometrics Journal*, **2**, 202–219.
- Hendry, D. F., and Krolzig, H.-M. (2001). *Automatic Econometric Model Selection*. London: Timberlake Consultants Press.
- Hendry, D. F., and Krolzig, H.-M. (2003). New developments in automatic general-to-specific modelling. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*, pp. 379–419. Princeton: Princeton University Press.
- Hendry, D. F., and Krolzig, H. M. (2004). We ran one regression. *Oxford Bulletin of Economics and Statistics*, **66**, 799–810.
- Hendry, D. F., and Krolzig, H.-M. (2005). The properties of automatic Gets modelling. *Economic Journal*, **115**, C32–C61.
- Hendry, D. F., and Morgan, M. S. (1989). A re-analysis of confluence analysis. *Oxford Economic Papers*, **41**, 35–52.
- Luukkonen, R., Saikkonen, P., and Teräsvirta, T. (1988). Testing linearity in univariate time series models. *Scandinavian Journal of Statistics*, **15**, 161–175.
- Maddala, G. S. (1977). *Econometrics*. New York: McGraw-Hill.
- Perez-Amaral, T., Gallo, G. M., and White, H. (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics*, **65**, 821–838.
- Perez-Amaral, T., Gallo, G. M., and White, H. (2005). A comparison of complementary automatic modelling methods: RETINA and PcGets. *Econometric Theory*, **21**, 262–277.
- Priestley, M. B. (1981). *Spectral Analysis and Time Series*. New York: Academic Press.
- Quandt, R. E. (1983). Computational problems and methods.. pp. 699–746. Amsterdam: North Holland.
- Sargan, J. D. (1982). On Monte Carlo estimates of moments that are infinite. In Basman, R. L., and Rhodes, G. F. (eds.), *Advances in Econometrics: A Research Annual*, Vol. 1, pp. 267–299. Greenwich, Connecticut: Jai Press Inc.
- Teräsvirta, T. (1994). Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**, 208–218.
- Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.