Sub-sample Model Selection Procedures in *Gets* Modelling

David F. Hendry and Hans-Martin Krolzig*
Economics Department, Oxford University.

April 16, 2003

Abstract

When the DGP is nested in the model, *PcGets* delivers high performance selection across different (unknown) states of nature. One of its steps involves sub-sample post-selection assessment, and here we consider its properties and investigate its practical application. The simulation results show that conditional on retaining a variable, sub-sample information cannot discriminate between substantive and adventitious significance. The Monte Carlo experiments also reveal that the sub-sample selection method suggested by Hoover and Perez (1999) is dominated by procedures selecting only on full-sample evidence, when both approaches are evaluated at a given size. Nevertheless, although the sub-sample procedures do not result in a genuinely beneficial trade-off between size and power, they are particularly successful in controlling the size for selection problems that were previously deemed almost intractable.

Contents

*Prepared for ESAM, Brisbane, 2002. Financial support from the U.K. Economic and Social Research Council under grant L11625015 is gratefully acknowledged. We are indebted to Julia Campos for helpful comments on an earlier draft.

1 Introduction

Despite the many difficulties intrinsic to model selection, viewed as searching for an unknown specification in a large class of models, recent automatic procedures have achieved high success rates in locating the data generation process (DGP) across a variety of simulation experiments such as Hoover and Perez (1999, 2000), Hendry and Krolzig (1999, 2002), Krolzig and Hendry (2001), Krolzig (2001, 2003), and Brüggemann, Krolzig and Lütkepohl (2002). Here we consider one of the selection strategies embodied in *PcGets*, namely its 'sub-sample significance evaluation' procedure. After a final model has been selected by the search process, its behaviour in overlapping sub-samples is evaluated, as a reliability check on the selected model.[1]

Hendry and Krolzig (2002) distinguish between the costs of inference, which are an inevitable consequence of non-zero significance levels and non-unit powers and apply even when the DGP is known, and the costs of search, which are additional to those faced when commencing from a model that is the DGP. In summarizing the Monte Carlo evidence on the performance of *PcGets* in a range of experiments, including those used to calibrate its settings, Hendry and Krolzig (2003) show that *PcGets* performs well — in the sense that the costs of search are low — but naturally varies across the (unknown) states of nature. Their simulation evidence also shows that the sub-sample assessment procedure substantially lowers the 'size' of the selection algorithm, defined as the average incorrect retention rate of irrelevant variables, with a small reduction in power. However, Lynch and Vital-Ahuja (1998) show that 'selecting variables that are significant on all three splits (the two sub-samples and overall)' delivers no gain over simply using a smaller nominal size. The Lynch and Vital-Ahuja (1998) argument applies to Hoover and Perez (1999, 2000) who retain variables at the selection stage only if they are significant in two overlapping sub-samples.

However, those approaches differ at first sight from selecting only on full-sample evidence, followed by evaluation on sub-samples, which is the *PcGets* approach investigated here. Nevertheless, the simulation evidence alone does not establish the efficacy of sub-sample selection: as Lynch and Vital-Ahuja (1998) express the matter, the key issue is whether the power loss of the sub-sample 'significance evaluation' procedure is smaller — given the size reduction achieved — than that resulting from just setting a tighter initial significance level. Unfortunately, power depends on both the unknown state of nature (through a non-centrality parameter) and on the significance level set for the null, and varies in a highly non-linear manner as a function of these. For example, if the power were close to unity, little loss could occur for small changes in nominal significance levels (called size as a shorthand below), whereas for smaller values of the non-centrality parameter, a large reduction in power might ensue.

The structure of the paper is as follows. Section 2 describes the various sub-sample selection procedures in Hoover and Perez (1999) and *PcGets*. Section 3 investigates by simulation, the distributional properties and the implied power-size trade-off of the Hoover-Perez sub-sample selection method and the *PcGets* sub-sample reliability assessment; different states of nature and various choices of the percentage of overlap are considered. Section 4 concludes.

[1] *PcGets* by Hendry and Krolzig (2001) is an Ox Package (see Doornik, 2001) based on the theory of reduction (Hendry, 1995, Ch.9) implementing automatic general-to-specific (*Gets*) modelling for linear models.

2 The form of sub-sample selection procedures

2.1 Hoover–Perez

Hoover and Perez (1999, 2000) select variables only if they are significant in two over-lapping sub-samples. In the former paper, they graph the trade-off between size and power (defined as the average retention probability of relevant variables) as the percentage of overlap varies from 50% to 90%, and find that a split at about 70–80% performs best, in that the slope of the trade-off is steeper below and flatter above. At first sight, that evidence looks persuasive; but the non-linear relation between size and power for a t-test also would show a similar shape as the size varied from (say) 10% to 0.1% for values of the non-centrality parameter in the neighbourhood of 2. To see this, consider a normal random variable:

$$x \sim \mathsf{N}\left[\mu, \sigma^2\right], \tag{1}$$

where $\mu = 2$ and $\sigma^2 = 1$ so:

$$\mathsf{P}\left(x \geq c_\alpha\right) = \mathsf{P}\left(x - 2 \geq c_\alpha - 2\right) = \frac{1}{\sqrt{2\pi}} \int_{c_\alpha - 2}^{\infty} \exp\left(-\frac{1}{2}\left[x-2\right]^2\right) dx.$$

Figure 1a plots the resulting power-size trade-off. The dashed line shows the evident slope change around 5%, suggesting that the trade-off 'worsens' sharply as the size falls, but that is simply the correct power cost of a smaller size, which should be determined by the relative losses on type 1 versus type 2 errors, not by the slope — which is an intrinsic feature of the test. To reinforce that point, the solid line shows a three-way division, with an intermediate slope in the region of 5%. Thus, to be of benefit, a split-sample evaluation would need to lose less power per reduction in size than the inherent trade-off.

There is a separate such trade-off line for each value of μ in (1), and in figure 1b, the trade-off from Hoover and Perez (1999) is shown with the corresponding lines for $\mathsf{t} = 3$ and $\mathsf{t} = 3.3$, between which it lies. While it is difficult to judge the mean t-value in their simulation study, the evidence of a steeper fall to the left, and a shallower rise to the right does not by itself suggest gains.

2.2 The *PcGets* approach

After selection, the relevance of variables in the final model selected by *PcGets* is explored by post-selection reliability checks to ascertain whether 'significance' is substantive or adventitious. Post-selection evaluation is an attempt to mimic the role in an automatic procedure of recursive estimation, aiming to evaluate whether apparently significant effects are substantive or chance. It is not a check on constancy, which has already been tested for the GUM, and checked by diagnostics at each potential reduction.

Under the null hypothesis H_0, using a 2-sided test, a t-value will exceed (in absolute value) a critical value c_α on $\alpha\%$ of the occasions, where α is the significance level, so:

$$\mathsf{P}\left(-c_\alpha \leq \mathsf{t} \leq c_\alpha \mid \mathsf{H}_0\right) = \alpha. \tag{2}$$

However, after selecting a model, the retained variables will have significant t-values by construction.[2] The selected set of variables in the final model thus comprises (on average) $\alpha\%$ of the initial set — which are significant by chance — and the remainder — which are significant by having non-central t-distributions. The issue is whether conditional on observing full-sample significance, there is a division

[2] We neglect the small percentage of the time where retained variables enter insignificantly because their elimination would induce a significant diagnostic test value.

