institute imdea social sciences

working papers series
in Economics and Social Sciences
2011-09

Automated model selection in finance: General-to-specific modelling of the mean and volatility specifications

by Genaro Sucarrat and Alvaro Escribano

June 2011

Ctra. Colmenar Viejo km. 14, 28049 Madrid, Spain

http://www.socialsciences.imdea.org/
http://repec.imdea.org
contacto.cienciassociales@imdea.org

www.imdea.org

Automated Model Selection in Finance: General-to-Specific Modelling of the Mean and Volatility Specifications[*]

Genaro Sucarrat[†]and Alvaro Escribano[‡]

23 June 2011

FORTHCOMING in the Oxford Bulletin of Economics and Statistics

Abstract

General-to-Specific (GETS) modelling has witnessed major advances over the last decade thanks to the automation of multi-path GETS specification search. However, several scholars have argued that the estimation complexity associated with financial models constitutes an obstacle to multi-path GETS modelling in finance. Making use of a recent result on log-GARCH Models, we provide and study simple but general and flexible methods that automate financial multi-path GETS modelling. Starting from a general model where the mean specification can contain autoregressive (AR) terms and explanatory variables, and where the exponential volatility specification can include log-ARCH terms, asymmetry terms, volatility proxies and other explanatory variables, the algorithm we propose returns parsimonious mean and volatility specifications. The finite sample properties of the methods are studied by means of extensive Monte Carlo simulations, and two empirical applications suggest the methods are very useful in practice.

JEL Classification: C32, C51, C52, E44
Keywords: General-to-specific, specification search, model selection, finance, volatility

[*] We are grateful to Luc Bauwens, Casper Christophersen, Jurgen Doornik, Neil R. Ericsson, David F. Hendry, Sebastien Laurent, André Alves Portela Santos, Enrique Sentana, the reviewers, and seminar participants at University of Oslo, BI Norwegian Business School, Statistics Norway, CORE and Universidad de Navarra, and conference participants at the XVII Foro de Finanzas, the XXXIV Simposio Economico, Forskermøtet 2010 and the 8th. OxMetrics User Conference, for helpful comments and suggestions. This research was supported by a Marie Curie Intra-European Fellowship within the 6th. European Community Framework Programme, and funding from the Bank of Spain Excellence Program is gratefully acknowledged.

[†] Corresponding author. Department of Economics, BI Norwegian Business School, email: genaro.sucarrat@bi.no, webpage: http://www.sucarrat.net/

[‡] Department of Economics, Universidad Carlos III de Madrid (Spain).

1 Introduction

Most financial models are highly non-linear and require complex optimisation algorithms and inference strategies in empirical application. Examples of some of the estimation and inference issues that may need careful attention include numerical approximation, multiple optima, convergence issues, negative variance, initial values, parameter constraints, finite sample approximations, and so on. For models with few parameters this may not pose unsurmountable problems. But as the number of parameters increases the resources and efforts needed for reliable estimation, inference and model validation multiply. Indeed, this may even become an obstacle to financial multi-path General-to-Specific (GETS) modelling, as for example argued by Granger and Timmermann (1999), and McAleer (2005) regarding automated GETS modelling of financial volatility.

A recent result in Sucarrat and Escribano (2010) means many of the estimation and inference complexity issues typically associated with financial models can readily be overcome. Specifically, they provide a result that enables consistent least squares estimation and inference of power log-GARCH models[1] for fixed power, under very general assumptions on the standardised error. Moreover, their simulations suggest Ordinary Least Squares (OLS) estimation of power log-ARCH models compares favourably with (Gaussian) Quasi Maximum Likelihood (QML) estimation when the standardised error is non-Gaussian. Since the exponential specification ensures positive variance and imposes fewer restrictions on the parameter space than standard ARCH models, this effectively means that estimation and inference are greatly simplified. Next, building on the work on automated GETS modelling by Hoover and Perez (1999), Hendry and Krolzig (2001, 2005), and Doornik (2009), and on the study by Bauwens and Sucarrat (2010) on GETS modelling of exchange rate volatility, we propose a simple but general and flexible model framework, and develop associated algorithms that automate multi-path GETS modelling of both the mean and volatility specifications. Starting from a general model where the mean specification (AR-X) can include autoregressive (AR) terms and explanatory variables (X), and where the exponential volatility specification can include log-ARCH, asymmetry terms, volatility proxies and other explanatory variables, our algorithm returns parsimonious mean and volatility specifications. The parameters of the variance specifications in the model we propose are consistently estimated by means of OLS, and inference regarding the parameters is performed by means of ordinary inference theory. Log-ARCH models can be viewed as nesting certain classes of stochastic volatility (SV) models, see Sucarrat and Escribano (2010), so we label the model a stochastic exponential ARCH (SEARCH) model. The acronym also connotates our main motivation for the model, namely that it facilitates specification search.

It is well known that ordinary ARCH models can be consistently estimated by

[1] "GARCH" is short for generalised autoregressive conditional heteroscedasticity, and the origins of the acronym are Engle (1982) and Bollerslev (1986).

means of OLS, and that heteroscedasticity robust inference strategies hold asymptotically. However, for several reasons, the least squares estimation and inference procedures associated with the power log-ARCH model have much better properties. First, the error term in the ARCH regression is heteroscedastic. By contrast, the error term in power log-GARCH regressions is IID. Second, the distribution of the error term in the ARCH regression has an exponential-like shape, and takes on values in $[-1, \infty)$. By contrast, in the power log-ARCH regression it is almost symmetric with the left-tail usually being "longer", and takes on values in $(-\infty, \infty)$. This means estimators and test-statistics are likely to correspond much closer to their asymptotic approximations in finite samples, since the convergence to their asymptotic counterparts will be much faster. Also, coefficient tests are likely to exhibit greater power under the alternative, since the error is "smaller" due to the log-transformation (this is the motivation for logarithmic Mincer and Zarnowitz (1969) regressions of volatility, see Pagan and Schwert (1990)). Finally, power log-GARCH regressions impose much weaker restrictions on the parameter space due to the exponential variance specification. In ARCH regressions, by contrast, strong parameter restrictions might be needed in order to ensure positive variance. The consequence of this is that a power log-GARCH regression is capable of depicting a much larger range of economic phenomena than an ARCH regression.

GETS modelling starts with a general unrestricted model (GUM) that is validated against a chosen set of misspecification tests. Next, simplification is undertaken by means of backwards step-wise regression, where each deletion is checked against the misspecification tests, and by a backtest (BaT) against the GUM.[2] Simplification stops when there are no more insignificant regressors, or when the possible deletions either do not pass the misspecification tests or the BaT. Multi-path GETS undertakes the simplification along several deletion paths, which may result in multiple terminal models. If so, then procedures for the selection between them are implemented. The main attraction of GETS modelling is that it takes the influence of all the potential variables into account already at the outset, which ideally should result in better estimation and inference, since each variable's impact is controlled for the impact of the others. Also, GETS model selection provides a predictable way of controlling the likeliness of retaining irrelevant variables via the regressor significance level.

GETS modelling has witnessed major advances over the last decade thanks to the development of *multi-path* GETS specification search software that automates the modelling process, see amongst others Hoover and Perez (1999), Hendry and Krolzig (2001, 2005), Krolzig (2003) and Doornik (2009).[3] An important result in this literature is that multi-path GETS can be vastly superior to single-path

[2] The BaT is also known as a parsimonious encompassing test.

[3] GETS specification search is closely related to, but not the same as, the GETS methodology, see Campos et al. (2005) for a comprehensive overview of the GETS methodology, and Mizon (1995) for a concise overview. Doornik (2008) discusses the relation between encompassing, an important aspect of the GETS methodology, and GETS specification search.

GETS and many other specification search algorithms. Key to the success is that estimation is essentially by means of OLS procedures, something which renders the automation of multi-path GETS specification search feasible in practice. Otherwise, if the multi-path GETS search were to be undertaken manually, practitioners and academics would often find it too time-consuming and cumbersome to implement.

The finite sample properties of the methods and algorithms that we propose are studied by means of extensive Monte Carlo simulation, and illustrated in two empirical applications. The simulations suggest that our algorithm compares well both in modelling the mean and volatility specifications, and the empirical applications confirm that the methods can be very useful in practice. The rest of the paper we organise in four sections. The next section, section 2, outlines the SEARCH model. Section 3 studies the properties of our multi-path GETS algorithm through extensive Monte Carlo simulation. Section 4 contains the two empirical applications and, finally, section 5 concludes.