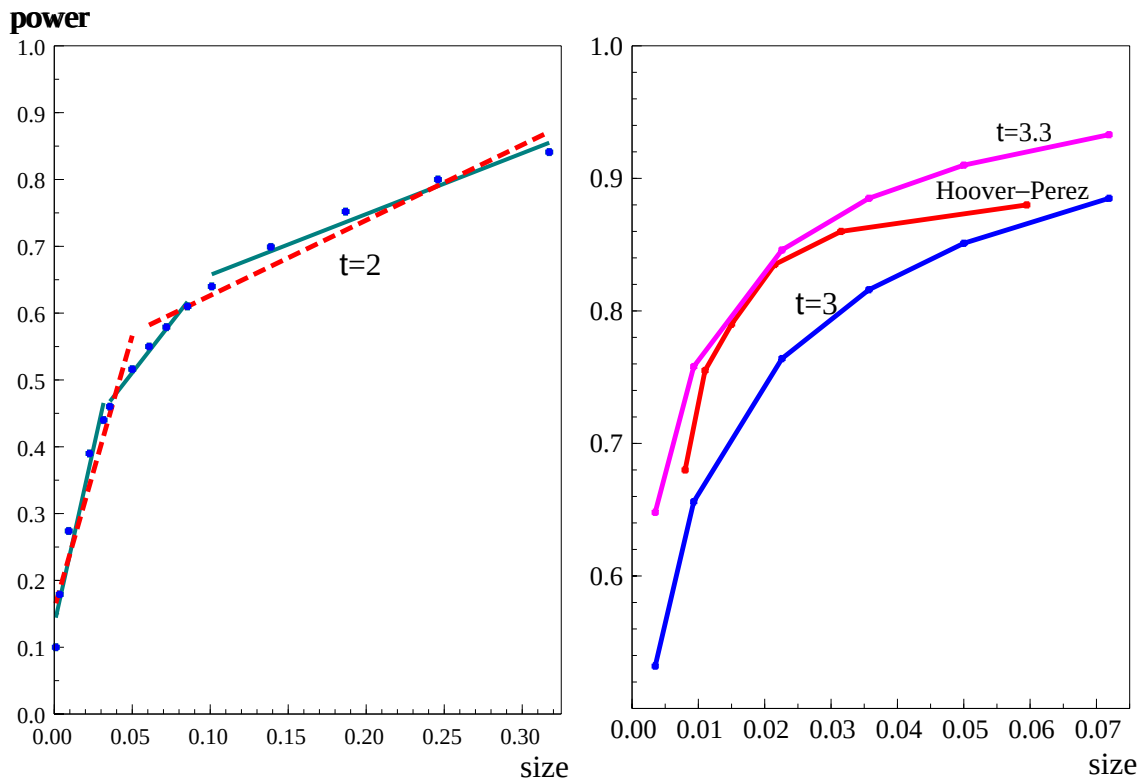

Figure 1 Power-size trade-off for a standard normal.

of the sample into sub-samples that would help discriminate between these, exploiting the fact that non-central t-values diverge, whereas central t-values are only significant by a chance value falling outside the range $[-c_\alpha, c_\alpha]$ at the end of the sample.

Our proposed filter between variables that really matter (non-central ts) and those that are adventitiously significant (central ts that happen to take large end-of-period values) is to check sub-sample reliability. The idea is that the central t-tests should be low in at least one of the two sub-periods, so revealing the actual irrelevance of the associated variable. However, because the sample sizes are smaller, less stringent critical values must be used to ensure a coherent inference procedure. *PcGets* centers on the Hoover and Perez (1999) split of 75–25 splits (so 50% of observations are in common), and adjusts the sub-sample nominal significance levels as a function of those selected for the full-sample selection.

It is clear from all the Monte Carlo studies that we have conducted that the reliability check reduces the size, and perhaps more importantly, has helped stabilize performance over different states of nature. Nevertheless, that by itself does not resolve the key issue of whether an equivalent size reduction achieved by lowering the initial significance level of every test would result in higher or lower power, and if so, how that changes across different DGPs. As noted above, the size-power trade-off is highly non-linear in both the significance level and the non-centrality parameters of the variables, and the analysis must be conditional on having retained each associated regressor at its observed t-value.

3 Assessing sub-sample-based model selection procedures

It is important to distinguish the reliability assessment of a model (which has been selected based on the full-sample information) from selection rules that are formulated in terms of sub-sample evidence. We now provide some Monte Carlo evidence indicating that the latter procedure is dominated by the former.

3.1 The curse of sub-samples

Both sub-sample-based selection rules rely on information from sub-sample t-tests, so it is useful to start by analyzing the properties of a simple sub-sample t-test (without conditioning on full-sample significance), and its relation to the full-sample t-test. Because of the difficulty induced by overlapping samples, in §3.1.1 we first consider when the sub-sample t-values are independent, so the sub-samples are non-overlapping. §3.1.2 discusses overlapping sub-samples.

3.1.1 Non-overlapping sub-samples

We consider the following approximation of the t-statistic:

$$\begin{aligned} \mathtt{t}_0 &= \frac{\widehat{\beta}}{\widehat{\sigma}_\beta} = \left(\frac{\widehat{\sigma}_\varepsilon^2}{T} \left[\frac{1}{T} \sum_{t=1}^{T} x_t^2 \right]^{-1} \right)^{-\frac{1}{2}} \frac{\frac{1}{T} \sum_{t=1}^{T} x_t y_t}{\frac{1}{T} \sum_{t=1}^{T} x_t^2} \\ &= \frac{\sqrt{T}}{\widehat{\sigma}_\varepsilon} \frac{\frac{1}{T} \sum_{t=1}^{T} x_t y_t}{\sqrt{\frac{1}{T} \sum_{t=1}^{T} x_t^2}} \simeq \frac{\sqrt{T}}{\sigma_\varepsilon \sigma_x} \left(\frac{1}{T} \sum_{t=1}^{T} x_t y_t \right). \end{aligned} \tag{3}$$

Under stationarity and ergodicity, sample moments are consistent for population, so replacing the sample second moments by their population counterparts will introduce an error, but should not bias the calculations. However, when the data second moments for the conditioning variables changes substantially over the sample, different outcomes could be obtained. We also assume a small number of regressors in the selected model such that degree-of-freedom corrections can be neglected, and focus the analysis on the scalar problem to highlight the key issues.

If the sample is split into J non-overlapping partitions, the full-sample t-value is then given approximately by:

$$\mathtt{t}_0 \simeq \frac{\sqrt{T}}{\sigma_\varepsilon \sigma_x} \left(\frac{1}{T} \sum_{j=1}^{J} \sum_{t \in \mathcal{T}_j} x_t y_t \right), \tag{4}$$

where the jth sub-sample t-value is given by:

$$\mathtt{t}_j \simeq \frac{\sqrt{\tau_j T}}{\sigma_\varepsilon \sigma_x} \left(\frac{1}{\tau_j T} \sum_{t \in \mathcal{T}_j} x_t y_t \right), \tag{5}$$

when τ_j is the fraction of observations belonging to the jth partition, with $\sum_j \tau_j = 1$. Hence:

$$\mathtt{t}_0 \simeq \sum_{j=1}^{J} \sqrt{\tau_j} \mathtt{t}_j > \sum_{j=1}^{J} \tau_j \mathtt{t}_j, \tag{6}$$

so the weighted sum of sub-sample t-values is less than the full-sample t-value. If the partitions are of equal size, $\tau_j = 1/J$, we have that

$$\mathsf{t}_0 \simeq \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \mathsf{t}_j. \tag{7}$$

and, hence, $\mathsf{E}[\mathsf{t}_0] \simeq \sqrt{J}\mathsf{E}[\mathsf{t}_j]$.

For non-overlapping sub-samples, the t_j-values are independently distributed so we can derive the average squared t-value in the full-sample from the sub-sample t_j^2-values as follows:

$$\mathsf{E}[\mathsf{t}_0^2] \simeq \sum_{j=1}^{J} \tau_j \mathsf{E}[\mathsf{t}_j^2] + \sum_{j=1}^{J} \sum_{i \neq j} \sqrt{\tau_i \tau_j} \mathsf{E}[\mathsf{t}_j]\mathsf{E}[\mathsf{t}_i]. \tag{8}$$

Again assuming equal-sized partitions:

$$\mathsf{E}[\mathsf{t}_0^2] \simeq \mathsf{E}[\mathsf{t}_j^2] + \left(1 - \frac{1}{J}\right) \psi^2. \tag{9}$$

for the given full-sample non-centrality parameter ψ (which for convenience is taken to be positive), so we get the following relationship between the average squared t-value in the full sample and the sub-samples:

$$\mathsf{E}[\mathsf{t}_j^2] \simeq \frac{1}{J}\psi^2, \tag{10}$$

which is confirmed by a comparison of (4) and (5). The higher the non-centrality ψ, the stronger the shrinkage of the expected sub-sample $|\mathsf{t}|$-value (compared to that of the full sample). This reduction in the information content of sub-sample $|\mathsf{t}|$-test might be referred to as the 'curse of sub-samples'. It indicates that sub-sample-based selection rules will find it hard to detect DGP variables (at a given size), especially for ψ^2 values near the critical region.