2 The SEARCH model

If automated multi-path GETS modelling is the objective, then there are limits[4] to what the mean and volatility specifications may contain. An example of a general and flexible structure that is amenable to automated multi-path GETS modelling is what we label the SEARCH model. In words, the SEARCH model can be described as an AR(M)-X specification in the mean, and as a power log-ARCH(P) specification with asymmetry terms, a volatility proxy and explanatory variables in the logarithmic volatility specification. Specifically, the SEARCH model is given by

$$r_t = \phi_0 + \sum_{m=1}^{M} \phi_m r_{t-m} + \sum_{n=1}^{N} \eta_n x_{nt} + \epsilon_t \qquad (1)$$

$$\epsilon_t = \sigma_t z_t, \quad z_t \sim IID(0,1), \qquad (2)$$

$$\begin{aligned} \log \sigma_t^\delta &= \alpha_0 + \sum_{p=1}^{P} \alpha_p \log|\epsilon_{t-p}|^\delta + \sum_{a=1}^{A} \lambda_a (\log|\epsilon_{t-a}|^\delta) I_{\{z_{t-a}<0\}} \\ &\quad + \omega_0 \log EqWMA_{t-1} + \sum_{d=1}^{D} \omega_d y_{dt}, \quad \delta > 0 \qquad (3) \end{aligned}$$

In the mean specification (1) ϕ_0 is the mean intercept, M is the number of autoregressive (AR) terms and N is the number of other conditioning variables that may be contemporaneous and/or lagged. Moving average (MA) terms are not in-

[4] One should maybe add the qualifier "current", because future developments are likely to broaden the class of models that are amenable to automated GETS modelling.

cluded in the mean specification in order to simplify estimation and thus specification search. However, the estimation and inference methods we employ for the log-volatility specification (equation (3)) will in general be applicable if MA terms or other non-linearities are included in the mean. One type of non-linearity that our methods does not admit, though, is GARCH-in-mean terms due to the dependence with the volatility specification.[5] The standardised errors $\{z_t\}$ are IID zero mean and unit variance, and can be both more or less fat-tailed than the normal, and possibly skewed.[6] In the logarithmic volatility specification (3), δ is the power. Throughout we will set δ to 2 for convenience, but our methods and algorithms can also be applied when δ differs from 2. Indeed, δ can in principle take on any real-valued number—integer or not—strictly greater than zero. P is the number of log-ARCH terms, and if sufficiently big then the log-ARCH(P) structure can be viewed as an approximation to a stationary log-GARCH(P,Q) specification. The λ_a are the impacts of logarithmic asymmetry term analogous to those of Glosten et al. (1993), but one may consider other asymmetry specifications instead, see Sucarrat and Escribano (2010). The question of which approach to asymmetry is most appropriate we leave for future research. The $\log EqWMA_{t-1}$ term is the natural logarithm of a volatility proxy that is equal to an equally weighted moving average of the past absolute residuals raised to the power δ. That is, given the residuals $\{\hat{\epsilon}_t\}$, $EqWMA_{t-1}$ is computed as $(1/T^*)\sum_{t^*=1}^{T^*}|\hat{\epsilon}_{t-t^*}|^\delta$ where T^* is the length of the moving average. It should be noted that the term $\log EqWMA_{t-1}$ can be viewed as a local approximation to $\log \sigma_{t-1}^\delta$, that is, a volatility proxy. The attractive properties of $\log EqWMA_{t-1}$ compared with $\log \sigma_{t-1}^\delta$, though, is that it is simpler to estimate the associated parameter of the former, and that ordinary least squares inference regarding the parameter ω_0 can be undertaken. D is the number of other conditioning variables that may be contemporaneous and/or lagged, and if $\lambda_1 = \cdots = \lambda_A = \omega_0 = \omega_1 = \cdots = \omega_D = 0$, then $|\sum_{p=1}^{P}\alpha_p| < 1$ is a sufficient condition for stability in the log-volatility specification. If z_t is distributed as a Normal, a Generalised Error Distribution (GED) with shape distribution greater than 1, or a Student's t with more than two degrees of freedom, then the unconditional variance will in general exist. Under standard assumptions (stability, etc.) the parameters of both the mean and volatility specifications can be estimated consistently by means of least squares. In particular, the log-volatility specification can be estimated via an AR-representation by means of an OLS procedure. Furthermore, if the mean is zero or if it is estimated with sufficiently high precision, then ordinary OLS inference in the AR-representation of the log-volatility specification is asymptotically valid for all the parameters apart from the constant α_0. See Sucarrat and Escribano (2010) for further details.

[5] This is not necessarily a serious drawback, since proxies for financial price variability (functions of past squared returns, bid-ask spreads, functions of high-low values, etc.) that can be included as regressors in the mean are readily available.

[6] Some moments may not exist if the density is too fat-tailed, see Sucarrat and Escribano (2010).

3 Financial multi-path GETS modelling

In this section we propose and study a simple and very flexible algorithm for financial multi-path GETS modelling of the SEARCH model. We underline that the algorithm might possibly be improved in numerous ways (we will briefly discuss some of them at various points), but we leave this for future research. The results in this section could therefore be viewed as a minimal starting point, or as a "lower bound" of what is possible.

The automated GETS algorithm we propose can be viewed as consisting of two stages,[7] whose starting point is an (overall) General Unrestricted Model (GUM). That is, a model with general unrestricted mean (MGUM) and volatility (VGUM) specifications. The first stage consists of multi-path GETS specification search of the MGUM specification, while the VGUM specification is kept unchanged. Of course, one could instead consider to model them simultaneously or alternatively the volatility specification first. But we leave this for future research. Our choice of modelling the mean specification first is simply motivated by technical and conceptual simplicity. The second stage of the algorithm we propose consists of multi-path GETS specification search of the VGUM specification, while the parsimonious mean specification is kept unchanged. Again, one could of course consider alternative search procedures, say, multi-path GETS specification search applied to each of the terminal specifications from the GETS search of the mean. But, again, we leave this for future research, and again our choice is based on technical and conceptual considerations. The purpose of this section is to study the properties of the first and second stages of our algorithm through Monte Carlo simulations.

In the Monte Carlo simulations we will focus on three statistics. Let k_0 denote the number of relevant variables in the GUM, and let k_1 denote the number of irrelevant variables in the GUM. The first statistic $\hat{p}(DGP)$ is simply the probability of recovering the DGP exactly, that is, the probability of selecting a model such that $\hat{k}_0 = k_0$ and $\hat{k}_1 = 0$. The statistic $\hat{p}(DGP)$ is intimately related to what in a multiple hypothesis testing context is called the Family-Wise Error (FWE), which is simply the probability of making one or more false rejections.[8] Specifically, in a GETS context the FWE is 1-$p(DGP)$, and consistent model selection takes place when $p(DGP)$ tends to 1 as the sample size goes to infinity, or alternatively that the FWE tends to 0. As pointed out by Romano et al. (2008), however, the FWE is rather conservative, and the FWE may in any case not be the error rate of greatest interest. The two statistics of (arguably) greatest interest in a GETS context are the average relevance proportion $M(\hat{k}_0/k_0)$, which is analogous to statistical power in a hypothesis testing context and which Doornik (2009) calls "potency", and the

[7] In earlier versions of this paper we included a third stage that modelled the density of the standardised error z_t. Due to space limitations we have taken this part out, and the issue is instead pursued in further detail in Marín and Sucarrat (2011).

[8] The methods of White (2000), Hansen (2005), and Romano and Wolf (2005) are examples of approaches that seek to control the FWE.

average irrelevance proportion $M(\hat{k}_1/k_1)$, which is analogous to statistical size in a hypothesis testing context and which Doornik (2009) terms "gauge". These two statistics can be viewed as a more detailed characterisation of the expected value of the False Discovery Proportion (FDP), see Romano et al. (2008).

3.1 A comparison of multi-path GETS algorithms

Three multi-path GETS specification search algorithms have previously been studied in the academic literature: The algorithm of Hoover and Perez (1999), henceforth HP, the PcGets algorithm of Hendry and Krolzig (1999, 2001, 2005), and the Autometrics algorithm (Doornik and Hendry 2007, Doornik 2009). The way our algorithm undertakes multi-path GETS specification search in stages 1 and 2 is essentially a straightforward improvement of the HP algorithm. But in order to distinguish our algorithm from that of Hoover and Perez we will refer to our algorithm as AutoSEARCH.[9] The purpose of this subsection is to compare the properties of AutoSEARCH with those of HP, PcGets and Autometrics. The latter three have all been developed for and studied in the modelling of a mean specification with homoscedastic errors, so the simulations in this subsection will exclusively focus on modelling the mean under the assumption of constant variance. In this case, AutoSEARCH proceeds as follows:

Step 1. Check whether a general unrestricted mean GUM (MGUM) of the form (1) produces serially uncorrelated residuals free from ARCH. By assumption, $k_0 \geq 0$ of the regressors are relevant, $k_1 \geq 0$ are irrelevant and the total number of regressors k is given by $k_0 + k_1 + 1 = k$. The "+1" is due to the constant, which is restricted from removal in the simulations of AutoSEARCH.