3.1.2 Overlapping sub-samples

Suppose, for the following, that $J = 2$. If the sub-samples are overlapping, *i.e.*, $\tau \in (0.5, 1)$, their t-values are no longer independent. To overcome the correlation problem, we partition the sample into three independent partitions (say, a, b and c) and construct from these, the two sub-samples and the full-sample t-values. The three generated t-distributed random variables are $\mathsf{t}_i \sim \mathsf{t}\left(\tau_i T, \sqrt{\tau_i}\psi\right)$ for $i = a$ and c and $\mathsf{t}_b \sim \mathsf{t}\left((2\tau - 1)T, \sqrt{2\tau - 1}\psi\right)$ such that $\mathsf{t}_0 \sim \mathsf{t}(T, \psi)$. It follows from (6) that the full-sample t-value is given by:

$$\mathsf{t}_0 \simeq \sqrt{1 - \tau T}\, \mathsf{t}_a + \sqrt{2\tau - 1}\, \mathsf{t}_b + \sqrt{1 - \tau T}\, \mathsf{t}_c. \tag{11}$$

The two sub-sample t-values result as follows:

$$\mathsf{t}_1 \simeq \sqrt{\frac{1 - \tau T}{\tau T}}\, \mathsf{t}_a + \sqrt{\frac{2\tau - 1}{\tau T}}\, \mathsf{t}_b, \tag{12}$$

$$\mathsf{t}_2 \simeq \sqrt{\frac{1 - \tau T}{\tau T}}\, \mathsf{t}_c + \sqrt{\frac{2\tau - 1}{\tau T}}\, \mathsf{t}_b. \tag{13}$$

It can be shown that the result in (6) holds: the weighted sum of sub-sample t-values is again less than the full-sample t-value.

The advantage of the above procedure is that t_0 as well as t_1 and t_2 can be generated as weighted sums of independently t-distributed random variables. Next, in §3.2, we use the framework laid out here to investigate the Hoover–Perez sub-sample selection rule, which evaluates the minimum of the two sub-sample t-values. Then, §3.3 examines the properties of the *PcGets* post-selection reliability check, which assesses the sub-sample evidence conditional on full-sample significance.

3.2 The Hoover–Perez approach

3.2.1 Selection rule

We first consider the selection rule of Hoover and Perez, namely include a regressor if and only if its coefficient is significant in both sub-samples. In other words, the minimum of the two sub-sample $|\mathsf{t}|$-values needs to be significant:

$$\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\min}_{\alpha,\tau T}, \tag{14}$$

where τ denotes the size of the sub-sample as a fraction of the (full) sample and α is the size of the test. For the simple framework considered here, we can define and control the size of the procedure as

$$\alpha = \Pr\left(\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\min}_{\alpha,\tau T} \,\middle|\, \psi = 0 \right). \tag{15}$$

So α is the nominal and empirical size of the procedure, which implies that the power of the selection procedure is given by:

$$\pi(\alpha, \tau, \psi) = \Pr\left(\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\min}_{\alpha,\tau T} \,\middle|\, \psi > 0 \right), \tag{16}$$

where ψ is the full-sample population $|\mathsf{t}|$-value of the DGP variable. The properties of the selection rule will ultimately depend on the distribution of $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$ for given ψ, which we will explore in the following.[3]

3.2.2 Simulating the distribution of $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$

*Design.*We investigate the properties of the $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$ statistic by simulation. The Monte Carlo study consists of $M = 5\ 000\ 000$ replications of an experiment with $\mathsf{t}(\tau T, \sqrt{\tau}\psi)$ distributed random variables with a (full-sample) non-centrality of $\psi \in \{0, 2, 3, 4, 5\}$ and a full-sample size of $T = 100$. The size of the sub-samples is $[\tau T]$, where $\tau \in [0.5, 1]$, such that $\tau = 0.5$ denotes the case of non-overlapping sub-samples, $\tau \in (0.5, 1)$ implies overlapping sub-samples and $\tau = 1$ is the borderline case with the sub-samples and the full sample coinciding.

In the case of non-overlapping sub-samples ($\tau = 0.5$), the experiment consists of two $\mathsf{t}(\nu, \frac{1}{\sqrt{2}}\psi)$ distributed random variables with $\nu = \frac{T}{2} = 50$ degrees of freedom and a (full-sample) non-centrality ψ. Let $\{\mathsf{t}_1, \mathsf{t}_2\}$ be $\mathsf{t}(\nu)$ distributed random variables. Then, the full-sample t-value is given by:

$$\mathsf{t}_0 = \frac{1}{\sqrt{2}} (\mathsf{t}_1 + \mathsf{t}_2).$$

For overlapping sub-samples we use the approach in equations (11) to (13).

Probability density function (pdf) of $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$. The probability density function of $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$ is illustrated in figure 2 for the case of non-overlapping sub-samples (*i.e.*, $\tau = 0.5$) and, overlapping sub-samples ($\tau = 0.75$). Furthermore, figure 2 compares the pdf of $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$ to the density of the simple full-sample t-value. It can be seen that the probability mass is shifted to the left. The shift is greater for the non-overlapping sub-samples and increases with a growing non-centrality. This indicates that the discrimination between DGP variables ($\psi > 0$) and nuisance variables ($\psi = 0$) is getting harder when the analysis is based on sub-sample information. This intuition will be confirmed in the analysis in section 3.2.3. We now continue to evaluate the properties of the distribution of $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$.

[3]Lynch and Vital-Ahuja (1998) analyzed the related problem whether the use of sub-sample evidence can mitigate the potential impact of data snooping on the distribution of test statistics. Comparing sub-sample and entire sample R^2 tests, Lynch and Vital-Ahuja found that the full-sample test has a less distorted size and more power than the multi-sample test.

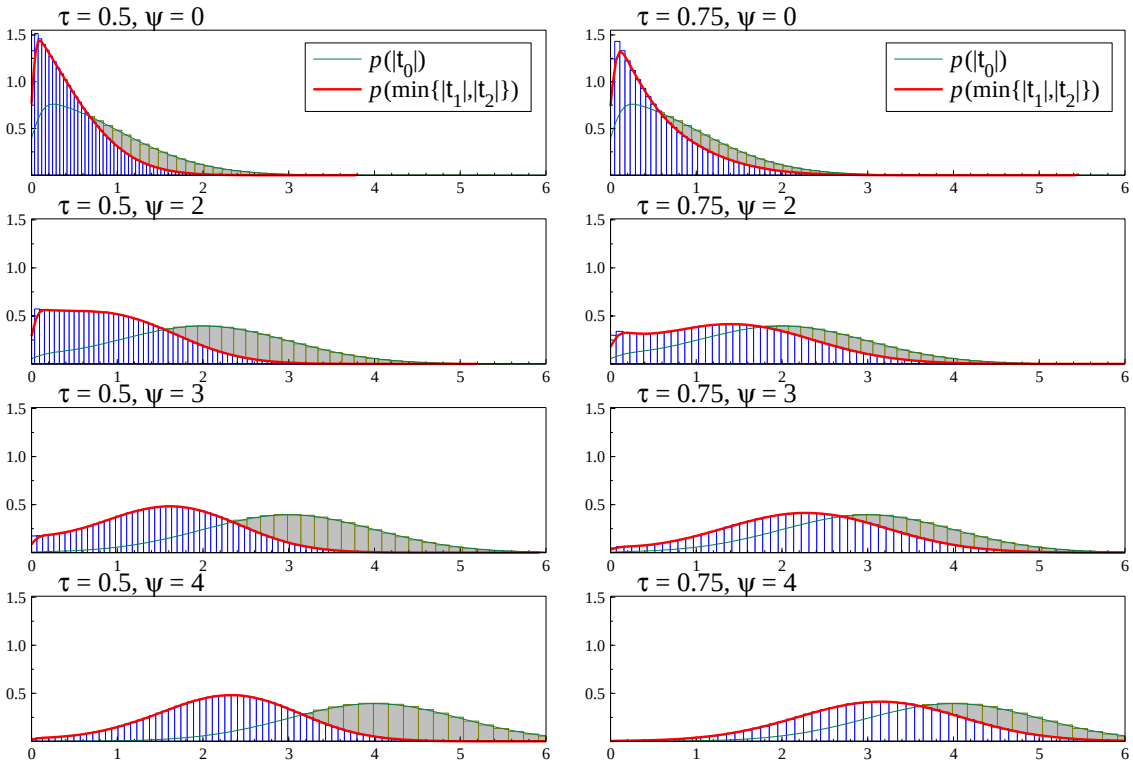

Figure 2 Density of the full-sample $|\mathrm{t}|$ and $\min\{|\mathrm{t}_1|, |\mathrm{t}_2|\}$ for $T = 100$.