Step 2. If the MGUM passes the diagnostic tests, then define the number of paths to be searched as equal to the number of insignificant variables in the GUM. In other words, just like PcGets, AutoSEARCH is not restricted to a maximum of ten paths as in the HP algorithm. The first insignificant variable constitutes the first variable to be removed in path 1, the second insignificant variable constitutes the first variable to be removed in path 2, and so on.

Step 3. After removal of the first variable in a path, subsequent simplification in each path is undertaken using "single-path" GETS search, where the regressor with highest p-value is sought deleted at each simplification. For each removal the standardised residuals are checked for serial correlation and ARCH using a Bonferroni correction,[10] and by a backtest (BaT) against the GUM. If removal induces either autocorrelation or heteroscedasticity (or both), or if removal does not

[9] We intend to make the code developed for this paper freely available as an (open source) R package with the name AutoSEARCH, see Sucarrat (2010).

[10] For example, if an overall nominal level of 5% is chosen for the diagnostic tests, then the autocorrelation and ARCH tests are each checked for significance using a level equal to the chosen

pass the BaT against the GUM, then the variable is re-included and subsequently restricted from removal in the simplification search in that path (but the variable is not restricted from removal in other paths). Simplification along the current path ends when there are no more insignificant variables, or if deletion of any of the insignificant variables does not pass the BaT against the GUM, or if one or more of the diagnostic tests fail. The single-path GETS search is undertaken for each of path.

Step 4. Form a list of models that contains the distinct terminal models of the search in steps 2 and 3. The GUM is always included in the list in order to ensure that the list is never empty.

Step 5. Select the best model from the list according to an information criterion (Schwarz is used in the simulations) that is computed using the Normal log-likelihood of the standardised residuals.

The AutoSEARCH algorithm can be viewed as a modified version of the HP algorithm of Hoover and Perez (1999), and the most important differences between the two algorithms are two. First, the HP algorithm is restricted to search a maximum of 10 paths, because this—in Hoover and Perez's view—resembled what users of the GETS methodology did in practice (prior to the existence of multi-path GETS specification search software). By contrast, just like in PcGets the number of paths in the AutoSEARCH algorithm is not limited to 10, but to the number of insignificant variables in the GUM (as in PcGets). This change improves the ability to detect relevant variables. The second important difference compared with HP concern which and how many diagnostic checks that are undertaken at each simplification.

The first main difference between the PcGets algorithm of Hendry and Krolzig (2005) on the one hand and the HP and AutoSEARCH algorithms on the other, is that PcGets is a "multi-round" algorithm, whereas HP and AutoSEARCH are "single-round" algorithms. Whereas HP and AutoSEARCH select between models from a first-round multi-path GETS simplification search, PcGets goes on to do further rounds if more than one model results from the first round. Starting from a GUM made up of the union of the models from the first round, PcGets continues the multi-round search until the resulting GUM does not change anymore. The main effect of multi-round search is an increased ability to retain relevant variables. However, it does to some extent come at the cost of excluding irrelevant variables. The Autometrics algorithm of Doornik (2009) is also a multi-round algorithm, and the basic principles are similar to those of PcGets. However, Autometrics searches more paths than PcGets by means of a "tree search" procedure.

In order to compare AutoSEARCH with HP, PcGets and Autometrics, we study AutoSEARCH in Monte Carlo experiments that have previously been run for the

nominal level divided by the number of diagnostic tests. Here, the number of tests is two and so the Bonferroni adjusted level is 2.5%. Simulations (not reported) suggests the Bonferroni correction is appropriate as long as the sample size is greater than 50.

three other algorithms. In Table 1 the experiments are labelled HP1, HP2' and HP7', and the results are contained in Table 2. The first important feature that emerges from the result is that AutoSEARCH is correctly calibrated in the sense that the average irrelevance proportion $M(\hat{k}_1/k_1)$ is approximately equal to the regressor significance level (5%) used in the simulations. This is the case for all three experiments. The second important feature of the results is that AutoSEARCH compares well overall in deleting irrelevant variables. In experiment HP1 where none of the regressors matter, AutoSEARCH recovers the DGP about 24% of time, which is lower than HP and PcGets. However, in HP2' and HP7' the values of $\hat{p}(DGP)$ are higher than those of HP and Autometrics. Admittedly though they are not as high as those of PcGets. A third important feature of the simulation results is that AutoSEARCH performs as well as the other algorithms in retaining relevant variables as measured by the average relevance proportion $M(\hat{k}_0/k_0)$. It should be pointed out though that in experiments HP2' and HP7' the signal of the variables that matter is relatively high. So possibly a different experimental design is needed in order to provide a more accurate comparison of the relative potency of the algorithms.

3.2 Multi-path GETS of the mean with heteroscedastic errors

When modelling financial returns, the errors $\{\epsilon_t\}$ of the mean specification very often remain heteroscedastic even after including explanatory information in the mean specification. So it is of interest to study the properties of multi-path GETS when the $\{\epsilon_t\}$ are heteroscedastic.[11] In doing so, we modify Steps 1-3 in our multi-path algorithm from the previous subsection in two straightforward ways. First, we use the White (1980) variance-covariance matrix for the coefficient test-statistics instead of the ordinary matrix.[12] Second, we turn off ARCH diagnostic checking and designate all the diagnostic checking significance level (5%) to the test for serial correlation.

The results of four Monte Carlo experiments, all with a reasonably persistent log-GARCH(1,1) specification on the errors of the mean $\{\epsilon_t\}$, are contained in Table 3. The first two experiments, HP1* and HP2'*, are essentially equal to HP1 and HP2' but for the heteroscedastic errors $\{\epsilon_t\}$.[13] At first sight the results of HP1* and HP2'* are not very encouraging, since the irrelevance proportion is about 15-16%,

[11] To our knowledge, no one has studied the properties of multi-path GETS specification search of the mean when the errors of the mean $\{\epsilon_t\}$ are heteroscedastic.

[12] There are several other heteroscedasticity consistent variance-covariance matrices that one might consider instead, and an interesting research question is to compare their relative merits. This we leave for future research, and here we choose White (1980) simply because it is the most common one in econometric software.

[13] In HP1* and HP2'* the value of α_0 has been calibrated. Specifically, the limit of the unconditional variance of a model with log-GARCH(1,1) specification $\log \sigma_t^2 = \alpha_0 + \alpha_1 \log \epsilon_{t-1}^2 + \beta_1 \log \sigma_{t-1}^2$

the triple of the nominal level. We do not know whether this is due to the—in a financial context—relatively small sample, the nature of the experiment, both, or some other reason(s). More investigation is needed. Nevertheless, one could argue that the results of experiments HP1* and HP2'* are not very indicative of how multi-path GETS actually performs in modelling financial returns at any rate. Firstly, the characteristics of the explanatory variables of the Hoover and Perez (1999) experiments do not correspond to the typical characteristics of explanatory variables in a financial context. Secondly, the DGP in HP2'* is a very persistent AR(1) process with an AR(1) coefficient as high as 0.75. In empirical finance by contrast the AR(1) coefficient is typically close to zero and rarely higher than 0.1 in absolute value. Finally, in finance one may argue that a sample size of 139 observations is actually relatively small. These reasons motivate two additional Monte Carlo experiments, namely SE1 and SE2, which better correspond to a modelling situation of financial returns. In the DGP of SE1 there are no relevant regressors, whereas in SE2 there is one, namely an AR(1) term with coefficient equal to 0.1. The results of these two experiments are much more encouraging, since the irrelevance proportion is approximately equal to the nominal level of 5%. Moreover, in experiment SE1 (where no regressors matter) AutoSEARCH recovers the DGP with a probability of about 0.60, whereas in SE2 both the relevance proportion and $\hat{p}(DGP)$ are relatively high in large samples. All in all, then, multi-path GETS works better for financial data than the HP1* and HP2* experiments suggests, since SE1 and SE2 better reflect the sample sizes and characteristics of financial modelling situations.