Critical values. Table 1 reports the critical values $c^{\min}_{\alpha,\tau T}$ of the $\min\{|\mathrm{t}_1|, |\mathrm{t}_2|\}$ statistic for given size α:

$$c^{\min}_{\alpha,\tau T} = \left\{ c \left| \Pr\left(\min\{|\mathrm{t}_1|, |\mathrm{t}_2|\} > c \right| \psi = 0 \right) = \alpha \right\}.$$

When compared to the critical values of a full-sample t-test ($\tau = 1.0$), the critical values have to be chosen much lower to reflect the shift of the probability mass to the left. The smaller τ, the stronger the shift. For $\alpha = 0.05$, the critical value $c^{\min}_{0.05,100\tau}$ drops from 1.984 for $\tau = 1.0$, over 1.556 for $\tau = 0.75$, to 1.232 for $\tau = 0.5$.

Table 1 Critical values $c^{\min}_{\alpha,\tau T}$ for the sub-sample $\min\{|\mathrm{t}|_1, |\mathrm{t}|_2\}$ test.

$\tau \setminus \alpha$	1%	2.5%	5%	7.5%	10%
0.50	1.677	1.434	1.232	1.106	1.012
0.65	1.985	1.667	1.410	1.249	1.131
0.70	2.082	1.754	1.484	1.315	1.189
0.75	2.167	1.832	1.556	1.381	1.250
0.80	2.244	1.906	1.624	1.446	1.313
0.85	2.320	1.977	1.691	1.511	1.376
1.00	2.623	2.275	1.984	1.799	1.660

Also, table 2 reports the corresponding nominal significance levels of a simple t-test (with $\nu = \tau T$). In the case of non-overlapping sub-samples ($\tau = 0.5$), sizes of 1%, 5% and 10% of the $\min\{|\mathrm{t}_1|, |\mathrm{t}_2|\}$ test would only require critical values associated with a significance level of a simple t-test at 9.7%, 22.1% and 31.4%. For $\tau = 0.75$, the required levels are reduced to 3.3%, 12.3% and 21.4%.

In table 3, we suppose that the critical values have been taken from the $\mathrm{t}(\tau T)$ distribution. As the probability mass of the $\min\{|\mathrm{t}_1|, |\mathrm{t}_2|\}$ statistic is shifted to the left of the $|\mathrm{t}_0|$-density, the test becomes

Table 2 Nominal t-probabilities $\eta(\alpha, \tau)$ for the critical values $c_{\eta,\tau T} = c^{\min}_{\alpha,\tau T}$.

$\tau \setminus \alpha$	1%	2.5%	5%	7.5%	10%
0.50	0.0966	0.1546	0.2206	0.2712	0.3137
0.65	0.0498	0.0985	0.1616	0.2143	0.2604
0.70	0.0397	0.0823	0.1409	0.1914	0.2371
0.75	0.0325	0.0697	0.1227	0.1701	0.2140
0.80	0.0269	0.0594	0.1074	0.1512	0.1921
0.85	0.0223	0.0506	0.0937	0.1337	0.1718
1.00	0.0100	0.0249	0.0499	0.0749	0.0999

dramatically undersized: For a nominal significance level of 1%, 5% and 10%, the resulting size of the $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$ test in non-overlapping sub-samples ($\tau = 0.5$) is 0.01%, 0.25% and 1%, respectively.

Table 3 Size $\alpha(\eta, \tau)$ of the $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c_{\eta,T}$ test.

$\tau \setminus \eta$	1%	2.5%	5%	7.5%	10%
0.50	0.0001	0.0006	0.0025	0.0057	0.0100
0.65	0.0011	0.0038	0.0096	0.0168	0.0249
0.70	0.0017	0.0053	0.0129	0.0217	0.0315
0.75	0.0024	0.0071	0.0163	0.0268	0.0383
0.80	0.0031	0.0089	0.0201	0.0323	0.0455
0.95	0.0040	0.0111	0.0243	0.0385	0.0534
1.00	0.0099	0.0249	0.0500	0.0749	0.0999

3.2.3 Power size trade-off

We now derive the *power size trade-off* of the $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\}$ test statistic for given size α with the sub-sample size being a fraction τ of the full-sample:

$$\begin{aligned} \pi(\alpha, \tau, \psi) &= \Pr\Big(\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\min}_{\alpha,\tau T} \,\Big|\, \psi > 0 \Big), \\ \text{where } \alpha &= \Pr\Big(\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\min}_{\alpha,\tau T} \,\Big|\, \psi = 0 \Big). \end{aligned}$$

Figure 3 reports the resulting power–size trade-off function $\pi(\alpha; \tau, \psi)$ for the given (full-sample) non-centrality parameter $\psi \in \{2, 3, 4, 5\}$, sub-sample size $\tau \in \{0.50, 0.55, \ldots, 1.00\}$ and, for greater numerical stability, $T = 1000$. The $(\alpha, \pi(\tau, \psi))$ functional is derived by parametric variation of the critical value $c^{\min}_{\alpha,\tau T} = c_{\eta(\alpha,\tau),\tau T}$ according to its simple t-test significance level η, resulting in sequences of $\alpha(\eta, \tau)$ and $\pi(\alpha(\eta, \tau); \tau, \psi)$. The power loss is quite substantial (up to 40% for $\tau = 0.5$), but it is worth noting that analyzing overlapping sub-samples can retrieve part of the power loss.

The power of the test relative to the full-sample case,

$$\frac{\pi(\alpha; \tau, \psi)}{\pi(\alpha; 1, \psi)} \text{ with } \psi > 0,$$

is illustrated in figure 4 for sub-sample sizes of $\tau = 0.5$ to 1.0. The power is found to be a monotonically increasing function in τ, so we can conclude that the sub-sample-based selection rule of Hoover and Perez (1999) is dominated by the simple full-sample t-test.

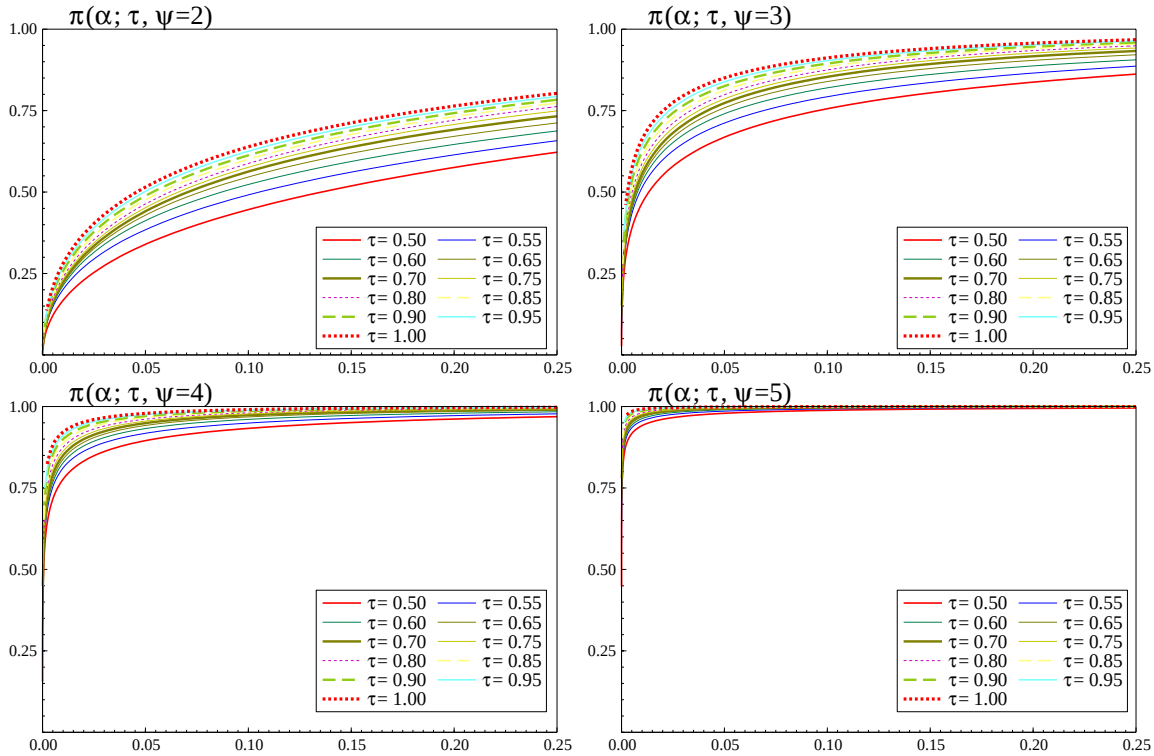

Figure 3 Power–size trade-off under the $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\min}_{\alpha,\tau T}$ selection rule ($T = 1000$).