3.3 Multi-path GETS of the log-volatility specification

The purpose of this subsection is to study the properties of AutoSEARCH in modelling the log-volatility specification. Specification search of the volatility proceeds in a similar way to specification search of the mean, but for one difference. AutoSEARCH undertakes diagnostic checks of the standardised residuals $\{\hat{z}_t\}$ instead of the residuals of the AR-representation of $\log \sigma_t^2$, see Sucarrat and Escribano (2010) for the relation between log-ARCH specifications and their AR-representations. Table 3 contains the simulation results of the two experiments SE3 and SE4. In

is

$$\exp\left(\frac{\alpha_0}{1-\alpha_1-\beta_1}\right) \cdot \lim_{Q\to\infty} \prod_{q=1}^{Q} E\left[z_{t-q}^{2\alpha_1(\alpha_1+\beta_1)^{q-1}}\right].$$

Numerical simulation suggests the limit of the power term is approximately equal to 0.3167 when $z_t \sim N(0,1)$ for $\alpha_1 = 0.1$ and $\beta_1 = 0.8$. In order to calibrate the DGP in HP1* such that its unconditional variance equals the constant variance of the DGP in HP1, we thus need to solve

$$130^2 = \exp\left(\frac{\alpha_0}{1-\alpha_1-\beta_1}\right) \cdot 0.3167$$

for α_0, which yields $\alpha_0 \approx 1.0885$. Similarly, in HP2'* we obtain $\alpha_0 \approx 1.0058$.

experiment SE3 the simulation DGP contains no relevant variables, whereas in SE4 the simulation DGP contains a single relevant variable, a log-ARCH(1) term, with $\alpha_1 = 0.2$. These two experiments start from a GUM that contains five log-ARCH terms, an asymmetry term, the contemporaneous and lagged variable of an unrelated but strongly persistent AR(1) process, two standard normal IID processes and two exponentially distributed IID processes with shape parameter equal to 1. The strongly persistent process may be interpreted as a volatility proxy (say, the level of trading volume or a high-frequency based measure), the normally distributed variables may be interpreted as proxying "short-term" relative changes in, say, trading volume, whereas the exponentially distributed variables may be interpreted as proxying the square of, say, stock market return, interest rate changes, or similarly.

The results of experiments SE3 and SE4 are contained in Table 3. The first main property is that the irrelevance proportion is remarkably stable (between 4.3% and 6.%) and close to the nominal level of 5% across the two experiments. Interestingly, the property is robust both across sample size and across the tail-thickness (determined by the shape parameter τ in the Generalised Error Distribution (GED)) of the standardised error. This is a very useful finding since it effectively means that one in empirical practice does not have to worry about the impact of these features on the irrelevance proportion. A second main property of the simulations is that the relevance proportion is reasonably high in experiment SE4, and that it tends to 1 as the sample size increases. Fatter tails than the normal reduces the relevance proportion, but not much since the difference reaches a maximum of 6 percentage points. Finally, a third property is that $\hat{p}(DGP)$ seems to depend on sample size when regressors matter, but not when no regressor matters.

3.4 A more conservative GETS algorithm for finance

Financial markets are notoriously difficult to predict *ex ante*, so falsely suggesting predictive or explanatory power by retaining irrelevant variables may potentially lead to huge losses and substantial systemic damage if it is used to guide investment or policy decisions. To achieve this one could simply reduce the regressor significance level α. However, this will affect the relevance proportion negatively, so one could also consider other modifications to the GETS algorithm that makes it more conservative, that is likely to have a less negative impact on the relevance proportion when variables matter. One such straightforward modification is simply to include the empty model in the list of terminal models (in Step 4) as long as it passes the diagnostic tests, irrespective of whether the multi-path GETS specification search finds the empty model or not (in Steps 1-3), and irrespective of whether the empty model passes the BaT against the GUM or not. We will refer to this modified algorithm as "GETS with empty" or "GETS w/empty" for short. Above, by contrast, AutoSEARCH included the empty model only if it was a proper terminal in Steps 1-3, and if it passed the BaT against the GUM.

The properties of GETS w/empty when $T \to \infty$ are straightforward when the empty model is the DGP, since then the probability of the empty model being included among the terminal models is $(1-a)$, where a is the overall asymptotic diagnostic significance level. Accordingly, $\lim_{T\to\infty}\hat{p}(DGP) = (1-a)$ and the irrelevance proportion will tend to a value that is either lower than or equal to the regressor significance level: $\lim_{T\to\infty} M(\hat{k}_1/k_1) \leq$ regressor significance level. In other words, the regressor significance level now becomes an upper bound rather than the target of the irrelevance proportion. When the DGP is not equal to the empty model, then the irrelevance proportion will also tend to a value that is either lower than or equal to the regressor significance level. But the effect on the relevance proportion and on $\hat{p}(DGP)$ is uncertain, since it will depend on the exact nature of the DGP. Nevertheless, it seems reasonable to conjecture that the relevance proportion is more likely to fall rather than to stay unchanged or rise. Overall, then, the consequence of the minor modification is that GETS w/empty becomes a more conservative algorithm than PcGets and Autometrics who start with a BaT of the empty model against the GUM: The empty model may not end up among the terminals if it does not parsimoniously encompass the GUM and the other terminals. So GETS w/empty is a more conservative algorithm under the same regressor and diagnostic significance levels.

Table 4 contains the results of applying GETS w/empty (implemented by modifying AutoSEARCH) on our previous experiments. The results suggest indeed that, when the DGP is empty, then $\hat{p}(DGP)$ increases substantially. The most dramatic improvement occurs in HP1* where $\hat{p}(DGP)$ goes up from 2% to 78%. Moreover, the irrelevance proportion falls from a value of 16%—which is three times higher than and incompatible with the value predicted by the underlying statistical theory, to a new value of 3.2% that is compatible with and within the interval predicted by theory. In the other experiments where the DGP is equal to the empty model, the increase in $\hat{p}(DGP)$ varies from about 15% points to about 35% points, and the fall in irrelevance proportion varies from about 0% points to about 4.5 % points. When the DGP is not equal to the empty model, then the irrelevance proportion either remains close to the theoretical target of 5%, or drops about 1-3% points. The only exception is the curious HP2'* whose irrelevance proportion remains unchanged at about 16%. As for the relevance proportion, we see indeed a fall as we conjectured. In HP2', HP2'* and HP7' there is no change, but in these experiments the signal is too strong to be very informative. The results of experiments SE2 and SE4 are more indicative of the properties of GETS w/empty, and in these experiments the relevance proportion either remains about the same (SE2) or falls substantially (SE4). As for $\hat{p}(DGP)$, it remains generally about the same but for experiment SE4, where it falls substantially in smaller samples.

4 Empirical applications

In this section we illustrate and assess the methods and algorithm further through two empirical applications.

4.1 Explanatory modelling of the mean and variance

In the empirical illustration of Sucarrat (2009) an explanatory model of weekly exchange rate return with homoscedastic errors outperforms standard volatility models (including realised volatility) in predicting variability *ex post*, both in-sample and out-of-sample. The principal source of the unusually high explanatory power (about 40% in terms of R^2) is forward order flow. Here, we revisit the model by extending the data sample and by adding more explanatory variables to the GUM, both in the mean and volatility specifications. The overall GUM we start from contains 14 deletable regressors in the mean and 14 in the volatility specification. Next, we undertake multi-path GETS w/empty specification search by means of the AutoSEARCH algorithm. In other words, we use the conservative configuration explored in subsection 3.4.

The overall GUM we start from is (p-values in square brackets)

$$\begin{aligned}
r_t &= \psi_0 + \psi_1 r_{t-1} + \psi_2 r_{t-1}^2 + \psi_3 \Delta x_t + \psi_4 \Delta x_{t-1} + \psi_5 \Delta(ir_t^{no} - ir_t^{eu}) \\
&\quad + \psi_6 \Delta(ir_{t-1}^{no} - ir_{t-1}^{eu}) + \psi_7 \Delta oilp_t + \psi_8 \Delta oilp_{t-1} + \psi_9 \Delta ose_t + \psi_{10} \Delta ose_{t-1} \\
&\quad + \psi_{11} \Delta sp100_t + \psi_{12} \Delta sp100_{t-1} + \psi_{13} \widehat{ECM}_{t-1} + \epsilon_t, \\
\widehat{ECM}_t &= s_t - 2.16 - 0.03 ir_t^{no} + 0.06 ir_t^{eu}, \\
\epsilon_t &= \sigma_t z_t, \qquad z_t \sim IID(0,1), \\
\log \sigma_t^2 &= \alpha_0 + \sum_{p=1}^{5} \alpha_p \log \epsilon_{t-p}^2 + \lambda(\log|\epsilon_{t-1}|^2) I_{z_{t-1}<0} + \omega_0 \log EqWMA(8)_{t-1} \\
&\quad + \omega_1 \Delta v_t + \omega_2 v_{t-1} + \omega_3 (\Delta ir_t^{no})^2 + \omega_4 (\Delta ir_t^{eu})^2 + \omega_5 (\Delta oilp_t)^2 + \omega_6 (\Delta ose_t)^2 \\
&\quad + \omega_7 (\Delta sp100_t)^2
\end{aligned}$$

$$R^2 : \ 0.35 \qquad AR_{1-2} : \ \underset{[0.45]}{1.60} \qquad ARCH_{1-6} : \ \underset{[0.96]}{1.55} \qquad JB : \ \underset{[0.03]}{7.37} \qquad T = 190$$

The sample goes from 2 October 2005 to 5 July 2009 (197 end-of-week observations), where $r_t = (s_t - s_{t-1}) \times 100$ is log-return in % of the NOK/EUR exchange rate (an increase means the NOK depreciates), Δx_t is the associated forward order flow (buy initiated volume - sell initiated volume) in billions of NOK, ir_t^{no} is the 1-week Norwegian interbank money market interest rate in %, ir_t^{eu} is the 1-week

Eurozone interbank money market interest rate in %, $oilp_t$ is the log of oilprice, ose_t is the log of the main stock market index at the Oslo stock exchange, $sp100_t$ is the log of the Standard and Poor's 100 index of US stocks, $\widehat{ECM}_t$ is the estimate of an error correction model, $EqWMA(8)_t$ is an eight week equally weighted moving average of the squared residuals $\{\hat{\epsilon}_t^2\}$, v_t is the log of weekly forward volume (buy initiated volume + sell initiated volume), AR_{1-2} is a Ljung and Box (1979) test for serial correlation in the standardised residuals up to the 2nd. order, $ARCH_{1-6}$ is a Ljung and Box (1979) test for serial correlation up to the sixth order in the squared standardised residuals, JB is the Jarque and Bera (1980) test for non-normality, and T is the number of observations used in the estimation.[14] GETS specification search of the mean is undertaken while holding the volatility specification fixed, with the constant in the mean restricted from deletion. Two diagnostic checks of the standardised residuals are undertaken after each regressor removal, AR_{1-2} and $ARCH_{1-6}$, both at 2.5%, and the Schwarz information criterion (made up of a Gaussian log-likelihood in the residuals of the mean) is used as tie-breaker. The search yields four terminal models, the most parsimonious one yielding the smallest information criterion and thus constituting the specific mean specification. Next, GETS specification search of the volatility specification while holding the mean specification fixed and the volatility constant excluded from deletion, yields three terminal volatility specifications. Again, two diagnostic checks of the standardised residuals are undertaken after each regressor removal, AR_{1-2} and $ARCH_{1-6}$, both at 2.5%. Also, the Schwarz information criterion is used as tie-breaker, but this time the Gaussian log-likelihood is computed in the standardised residuals.[15] The overall specific model that we obtain is (t-ratios in parentheses and p-values in square brackets):

$$\begin{aligned}
\hat{r}_t &= 0.019 + \underset{(2.37)}{0.180} r_{t-1} - \underset{(-2.31)}{0.033} \Delta x_{t-1} - \underset{(-6.20)}{0.099} \Delta oilp_t - \underset{(-2.56)}{6.387} \widehat{ECM}_{t-1}, \\
\log \hat{\sigma}_t^2 &= -3.765 + \underset{(2.86)}{1.439} \Delta v_t + \underset{(2.43)}{0.886} v_{t-1}
\end{aligned}$$

$$R^2:\ 0.27 \quad AR_{1-2}:\ \underset{[0.76]}{0.56} \quad ARCH_{1-6}:\ \underset{[0.92]}{2.00} \quad JB:\ \underset{[0.06]}{5.59} \quad T = 190$$

The mean specification differs in two ways compared with Sucarrat (2009). The lagged order flow Δx_{t-1} is retained instead of the contemporaneous order flow, and the contemporaneous relative change in oil price is now found to be significant.

[14] The rawdata of s_t, ir_t^{no}, ir_t^{eu}, $oilp_t$, ose_t and $sp100_t$ are the daily series ew:nor19101, ew:nor14265, ew:emu14313, ew:com20220, ew:nor15565 and ew:usa15100200, respectively, from Reuters - EcoWin. The Norwegian order flow data are from Norges Bank (the Norwegian central bank) and can be downloaded via the url `http://www.norges-bank.no/templates/reportroot____60389.aspx`. The data are described in more detail in Meyer and Skjelvik (2006).

[15] This resulted in a situation where two specifications attained the lowest value on the information criterion. In order to choose among them we simply selected the one that we found more interesting from an economic point of view.

Further investigation suggests the most important reasons for the difference is that the impact of contemporaneous forward order flow changed in the course of the financial events of 2008-2009, and that both the oil price and the value of the NOK fell substantially over the same period. The specific volatility specification contains two terms: Changes in market activity as measured by the relative change in forward volume, and the lagged (log-)level of volume. The latter is strongly autocorrelated and may thus be the source of any possible ARCH. Finally, the Jarque and Bera (1980) test suggest there is a slight departure of normality in the standardised residuals.

4.2 How well do volatility proxies forecast variability?

Volatility is by definition a conditional forecast of price variability when the conditional mean is zero, and a common economic interpretation of a mean equal to zero is that the direction of the financial price change is unpredictable. This explains the importance of volatility forecasts in derivative pricing. Indeed, volatility forecasts are arguably the most important inputs in derivative pricing, and so volatility forecasting is of great importance in the financial industry.

The volatility forecasting literature has experienced major developments over the last decade or so. One of the developments is the increased production, dispersion and availability of high-frequency data, and the increased and cheaper computing power to handle the larger datasets. A second development of great importance is theoretical. The last ten years have witnessed many theoretical contributions that enables efficient volatility forecasting by making use of high-frequency data. The most well-known of the estimators is realised volatility (RV, sums of squared intra-period high-frequency returns), but numerous relatives have also been proposed and studied. How well do all these volatility proxies actually forecast price variability? If the underlying continuous time model is a valid or “true” representation of the DGP in some appropriate sense—this is effectively the assumption that RV and its cousins rely upon, then this has three important implications. First, the standardised residuals defined as $\hat{z}_t = r_t/\sqrt{RV}_t$ should be serially uncorrelated and exhibit no ARCH. Second, the coefficient restrictions $\alpha_0 = 0$ and $\omega = 1$ in the SEARCH specification $\log \sigma_t^2 = \alpha_0 + \omega \log RV_t$ should not be rejected. Third, RV_t should parsimoniously encompass models that make use of the same data. If it does not, then this means the other models make more efficient use of the data.

The first two implications are readily investigated via logarithmic Mincer-Zarnowitz regressions (MZ), which amounts to fitting

$$\log \sigma_t^2 = \alpha_0 + \omega \log RV_t. \tag{4}$$

Hansen and Lunde (2006) and Patton and Sheppard (2009) have argued against the use of logarithmic Mincer and Zarnowitz (1969) regressions. However, the problems they point to are essentially resolved by the results in Sucarrat and Escribano (2010).

Next, the hypotheses of no serial correlation and ARCH in the standardised error, and whether $\alpha_0 = 0$ and $\omega = 1$, can readily be tested. Table 5 contains logarithmic MZ-regressions of daily stock return (IBM) on three different volatility proxies. This data series is of interest because Patton (2011) uses them to illustrate how volatility proxies can improve volatility forecast evaluation. However, Table 5 shows that all the three proxies invalidate the hypothesis of no ARCH in the standardised residuals, and at 6% percent or lower the joint coefficient restrictions $\alpha_0 = 0$ and $\omega = 1$ are rejected for all three proxies. In other words, the basic diagnostic tests and the coefficient restriction tests do not convincingly suggest that the theory upon which the volatility proxies is based on holds empirically. The candidate that comes closest to satisfying the basic diagnostics is the third proxy, that is, RV made up of 5-minute intra-day returns.