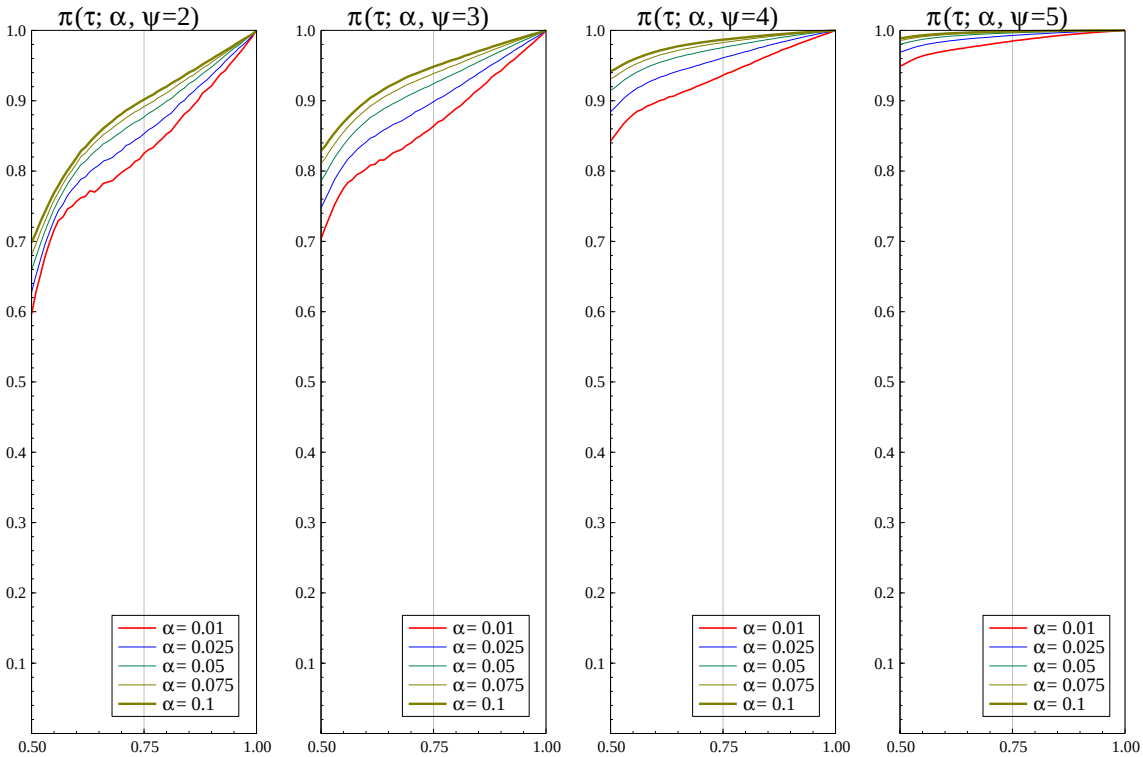

Figure 4 Relative power for given τ under the $\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\min}_{\alpha,\tau T}$ selection rule ($T = 1000$).

3.3 *PcGets* approach

3.3.1 Reliability statistic

In *PcGets*, a variable is selected if it is significant in the full sample, *i.e.*, $|\mathsf{t}_0| > c_{\gamma,T}$.[4] After selection, the relevance of variables in the final model is explored by post-selection reliability checks to ascertain whether 'significance' is substantive or adventitious.

The reliability of a regressor, which is normalized to be bounded between zero (no reliability) and one (full reliability), is a function of the full-sample $|\mathsf{t}_0|$-value and the significance of that regressor in the two sub-samples:

$$r(|\mathsf{t}_0|, |\mathsf{t}_1|, |\mathsf{t}_2|) \in [0, 1],$$

where the partial derivatives $r_i \geq 0$ for $i = 1, 2, 3$, $r(|\mathsf{t}_0|, \cdot, \cdot) = 0$ if $|\mathsf{t}_0| < c_{\gamma,T}$ and $r(\cdot) = 1$ if $|\mathsf{t}_i| > c^{\rm sub}_{\delta,\tau T}$ for all i. In the following, we consider parameterizations of the reliability function which are based on a constant penalty ρ for insignificance in sub-samples:

$$r(|\mathsf{t}_0|, |\mathsf{t}_1|, |\mathsf{t}_2|) = \mathrm{I}(|\mathsf{t}_0| > \mathrm{c}_{\gamma,\mathrm{T}}) \left[1 - \rho \mathrm{I}(|\mathsf{t}_1| < \mathrm{c}^{\rm sub}_{\delta,\tau\mathrm{T}}) - \rho \mathrm{I}(|\mathsf{t}_2| < \mathrm{c}^{\rm sub}_{\delta,\tau\mathrm{T}})\right], \tag{17}$$

where $\mathrm{I}(\cdot)$ is an indicator function with $\mathrm{I}(\mathcal{C}) = \infty$ if $\mathcal{C}$ is true and 0 otherwise. We allow here for different significance levels for the full sample (γ) and the sub-samples (δ). *PcGets* sets $\rho = 0.3$ and $c^{\rm sub}_{\delta,\tau T} = c_{1.5\gamma,\tau T}$, where — for typical macro-economic sample sizes — the significance level γ is 0.05 for the liberal and 0.01 for the conservative strategy.

Note that we can write (17) as:

$$r(|\mathsf{t}_0|, |\mathsf{t}_1|, |\mathsf{t}_2|) = \mathrm{I}(|\mathsf{t}_0| > \mathrm{c}_{\gamma,\mathrm{T}}) \left[1 - \rho \mathrm{I}(\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} < \mathrm{c}^{\rm sub}_{\delta,\tau\mathrm{T}}) - \rho \mathrm{I}(\max\{|\mathsf{t}_1|, |\mathsf{t}_2|\} < \mathrm{c}^{\rm sub}_{\delta,\tau\mathrm{T}})\right],$$

which can be easily compared to the Hoover–Perez rule:

$$r^{\rm HP}(|\mathsf{t}_1|, |\mathsf{t}_2|) = \mathrm{I}(\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > \mathrm{c}^{\rm min}_{\alpha,\tau\mathrm{T}}).$$

For $r^{\rm HP}(|\mathsf{t}_1|, |\mathsf{t}_2|)$, we defined size as:

$$\alpha = \Pr\left(\min\{|\mathsf{t}_1|, |\mathsf{t}_2|\} > c^{\rm min}_{\alpha,\tau T} \;\middle|\; \psi = 0\right) = \mathsf{E}\left[r^{\rm HP}(|\mathsf{t}_0|, |\mathsf{t}_1|, |\mathsf{t}_2|) \middle| \psi = 0\right].$$

In an analogous fashion, we can define size and power for the *PcGets* approach as follows:
Size ($\psi = 0$):

$$\begin{aligned}
\alpha(\gamma, \delta) &= \mathsf{E}\left[r(|\mathsf{t}_0|, |\mathsf{t}_1|, |\mathsf{t}_2|) \middle| \psi = 0\right] \\
&= \Pr\left(|\mathsf{t}_0| > c_{\gamma,T} \;\middle|\; \psi = 0\right) \left[1 - \rho \sum_{i=1}^{2} \Pr\left(|\mathsf{t}_i| < c^{\rm sub}_{\delta,\tau T} \;\middle|\; |\mathsf{t}_0| > c_{\gamma,T}, \psi = 0\right)\right];
\end{aligned}$$

Power ($\psi > 0$):

$$\begin{aligned}
\pi(\gamma, \delta, \tau, \psi) &= \mathsf{E}\left[r(|\mathsf{t}_0|, |\mathsf{t}_1|, |\mathsf{t}_2|) \middle| \psi\right] \\
&= \Pr\left(|\mathsf{t}_0| > c_{\gamma,T} \;\middle|\; \psi\right) \left[1 - \rho \sum_{i=1}^{2} \Pr\left(|\mathsf{t}_i| < c^{\rm sub}_{\delta,\tau T} \;\middle|\; |\mathsf{t}_0| > c_{\gamma,T}, \psi\right)\right],
\end{aligned}$$

[4] We abstract here from the possibility that a variable might be selected to ensure congruence, although it is not significant in the full sample.