Whether a volatility proxy parsimoniously encompasses other models that make use of the same data is readily investigated by means of automated multi-path GETS modelling. Table 6 contains the results of an analysis for the third volatility proxy. MGUM is the general and unrestricted mean specification, whereas VGUM1 and VGUM2 are two different volatility GUMs. VGUM1 contains only the constant and $\log RV_t^{5m}$ as regressor, and the ARCH diagnostic test suggests $\log RV_t^{5m}$ does not capture all the volatility persistence given the MGUM. This motivates VGUM2 where we add log-ARCH lags and an asymmetry term to the VGUM. This improves the ARCH diagnostics. MSPEC and VSPEC are the specifications obtained after multi-path GETS specification search of MGUM and VGUM2: First MSPEC is obtained by holding the volatility specification fixed and equal to VGUM2, and next VSPEC is obtained by holding the mean specification fixed and equal to MSPEC. In the specification search of the mean the standardised residuals are checked after each deletion for serial correlation up to the 3rd. order, and the squared standardised residuals are checked for serial correlation up to the 5th. order. Schwarz's information criterion, computed in terms of a Gaussian log-likelihood made up of the residuals of the mean, is used as tie-breaker for terminal models in the mean. The Schwarz criterion is also used as a tie-breaker between log-volatility specifications. However, in this case the standardised residuals are used for the Gaussian log-likelihood. The constants in both the mean and log-volatility specifications are restricted from deletion during the specification search. The conclusion is that RV_t^{5m} does not parsimoniously encompass all the other hypothesised effects, since three day-of-the week dummies are retained in addition to $\log RV_t^{5m}$ in the log-variance specification. Indeed, the estimates suggest that there are some substantial periodicity effects (volatility lower on Wednesday, Thursday and Friday) that are not accounted for by realised volatility.

5 Conclusions

By making use of the recent results in Sucarrat and Escribano (2010), we have proposed methods and algorithms that resolves many of the problems earlier faced in the implementation of automated multi-path General-to-Specific (GETS) specification search of financial models. The simulations and empirical applications suggest the methods can be of great value in financial practice for the following reasons: First, our simulations show that the GETS algorithm we propose compares well with the other multi-path GETS algorithms that are currently available. The irrelevance proportion is generally equal to the nominal regressor level across experiments (the exceptions are in experiments of little relevance for finance), sample sizes and density shapes, and the relevance proportions are sufficiently high. Second, the slight modification to multi-path GETS modelling that we propose for finance in order to make GETS model selection more conservative, improves substantially the probability to recover the DGP when it is equal to the empty model (the cost is that it affects the relevance proportion negatively only when variables are marginally significant), and reduces the irrelevance proportion to a value that is lower than the nominal regressor significance level. This increased capacity to delete irrelevant variables is particularly desirable for financial economics and business finance. Third, in our first empirical application it takes us only a couple of seconds on an ordinary computer to undertake multi-path GETS modelling of a Stochastic Exponential ARCH (SEARCH) model with a total of 28 deletable regressors in the mean and volatility specifications. By contrast, automated multi-path GETS specification search of the GARCH counterpart of our SEARCH model, with joint ML estimation of the mean, volatility and density of the standardised errors, may not be feasible in practice, and would in any case require substantial effort and time by the modeller, in addition to numerous subjective decisions throughout the modelling process about starting values, convergence criteria, multiple optima, sensible estimates and so on. Fourth, our second empirical application show that our methods can be of great use in both evaluating the forecasts of volatility proxies, and in improving them. Nevertheless, there is still room for further improvement, generalisation and exploration of the methods and algorithms we have proposed in this paper. For example, the efficiency of the estimation procedures may be improved through (say) feasible generalised least squares (FGLS) procedures and/or iterative least squares procedures, and common outlier detection algorithms (developed for the identification of outliers in the mean) can readily be adapted to search for Bernoulli jumps in the log-volatility specification.

References

Bauwens, L. and G. Sucarrat (2010). General to Specific Modelling of Exchange Rate Volatility: A Forecast Evaluation. *International Journal of Forecasting 26*, 885–907.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics 31*, 307–327.

Campos, J., D. F. Hendry, and N. R. Ericsson (Eds.) (2005). *General-to-Specific Modeling. Volumes 1 and 2.* Cheltenham: Edward Elgar Publishing.

Doornik, J. (2008). Encompassing and Automatic Model Selection. *Oxford Bulletin of Economics and Statistics 70*, 915–925.

Doornik, J. (2009). Autometrics. In J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry.* Oxford: Oxford University Press.

Doornik, J. A. and D. F. Hendry (2007). *Empirical Econometric Modelling - PcGive 12: Volume I.* London: Timberlake Consultants Ltd.

Engle, R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflations. *Econometrica 50*, 987–1008.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance 48*, 1779–1801.

Granger, C. W. and A. Timmermann (1999). Data mining with local model specification uncertainty: a discussion of Hoover and Perez. *Econometrics Journal 2*, 220–225.

Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business and Economic Statistics 23*, 365–380.

Hansen, P. R. and A. Lunde (2006). Consistent ranking of volatility models. *Journal of Econometrics 131*, 97–121.

Hendry, D. F. and H.-M. Krolzig (1999). Improving on 'Data Mining Reconsidered' by K.D. Hoover and S.J. Perez. *Econometrics Journal 2*, 202–219.

Hendry, D. F. and H.-M. Krolzig (2001). *Automatic Econometric Model Selection using PcGets.* London: Timberlake Consultants Press.

Hendry, D. F. and H.-M. Krolzig (2005). The Properties of Automatic Gets Modelling. *Economic Journal 115*, C32–C61.

Hoover, K. D. and S. J. Perez (1999). Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search. *Econometrics Journal 2*, 167–191. Dataset and code: `http://www.csus.edu/indiv/p/perezs/Data/data.htm`.

Jarque, C. and A. Bera (1980). Efficient Tests for Normality, Homoskedasticity, and Serial Independence. *Economics Letters 6*, pp. 255–259.

Krolzig, H.-M. (2003). General-to-Specific Model Selection Procedures for Structural Vector Autoregressions. *Oxford Bulletin of Economics and Statistics 65*, 803–819.

Ljung, G. and G. Box (1979). On a Measure of Lack of Fit in Time Series Models. *Biometrika 66*, 265–270.

Marín, J. M. and G. Sucarrat (2011). Modelling the Skewed Exponential Power Distribution in Finance. In C. Perna and M. Sibill (Eds.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance*. Springer, in press.

McAleer, M. (2005). Automated inference and learning in modeling financial volatility. *Econometric Theory 21*, 232–261.

Meyer, E. and J. Skjelvik (2006). Statistics on foreign exchange transactions — new insight into foreign exchange markets. *Norges Bank Economic Bulletin* (2/06), 80–88.

Mincer, J. and V. Zarnowitz (1969). The Evaluation of Economic Forecasts. In J. Zarnowitz (Ed.), *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Mizon, G. (1995). Progressive Modeling of Macroeconomic Time Series: The LSE Methodology. In K. D. Hoover (Ed.), *Macroeconometrics. Developments, Tensions and Prospects*. Kluwer Academic Publishers.

Pagan, A. and G. Schwert (1990). Alternative models for conditional volatility. *Journal of Econometrics 45*, 267–290.

Patton, A. J. (2011). Volatility Forecast Evaluation and Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics 160*, 246–256. Code and data: `http://econ.duke.edu/~ap172/Patton_robust_loss_apr06.zip`.

Patton, A. J. and K. Sheppard (2009). Evaluating Volatility and Correlation Forecasts. In T. Andersen, R. Davies, J. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*. Berlin: Springer Verlag.

Romano, J. P., A. Shaikh, and M. Wolf (2008). Formalized data snooping based on generalized error rates. *Econometric Theory 24*, 404–447.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*, 1237–1282.

Sucarrat, G. (2009). Forecast Evaluation of Explanatory Models of Financial Variability. *Economics – The Open-Access, Open-Assessment E-Journal 3*. `http://www.economics-ejournal.org/economics/journalarticles/2009-8`.

Sucarrat, G. (2010). AutoSEARCH: An R Package for Automated Financial Modelling. `http://www.sucarrat.net/`.

Sucarrat, G. and Á. Escribano (2010). The Power Log-GARCH Model. `http://www.sucarrat.net/`.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica 48*, 817–838.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica 68*, 1097–1126.

Table 1: List of experiments

Label	k_0	k_1	Simulation DGP	GUM
HP1	0	40	$r_t = \epsilon_t, \quad \epsilon_t = 130 z_t, \quad z_t \overset{IID}{\sim} N(0,1)$	$r_t = \sum_{k=1}^{41} \psi_k x_{kt}^{HP} + \epsilon_t$, where $x_{37t}^{HP} = r_{t-1}$, $x_{38t}^{HP} = r_{t-2}$, $x_{39t}^{HP} = r_{t-3}$, $x_{40t}^{HP} = r_{t-4}$ and $x_{41t}^{HP} = 1$
HP1*	0	40	$r_t = \epsilon_t, \quad \epsilon_t = \sigma_t z_t, \quad z_t \overset{IID}{\sim} N(0,1), \quad E(\epsilon_t^2) \approx 130^2$ $\log \sigma_t^2 = 1.0885 + 0.1 \log \epsilon_{t-1}^2 + 0.8 \log \sigma_{t-1}^2$	Same as in HP1
HP2'	1	39	$r_t = 0.75 r_{t-1} + \epsilon_t, \quad \epsilon_t = 85.99 z_t, \quad z_t \overset{IID}{\sim} N(0,1)$	Same as in HP1
HP2'*	1	39	$r_t = 0.75 r_{t-1} + \epsilon_t, \quad \epsilon_t = \sigma z_t, \quad z_t \overset{IID}{\sim} N(0,1)$, $E(\epsilon_t^2) \approx (85.99)^2, \quad \log \sigma_t^2 = 1.0058 + 0.1 \log \epsilon_{t-1}^2 + 0.8 \log \sigma_{t-1}^2$	Same as in HP1
HP7'	3	37	$r_t = 0.75 r_{t-1} + 1.33 x_{11t}^{HP} - 0.9975 x_{29t}^{HP} + \epsilon_t$, $\epsilon_t = 6.44 z_t, \quad z_t \overset{IID}{\sim} N(0,1)$	Same as in HP1
SE1	0	9	$r_t = \epsilon_{1t}, \quad \epsilon_{1t} = \sigma_t z_t, \quad z_t \overset{IID}{\sim} N(0,1)$ $\log \sigma_t^2 = 0.1 \log \epsilon_{1t-1}^2 + 0.8 \log \sigma_{t-1}^2$	$r_t = \phi_0 + \sum_{m=1}^{2} \phi_1 r_{t-m} + \sum_{n=0}^{2} \eta_n x_{t-n} + \eta_3 x_{1t}^{N(0,1)} + \eta_4 x_{2t}^{N(0,1)} + \eta_5 x_{1t}^{EXP(1)} + \eta_6 x_{2t}^{EXP(1)} + \epsilon_{1t}$, where $x_t = 0.1 + 0.9 x_{t-1} + \epsilon_{2t}$ with $\epsilon_{2t} \overset{IID}{\sim} N(0,1)$
SE2	1	8	$r_t = 0.1 r_{t-1} + \epsilon_{1t}, \quad \epsilon_{1t} = \sigma_t z_t, \quad z_t \overset{IID}{\sim} N(0,1)$ $\log \sigma_t^2 = 0.1 \log \epsilon_{1t-1}^2 + 0.8 \log \sigma_{t-1}^2$	Same as in SE1
SE3	0	12	$r_t = \epsilon_{1t}, \quad \epsilon_{1t} = \sigma_t z_t, \quad \sigma_t = 1, \quad z_t \overset{IID}{\sim} GED(\tau)$, $\tau \in \{1.1, 2\}$	$\log \sigma_t^2 = \alpha_0 + \sum_{p=1}^{P} \alpha_p \log \epsilon_{1t-p}^2 + \lambda_1 (\log \epsilon_{1t-1}^2) I_{z_{t-1}<0} + \omega_1 y_t + \omega_2 y_{t-1} + \omega_3 x_{1t}^{EXP(1)} + \omega_4 x_{2t}^{EXP(1)} + \omega_5 x_{3t}^{N(0,1)} + \omega_6 x_{4t}^{N(0,1)}$, where $y_t = 0.9 y_{t-1} + \epsilon_{2t}$ with $\epsilon_{2t} \overset{IID}{\sim} N(0,1)$
SE4	1	11	$r_t = \epsilon_{1t}, \quad \epsilon_{1t} = \sigma_t z_t, \quad z_t \overset{IID}{\sim} GED(\tau), \quad \tau \in \{1.1, 2\}$, $\log \sigma_t^2 = 0.2 \log \epsilon_{1t-1}^2$	Same as in SE3

The design of the experiments HP1, HP1*, HP2', HP2'* and HP7' are based on Hoover and Perez (1999, see Table 3 on p. 174), and make use of their data $x_{1t}^{HP}, \ldots, x_{36t}^{HP}$ (available via `http://www.csus.edu/indiv/p/perezs/Data/data.htm`). It should be noted that there are two typos in their Table 3. The value $\sqrt{(7/4)}$ should instead be $\sqrt{7}/4$ in the model of the autoregressive error, and the value 6.73 should instead be 6.44 in model 7', see also Doornik (2009). GED(τ) is short for Generalised Error Distribution with shape parameter τ, where the Normal is obtained when $\tau = 2$, whereas more (less) heavy-tailed densities are obtained when $\tau < 2$ ($\tau > 2$). In the simulations of the DGPs of HP1*, HP2', HP2'*, SE1, SE2 and SE4, a prior burn-in sample of 100 observations is discarded in each replication in order to handle the initial value issue. The number of relevant variables in the GUM is k_0, the number of irrelevant variables in the GUM is k_1 and the total number of variables (the constant included) in the GUM is $k = k_0 + k_1 + 1$.

Table 2: Comparison of GETS algorithms: Specification search in the mean with Gaussian homoscedastic errors $\{\epsilon_t\}$, using a nominal regressor significance level of 5%

Experiment	k_0	k_1	Algorithm	T	$M(\hat{k}_0/k_0)$	$M(\hat{k}_1/k_1)$	$\hat{p}(DGP)$
HP1	0	40	AutoSEARCH	139		0.049	0.239
			HP			0.045	0.292
			PcGets			≈ 0.04	≈ 0.45
HP2'	1	39	AutoSEARCH	139	1.000	0.050	0.252
			HP		1.000	0.107	0.000
			PcGets		≈ 0.97	≈ 0.05	≈ 0.32
			Autometrics		1.000	0.063	0.119
HP7'	3	37	AutoSEARCH	138	1.000	0.051	0.232
			HP		0.967	0.082	0.040
			PcGets		≈ 1.00	≈ 0.04	≈ 0.37
			Autometrics		0.999	0.066	0.111

Simulations of the AutoSEARCH algorithm are in R with 1000 replications. $M(\hat{k}_0/k_0)$ is the average proportion of relevant variables $\hat{k}_0$ retained relative to the actual number of relevant variables k_0 in the DGP. $M(\hat{k}_1/k_1)$ is the average proportion of irrelevant variables $\hat{k}_1$ retained relative to the actual number of irrelevant variables k_1 in the GUM. The estimate $\hat{k_1}$ includes both significant and insignificant retained irrelevant variables (this is in line with Hendry and Krolzig (2005), and Doornik (2009), but counter to HP which only includes significant irrelevant variables in the estimate). $\hat{p}(DGP)$ is the proportion of times the DGP is found exactly. The properties of the HP algorithm are from Hoover and Perez (1999, Table 4 on p. 179). The properties of the PcGets algorithm are from Hendry and Krolzig (2005, Figure 1 on p. C39), and the properties of the Autometrics algorithm are from Doornik (2009, section 6). For AutoSEARCH, a constant is included in all the regressions but ignored in the evaluation of successes and failures (this is in line with Hoover and Perez (1999) but counter to Hendry and Krolzig (2005), and Doornik (2009)), in the diagnostic checks both the AR and ARCH test of the standardised residuals were undertaken at lag 2 using a significance level of 2.5% for each, and as tiebreaker the Schwarz information criterion is used with a Gaussian log-likelihood made up of the standardised residuals of the mean specification.

Table 3: Properties of AutoSEARCH: Specification search in the mean with heteroscedastic errors $\{\epsilon_t\}$ and in the log-volatility specification, using a nominal regressor significance level of 5%

Experiment	DGP	k_0	k_1	T	τ	$M(\hat{k}_0/k_0)$	$M(\hat{k}_1/k_1)$	$\hat{p}(DGP)$
HP1*	Empty	0	40	139	2.0		0.160	0.015
HP2'*	AR(1)	1	39	139	2.0	1.000	0.156	0.021
SE1	Empty	0	9	139	2.0		0.065	0.591
				200	2.0		0.057	0.622
				500	2.0		0.053	0.615
				1000	2.0		0.058	0.613
SE2	AR(1)	1	8	139	2.0	0.217	0.065	0.147
				200	2.0	0.274	0.054	0.184
				500	2.0	0.549	0.048	0.379
				1000	2.0	0.821	0.049	0.567
SE3	Empty	0	12	139	2.0		0.043	0.625
					1.1		0.048	0.607
				200	2.0		0.047	0.612
					1.1		0.052	0.562
				500	2.0		0.044	0.624
					1.1		0.044	0.628
				1000	2.0		0.047	0.594
					1.1		0.050	0.573
SE4	Log-ARCH(1)	1	11	139	2.0	0.477	0.057	0.316
					1.1	0.425	0.063	0.267
				200	2.0	0.643	0.059	0.421
					1.1	0.583	0.063	0.355
				500	2.0	0.949	0.048	0.608
					1.1	0.947	0.048	0.615
				1000	2.0	1.000	0.045	0.638
					1.1	0.999	0.050	0.601

Simulations of the AutoSEARCH algorithm are in R with 1000 replications. In HP1*, HP2'*, SE1 and SE2 only one diagnostic check (AR) of the standardised residuals is undertaken at lag 2 using a significance level of 5%, and as tiebreaker the Schwarz information criterion is used with a Gaussian log-likelihood made up of the mean residuals $\{\hat{\epsilon}_t\}$. In SE3 and SE4 the AR and ARCH tests of the standardised residuals are undertaken at lag 2 using a nominal significance level of 2.5% for each, and as tiebreaker the Schwarz information criterion is used with a Gaussian log-likelihood made up of the standardised residuals $\{\hat{z}_t\}$.