which can be rewritten, for the size say, as:

$$\begin{aligned}
\alpha(\gamma,\delta) &= \mathsf{E}\left[r(|\mathsf{t}_0|,|\mathsf{t}_1|,|\mathsf{t}_2|)\middle|\,\psi=0\right] \\
&= \Pr\left(|\mathsf{t}_0|>c_{\gamma,T}\;\middle|\;\psi=0\right)\;\left[1-\rho\Pr\left(\min\{|\mathsf{t}_1|,|\mathsf{t}_2|\}<c^{\text{sub}}_{\delta,\tau T}\middle|\,|\mathsf{t}_0|>c_{\gamma,T},\psi=0\right)\right. \\
&\qquad\left.-\rho\Pr\left(\max\{|\mathsf{t}_1|,|\mathsf{t}_2|\}<c^{\text{sub}}_{\delta,\tau T}\middle|\,|\mathsf{t}_0|>c_{\gamma,T},\psi=0\right)\right] \\
&\simeq \Pr\left(|\mathsf{t}_0|>c_{\gamma,T}\;\middle|\;\psi=0\right)\left[(1-\rho)+\rho\Pr\left(\min\{|\mathsf{t}_1|,|\mathsf{t}_2|\}>c^{\text{sub}}_{\delta,\tau T}\;\middle|\,|\mathsf{t}_0|>c_{\gamma,T},\;\psi=0\right)\right] \\
&= (1-\rho)\Pr\left(|\mathsf{t}_0|>c_{\gamma,T}\;\middle|\;\psi=0\right)+\rho\Pr\left(\min\{|\mathsf{t}_1|,|\mathsf{t}_2|\}>c^{\text{sub}}_{\delta,\tau T},|\mathsf{t}_0|>c_{\gamma,T}\middle|\,\psi=0\right)
\end{aligned}$$

since $\Pr\left(\max\{|\mathsf{t}_1|,|\mathsf{t}_2|\}<c^{\text{sub}}_{\delta,\tau T}\middle|\,|\mathsf{t}_0|>c_{\gamma,T},\psi\right)\simeq 0$ for $\delta\simeq\gamma$.

Before investigating the power-size trade-off implied by the *PcGets* reliability statistic (17) in section 3.3.3, we proceed by analyzing the properties of the density of the sub-sample $|\mathsf{t}_i|$-value given its significance in full sample, *i.e.*, $|\mathsf{t}_0|>c_{\gamma,T}$.

3.3.2 Simulating the conditional distribution of $|\mathsf{t}_i|$ given $|\mathsf{t}_0|>c_{\gamma,T}$

Design. Using the same framework as in section 3.2.2, we now investigate the sub-sample properties of a single t-test when the analysis is conditioned on its significance in the full sample. The Monte Carlo study again consists of $M=5\,000\,000$ replications of the experiment with $\mathsf{t}(\tau T,\sqrt{\tau}\psi)$ distributed random variables with a full-sample non-centrality $\psi\in\{0,2,3,4,5\}$ and sample size $T=100$. The size of the sub-samples is $[\tau T]$, where $\tau\in[0.5,1]$, such that $\tau=0.5$ denotes the case of non-overlapping sub-samples, $\tau\in(0.5,1)$ implies overlapping sub-samples and $\tau=1$ is the borderline case with the sub-samples and the full sample coinciding.

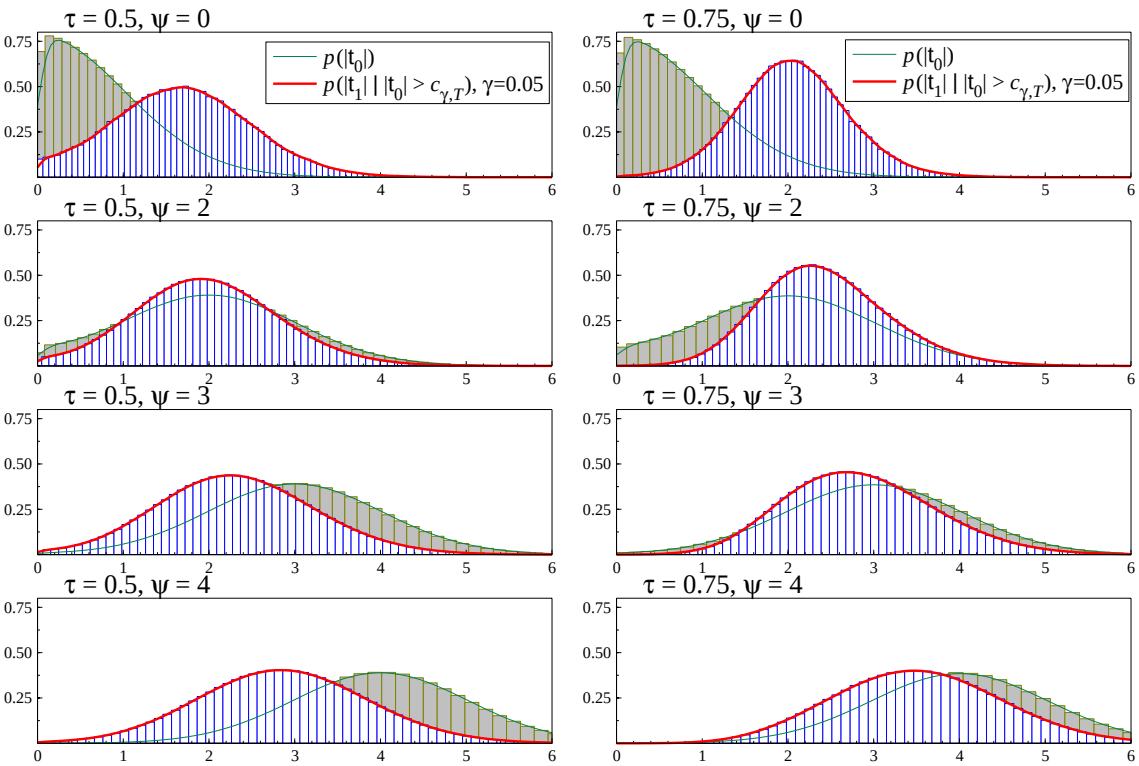

Figure 5 The density of $|\mathsf{t}_i|$ and $\min_i\{|\mathsf{t}_i|\}$ conditional on significance in the full sample.

Figure 5 plots the conditional density of $|\mathsf{t}_i|$ in non-overlapping ($\tau = 0.5$) and overlapping ($\tau = 0.75$) sub-samples conditional on significance in the full sample. When compared to the density of simple (full-sample) t-test, two effects become evident:

(i) For non-DGP variables, conditioning on significance in the full sample makes the pdf of its sub-sample $|\mathsf{t}|$-value more similar to the unconditional density of a DGP variable with non-centrality close to the critical value of the full-sample test, $\psi \approx c_{\gamma,T}$. Thus, probability mass is dramatically shifted to the right.

(ii) For DGP variables with a sufficiently high population $|\mathsf{t}|$-value, $\psi > c_{\gamma,T}$, the probability of being selected is close to one. So knowing the fact that the variable is significant in the full-sample does not have any significant information value attached. Thus, the effect just described, which is so powerful for non-DGP variables, does not play a role here. Instead the 'curse of sub-samples' is due to shifting the probability mass to the left.

The two effects greatly complicate the selection problem: if a regressor is significant in the full sample, $|\mathsf{t}_0| > c_{\gamma,T}$, the distribution of the sub-sample $|\mathsf{t}|$-values of a variable that matters ($\psi > 0$) is hardly distinguishable from that of a nuisance variable ($\psi = 0$). A comparison of the two depicted cases ($\tau = 0.5$ versus $\tau = 0.75$) suggests the use of information from overlapping sub-samples for the reliability statistic.

The resulting size of the conditional sub-sample $|\mathsf{t}_i|$ test at critical values corresponding to the reported *nominal* significance levels of a simple t-test is reported in table 4 for $\gamma = 0.05$ and in table 5 for a full-sample significance level of $\gamma = 0.01$. In the split-sample analysis of *PcGets*, the size of the sub-sample is $0.75T$ and the *nominal* significance level is 1.5γ, where γ is the significance level in the full sample. Thus, a nuisance parameter which is significant in the full sample has a 64.97% probability of passing the sub-sample test using the *PcGets* liberal strategy (54.1% for the conservative strategy).

Table 4 Size δ of $|\mathsf{t}_1| > c_{\eta,\tau T}$ given $|\mathsf{t}_0| > c_{0.05,T}$.

$\tau \setminus \delta$	1%	2.5%	5%	7.5%	10%
0.50	0.1108	0.2174	0.3404	0.4299	0.5006
0.65	0.1536	0.2985	0.4559	0.5618	0.6396
0.70	0.1641	0.3236	0.4947	0.6068	0.6858
0.75	0.1720	0.3465	0.5322	0.6497	0.7309
0.80	0.1804	0.3730	0.5751	0.6981	0.7787
0.85	0.1881	0.4011	0.6254	0.7532	0.8307
1.00	0.1989	0.4993	1.0000	1.0000	1.0000

Table 5 Size δ of $|\mathsf{t}_1| > c_{\eta,\tau T}$ given $|\mathsf{t}_0| > c_{0.01,T}$.