Table 4: Properties of AutoSEARCH: Conservative GETS model selection, using a nominal regressor significance level of 5%

Experiment	DGP	k_0	k_1	T	$M(\hat{k}_0/k_0)$	$M(\hat{k}_1/k_1)$	$\hat{p}(DGP)$
HP1	Empty	0	40	139		0.032	0.470
HP2’	AR(1)	1	39	139	1.000	0.051	0.253
HP7’	AR(1) + $x^{HP}_{11t}, x^{HP}_{29t}$	3	37	138	1.000	0.052	0.240
HP1*	Empty	0	40	139		0.032	0.783
HP2’*	AR(1)	1	39	139	1.000	0.160	0.016
SE1	Empty	0	9	139		0.031	0.788
				200		0.034	0.767
				500		0.029	0.793
				1000		0.023	0.840
SE2	AR(1)	1	8	139	0.224	0.042	0.132
				200	0.287	0.039	0.212
				500	0.568	0.033	0.405
				1000	0.838	0.039	0.614
SE3	Empty	0	12	139		0.006	0.954
				200		0.004	0.955
				500		0.004	0.962
				1000		0.005	0.967
SE4	Log-ARCH(1)	1	11	139	0.205	0.024	0.141
				200	0.335	0.026	0.222
				500	0.863	0.041	0.536
				1000	0.998	0.048	0.603

Simulations of the AutoSEARCH algorithm are in R with 1000 replications. The standardised errors $\{z_t\}$ are $IIN(0,1)$ in all simulations (that is, $\tau = 2$).

Table 5: Logarithmic Mincer-Zarnowitz regressions of variability (squared return) on volatility proxies

Model	$\hat{\alpha}_0$ $[p{-}val]$	$\hat{\omega}$ $[p{-}val]$	$\chi^2(2)$ $[p{-}val]$	$AR(1)$ $[p{-}val]$	$ARCH(5)$ $[p{-}val]$	JB $[p{-}val]$
$\log\sigma_t^2 = \alpha_0 + \omega \log RV_t^{65m}$	-0.04 [0.85]	2.17 [0.00]	72.81 [0.00]	-0.04 [0.04]	0.06 [0.00]	6301.47 [0.00]
$\log\sigma_t^2 = \alpha_0 + \omega \log RV_t^{15m}$	-0.58 [0.03]	1.87 [0.00]	19.22 [0.00]	-0.01 [0.45]	0.11 [0.00]	53.80 [0.00]
$\log\sigma_t^2 = \alpha_0 + \omega \log RV_t^{5m}$	-0.64 [0.05]	1.55 [0.02]	5.69 [0.06]	-0.01 [0.76]	0.11 [0.00]	4.68 [0.10]

All computations in R. The estimates and tests are based on the assumptions that $r_t = \sigma_t z_t, z_t \sim IID(0,1)$ and $\log\sigma_t^2 = \alpha_0 + \omega \log RV_t^{(\cdot)}$, where r_t is daily IBM return 4 January 1993 - 31 December 2003 (2772 observations). The data are from Patton (2011), where RV_t^{65m}, RV_t^{15} and RV_t^{5m} are realised volatilities made up of 65-minute, 15-minute and 5-minute intra-day returns. The p-values in the $\hat{\alpha}_0$ and $\hat{\omega}$ columns are from Wald coefficient restriction tests of $\alpha_0 = 0$ and $\omega = 1$, respectively, whereas the p-values in the $\chi^2(2)$ column are from their joint test. The ordinary variance-covariance matrix is used for the Wald tests. $AR(1)$ is a Ljung and Box (1979) test of 1st. order serial correlation in the standardised residuals $\{\hat{z}_t\}$, $ARCH(5)$ is a Ljung and Box (1979) test of 5th. order serial correlation in the squared standardised residuals $\{\hat{z}_t^2\}$, and JB is the Jarque and Bera (1980) test for non-normality.

Table 6: Parsimonious encompassing test of RV_t^{5m}

$$\begin{aligned}
\text{MGUM:}\quad \underset{[p.val.]}{\hat{r}_t} &= \underset{[0.02]}{0.216} - \underset{[0.14]}{0.035} r_{t-1} - \underset{[0.20]}{0.029} r_{t-2} + \underset{[0.06]}{0.005} RV_{t-1}^{5m} + \underset{[0.62]}{0.002} RV_{t-2}^{5m} - \underset{[0.62]}{0.060} Tue_t \\
&\quad - \underset{[0.01]}{0.327} Wed_t - \underset{[0.27]}{0.152} Thu_t - \underset{[0.02]}{0.291} Fri_t \\
\text{VGUM 1:}\quad \underset{[p.val.]}{\log \hat{\sigma}_t^2} &= -\underset{[0.03]}{0.154} + \underset{[0.00]}{1.046} \log RV_t^{5m} \\
&\quad \underset{[p.val.]}{AR(3)} : \underset{[0.50]}{-0.01} \qquad \underset{[p.val.]}{ARCH(5)} : \underset{[0.01]}{0.05} \\
\text{VGUM 2:}\quad \underset{[p.val.]}{\log \hat{\sigma}_t^2} &= \underset{[0.31]}{0.133} + \underset{[1.00]}{0.000} \log \hat{\epsilon}_{t-1}^2 - \underset{[0.52]}{0.012} \log \hat{\epsilon}_{t-2}^2 - \underset{[0.41]}{0.015} \log \hat{\epsilon}_{t-3}^2 + \underset{[0.12]}{0.028} \log \hat{\epsilon}_{t-4}^2 \\
&\quad + \underset{[0.37]}{0.039} \log \hat{\epsilon}_{t-5}^2 - \underset{[0.93]}{0.003} (\log \hat{\epsilon}_{t-1}^2) \hat{I}_{\{\epsilon_{t-1}<0\}} - \underset{[0.45]}{0.051} \log EqWMA(20)_{t-1} \\
&\quad + \underset{[0.00]}{1.052} \log RV_t^{5m} - \underset{[0.33]}{0.130} Tue_t - \underset{[0.05]}{0.261} Wed_t - \underset{[0.01]}{0.333} Thu_t - \underset{[0.00]}{0.377} Fri_t \\
&\quad \underset{[p.val.]}{AR(3)} : \underset{[0.46]}{-0.01} \qquad \underset{[p.val.]}{ARCH(5)} : \underset{[0.78]}{-0.00} \\
\text{MSPEC:}\quad \underset{[p.val.]}{\hat{r}_t} &= \underset{[0.07]}{0.075} \\
\text{VSPEC:}\quad \underset{[p.val.]}{\log \hat{\sigma}_t^2} &= -\underset{[0.66]}{0.036} + \underset{[0.00]}{1.094} \log RV_t^{5m} - \underset{[0.03]}{0.237} Wed_t - \underset{[0.01]}{0.299} Thu_t - \underset{[0.00]}{0.320} Fri_t \\
&\quad \underset{[p.val.]}{AR(3)} : \underset{[0.19]}{-0.02} \qquad \underset{[p.val.]}{ARCH(5)} : \underset{[0.04]}{0.04}
\end{aligned}$$

All computations in R. MGUM is short for mean GUM, VGUM is short for variance GUM, MSPEC is short for specific mean specification, VSPEC is short for specific variance specification and Tue_t, Wed_t Thu_t and Fri_t are week of the day dummies. In the mean specifications White (1980) standard errors are used, whereas in the volatility specifications ordinary standard errors are used. $AR(3)$ is a Ljung and Box (1979) test of serial correlation in the standardised residuals $\{\hat{z}_t\}$ up to order 3 and $ARCH(5)$ is a Ljung and Box (1979) test of serial correlation in the squared standardised residuals $\{\hat{z}_t^2\}$ up to order 5.