$\tau \setminus \eta$	1%	1.5%	2%	2.5%	5%
0.50	0.2401	0.3051	0.3545	0.3962	0.5423
0.65	0.3613	0.4449	0.5075	0.5577	0.7130
0.70	0.4002	0.4912	0.5582	0.6102	0.7645
0.75	0.4439	0.5412	0.6129	0.6664	0.8160
0.80	0.4881	0.5933	0.6676	0.7225	0.8623
0.85	0.5438	0.6569	0.7320	0.7834	0.9047
1.00	1.0000	1.0000	1.0000	1.0000	1.0000

To illustrate the procedure, we also report here the results for the hypothetical case of $\tau = 1$. This results in a two-stage test, where on the first stage a simple t-test is performed at a significance level

of 0.05. Conditional on the outcome of that test, a further t test is applied to significant variables at a nominal size of δ. Clearly all t_0-values with $|\mathrm{t}_0| > c_{\gamma,T}$ are going to pass this test if $\delta \geq \gamma$.

Table 6 reports the critical value $c^{\mathrm{sub}}_{\delta,\tau T}$ of the sub-sample t-test conditional on significance in the full sample $|\mathrm{t}| > c_{\gamma,T}$, when the size of the sub-sample test is calibrated to equalize the size in the full sample, *i.e.*, $\delta = \gamma$. It illustrates the shift of the pdf to the right, when compared to the pdf of an unconditional t-test.

Table 6 Critical values $c^{\mathrm{sub}}_{\gamma,\tau T}$ of the sub-sample t-test conditional on $|\mathrm{t}_0| > c_{\gamma,T}$.

$\tau \setminus \gamma$	1%	2.5%	5%	7.5%	10%	20%	30%	40%	50%
0.50	4.148	3.526	3.049	2.743	2.520	1.926	1.522	1.194	0.912
0.65	4.148	3.588	3.130	2.840	2.619	2.034	1.635	1.311	1.021
0.70	4.167	3.590	3.136	2.846	2.630	2.051	1.657	1.337	1.051
0.75	4.141	3.583	3.132	2.848	2.634	2.065	1.677	1.360	1.078
0.80	4.120	3.569	3.128	2.847	2.636	2.075	1.690	1.379	1.102
0.85	4.113	3.570	3.127	2.846	2.638	2.082	1.703	1.396	1.124
1.00	4.089	3.532	3.099	2.828	2.624	2.081	1.712	1.416	1.157

Table 7 corresponds to the previous table. It reports the nominal significance level of a simple t-test when the critical values $c^{\mathrm{sub}}_{\gamma,\tau T}$ given by table 6 are used. For reference, we also report the results for the sequential t-test implied by $\tau = 1$.

Table 7 Nominal $\mathrm{t}(\tau T)$-tail probability $\eta(\gamma,\tau)$ for the critical values $c^{\mathrm{sub}}_{\gamma,\tau T}$.

$\tau \setminus \gamma$	1%	2.5%	5%	7.5%	10%	20%	30%	40%	50%
0.50	0.0001	0.0009	0.0037	0.0084	0.0150	0.0598	0.1344	0.2381	0.3660
0.65	0.0001	0.0006	0.0026	0.0060	0.0110	0.0460	0.1069	0.1945	0.3110
0.70	0.0001	0.0006	0.0025	0.0058	0.0105	0.0440	0.1019	0.1855	0.2970
0.75	0.0001	0.0006	0.0025	0.0057	0.0102	0.0423	0.0977	0.1778	0.2846
0.80	0.0001	0.0006	0.0025	0.0056	0.0101	0.0412	0.0949	0.1717	0.2736
0.85	0.0001	0.0006	0.0024	0.0055	0.0099	0.0403	0.0922	0.1664	0.2642
1.00	0.0001	0.0006	0.0025	0.0057	0.0101	0.0400	0.0900	0.1599	0.2501

Analogously to tables 6 and 7, the two following tables 8 and 9 report critical values $c^{\mathrm{sub}}_{\delta,\tau T}$ and nominal simple t-test significance levels of the conditional sub-sample t-test, but now under the assumption that the full-sample evidence has been evaluated at a given significance level of $\gamma = 0.05$. For $\tau = 0.75$, an actual size of 0.05 requires a critical value of 3.132, which corresponds to a nominal size of 0.25% in a simple t-test. For $\tau = 1$, the critical values can be taken from a $\mathrm{t}(T,\psi)$-distribution evaluated at the two-sided tail-probability $\eta = \delta\gamma$.

Table 8 Critical values $c^{\mathrm{sub}}_{\delta,\tau T}$ of the sub-sample t-test conditional on $|\mathrm{t}_0| > c_{0.05,T}$.

$\tau \setminus \delta$	1%	2.5%	5%	7.5%	10%	20%	30%	40%	50%
0.50	3.662	3.324	3.049	2.866	2.728	2.362	2.101	1.880	1.677
0.65	3.689	3.382	3.130	2.970	2.850	2.522	2.291	2.098	1.919
0.70	3.688	3.383	3.136	2.981	2.865	2.553	2.336	2.155	1.986
0.75	3.668	3.364	3.132	2.983	2.873	2.574	2.369	2.198	2.042
0.80	3.649	3.361	3.128	2.986	2.877	2.593	2.402	2.243	2.097
0.85	3.652	3.363	3.127	2.985	2.880	2.610	2.430	2.283	2.151
1.00	3.598	3.319	3.099	2.967	2.868	2.624	2.474	2.364	2.275

Table 9 Nominal $\mathtt{t}(\tau T)$-tail probability $\eta(\delta,\tau)$ for the critical values $c^{\rm sub}_{\delta,\tau T}$.

$\tau \setminus \delta$	1%	2.5%	5%	7.5%	10%	20%	30%	40%	50%
0.50	0.0006	0.0017	0.0037	0.0061	0.0088	0.0221	0.0407	0.0659	0.0998
0.65	0.0005	0.0012	0.0026	0.0042	0.0058	0.0141	0.0252	0.0398	0.0594
0.70	0.0004	0.0012	0.0025	0.0040	0.0055	0.0129	0.0224	0.0346	0.0510
0.75	0.0005	0.0012	0.0025	0.0039	0.0053	0.0120	0.0204	0.0310	0.0447
0.80	0.0005	0.0012	0.0025	0.0037	0.0052	0.0113	0.0186	0.0277	0.0391
0.85	0.0004	0.0012	0.0024	0.0037	0.0050	0.0107	0.0172	0.0249	0.0343
1.00	0.0005	0.0013	0.0025	0.0038	0.0050	0.0101	0.0151	0.0200	0.0250

3.3.3 Power size trade-off

We now derive the *power* of the reliability statistic for given size α and the sub-sample size being a fraction τ of the full-sample:

$$\pi(\gamma,\tau,\psi) \;=\; \Pr\Big(|\mathtt{t}_0|>c_{\gamma,T}\;\Big|\;\psi\Big)\left[1-0.3\sum_{i=1}^{2}\Pr\Big(|\mathtt{t}_i|<c_{1.5\gamma,\tau T}\;\Big|\;|\mathtt{t}_0|>c_{\gamma,T},\psi\Big)\right],$$

$$\text{where } \alpha(\gamma,\tau) \;=\; \Pr\Big(|\mathtt{t}_0|>c_{\gamma,T}\;\Big|\;\psi=0\Big)\left[1-0.3\sum_{i=1}^{2}\Pr\Big(|\mathtt{t}_i|<c_{1.5\gamma,\tau T}\;\Big|\;|\mathtt{t}_0|>c_{\gamma,T},\psi=0\Big)\right].$$

For the derivation of the power-size trade-off $\pi(\alpha;\psi,\tau)$ shown in figure 6, we use the same approach as before. The $(\alpha,\pi(\psi,\tau))$ functional is produced by parametric variation of the nominal significance level γ.

Figure 6 reports the resulting power–size trade-off for $T=1000$. The efficient frontier is again given by the full-sample analysis ($\tau=1$). While using non-overlapping sub-samples ($\tau=0.5$) delivers the worst power at any size α, analyzing overlapping sub-samples can retrieve part of the power loss. This is illustrated in figure 7, which plots the power of the *PcGets* reliability statistic $\pi(\alpha;\tau,\psi)$ relative to the power of the full-sample analysis $\pi(\alpha;1,\psi)$ for sub-sample sizes of $\tau=0.5$ to 1.0. The power is found to be a monotonically increasing function in τ. For the sub-sample size used by *PcGets* (*i.e.*, $\tau=0.75$), the power loss is less than 20% for $\psi>2$.

While the loss in power is as severe as in the case of the Hoover and Perez (1999) sub-sample-based selection rule, it is less damaging, since the reliability statistics are only provided as an additional information source: The *PcGets* model selection process proceeds on the basis of the full-sample evidence; then, the reliability of the selected variables is reported, and the user's own model choice might take this into consideration. For the size and power calculations presented here, we assumed that the reliability statistics are translated into retention probabilities in a linear fashion. It is also worth noting, that we derived the simulation results under the assumption of structural stability. In practice, models are subject to structural breaks, so gains from analyzing sub-sample information can be expected in that setting.

4 Conclusion

Model selection is an important part of a progressive research strategy, and itself is progressing rapidly. The sub-sample reliability procedure appears in Monte Carlo studies to reduce size at a small cost in power, but does not in fact result in a trade-off that is genuinely beneficial, although it certainly seems relatively costless, and has successfully controlled the null rejection frequency for selection problems that were previously deemed almost intractable (see e.g., Lovell, 1983).

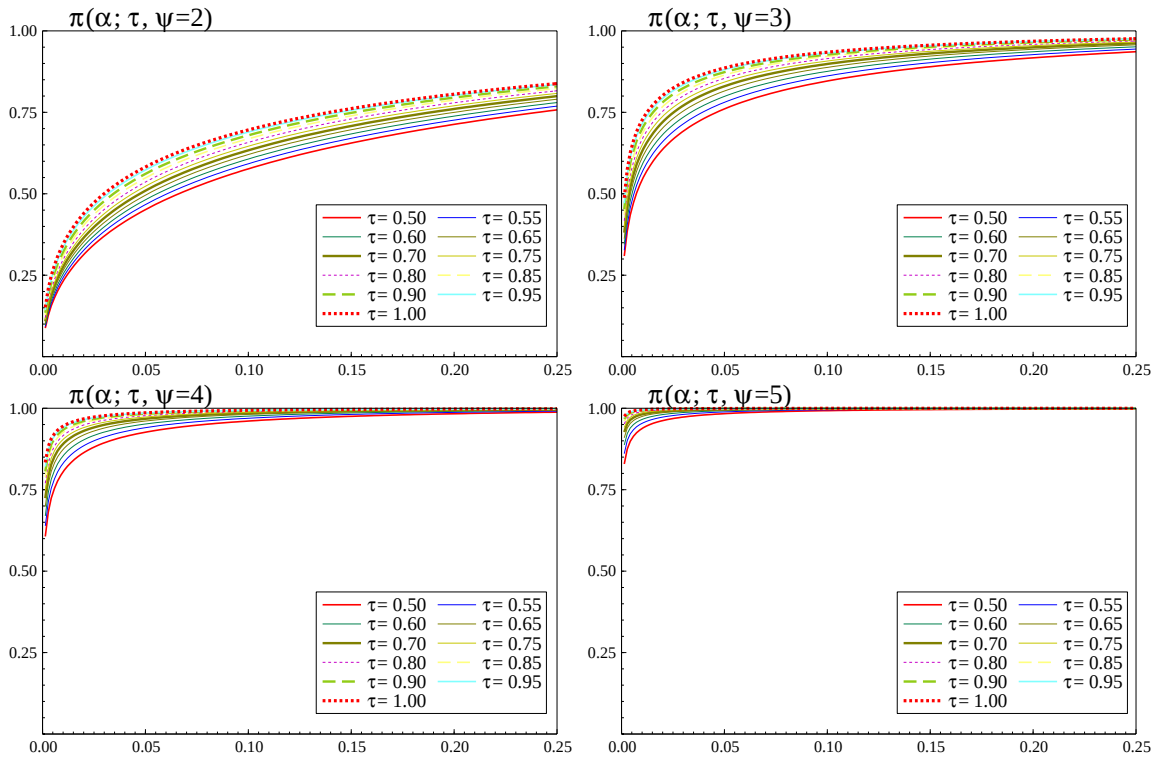

Figure 6 Power-size trade-off for the *PcGets* reliability function ($T = 1000$).

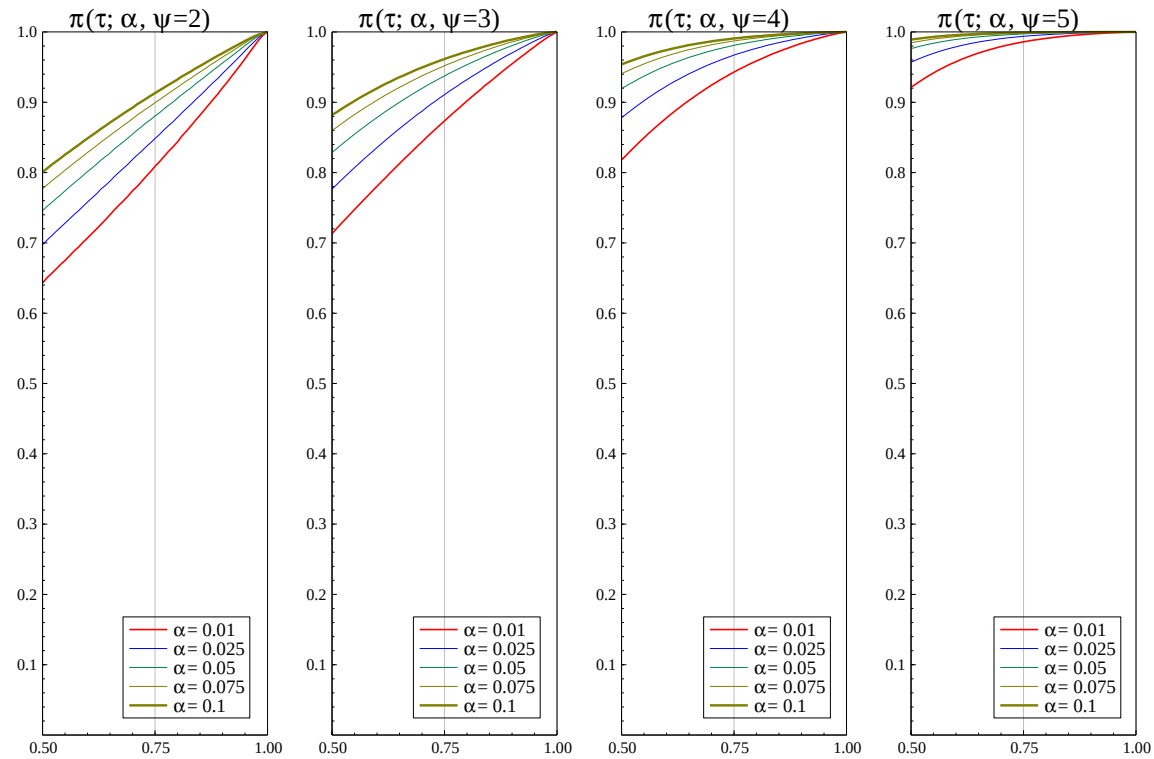

Figure 7 Relative Power for given τ for the *PcGets* reliability function ($T = 1000$).

References

Brüggemann, R., Krolzig, H.-M., and Lütkepohl, H. (2002). Comparison of model selection procedures for VAR processes. Discussion Paper, Humboldt–University, Berlin.

Doornik, J. A. (2001). *Object-Oriented Matrix Programming using Ox* 4th edn. London: Timberlake Consultants Press.

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

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 with PcGets*. London: Timberlake Consultants Press.

Hendry, D. F., and Krolzig, H.-M. (2002). New developments in automatic general-to-specific modelling. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*. Princeton: Princeton University Press. forthcoming.

Hendry, D. F., and Krolzig, H.-M. (2003). The properties of automatic *Gets* modelling. Discussion Paper, Department of Economics, University of Oxford.

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

Hoover, K. D., and Perez, S. J. (2000). Truth and robustness in cross-country growth regressions. unpublished paper, Economics Department, University of California, Davis.

Krolzig, H.-M. (2001). General-to-specific reductions in vector autoregressive processes. In Friedmann, R., Knüppel, L., and Lütkepohl, H.(eds.), *Econometric Studies - A Festschrift in Honour of Joachim Frohn*, pp. 129–157. Münster: LIT Verlag.

Krolzig, H.-M. (2003). General-to-specific model selection procedures for structural vector autoregressions. Discussion Paper, Department of Economics, University of Oxford.

Krolzig, H.-M., and Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, **25**, 831–866.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, **65**, 1–12.

Lynch, A. W., and Vital-Ahuja, T. (1998). Can subsample evidence alleviate the data-snooping problem? A comparison to the maximal R^2 cutoff test. Discussion paper, Stern Business School, New York University.