

SEARCH FOR A STRUCTURAL SPECIFICATION OF THE EARNINGS-RETURNS RELATION

JORDAN G. MILEV

Keywords: earnings response coefficient, unexpected earnings, nonlinearity, data snooping, genetic programming algorithms

JEL classification: C51 - Model Construction and Estimation; G12 - Asset Pricing; M41 - Accounting

Draft. Comments welcome.

The author wishes to thank participants at the Yale University Summer Workshop in Economics and the 11th Symposium of the Society for Nonlinear Dynamics and Econometrics for helpful comments on earlier drafts. Permission to use I/B/E/S data provided by Thompson Financial is gratefully acknowledged.

ABSTRACT. The task of model identification is present in most applied econometric exercises. In the simplest case, researchers are faced with empirical data for two economic variables and have the task of formulating a structural model characterizing their relationship. The selection of the appropriate functional form rests with the researcher and has profound implications about the consistency and significance of estimated model parameters and about predictions obtained from it. I advocate the use of a flexible parametric estimation approach using a stochastically driven optimization algorithm. In contrast to traditional hill climbing algorithms that proceed from one point in the search space to another, the algorithm operates on a set of points in the parameter space.

I use the algorithm to model how stock prices react to unanticipated accounting earnings. More particularly, I suggest several nonlinear parametric specifications of the reaction of excess current-period stock price returns to the unexpected component of quarterly earnings. My motivation is the well-known misspecification problem in studies that ignore nonlinearity, as it presents a source of systematic error in motivation is the well-known misspecification problem in studies that ignore nonlinearity, as it presents a source of systematic error in earnings response regressions (see Freeman and Tse 1992 and Beneish and Harvey 1998, for a discussion of nonlinearity in this empirical context). Using the described model identification procedure I am able to confirm the existence of a nonlinear model that has superior in-sample and out-of-sample predictive power over the linear model. The approach presents an advantage over alternative nonparametric techniques as it can be used to derive a parametric model for further empirical work on the earnings-returns relation.

1. ESTIMATING EARNINGS RESPONSE REGRESSIONS

In this paper we derive several structural specifications of the relation between current-period unexpected accounting earnings and abnormal stock returns. The functional form of the link between these two variables has been a key issue in the accounting literature (where issues regarding the information content of earnings is of interest), and in the econometric literature (which has been more concerned with issues of specification and estimation). Our nonlinear specifications extend the traditional linear regression models usually employed to study the earnings-returns relationship. They suggest particular functional forms of the earnings-returns relation suitable for further empirical work. To accomplish our goal of model selection, we advocate the use of a procedure based on genetic algorithms (GP) to systematically seek out and identify a structural model for a nonlinear data generating process among candidate models from a well-defined (and vast) set of functions.

The task of model identification is present in most applied econometric exercises. In the simplest case, researchers are faced with empirical data for two economic variables and have the task of formulating a structural model characterizing their relationship. Until now, researchers have noted the shape of the observed relationship and have employed an *ad hoc* functional relationship having the desired derivative signs. Very often such models are good at in-sample prediction and fare poorly after further scrutiny reveals their out-of-sample performance. Nonparametric approaches also fare poorly in out-of-sample prediction. More importantly, the use of ad hoc model selection by researchers has an effect on our ability to draw inferences from the model. As Merton (1987, p. 107) puts it, “Is it reasonable to use the standard t-statistic as a valid measure of significance when the test is conducted on the

same data used by many earlier studies whose results influenced the choice of theory to be tested?”

We demonstrate the use of an algorithmic model selection procedure that enables researchers to identify a plausible data generating process from a given set of functions, with minimal researcher intervention beyond specifying the set of functions itself and the metric that needs to be maximized. We employ a variant of the test of data snooping recently developed by White (2000) to evaluate whether our models have better predictive power than the benchmark model currently accepted by the literature, after accounting for the degree of data mining performed by the genetic programming algorithm. With the new availability of such “data snooping” tests and the recent advances of computationally intensive search algorithms, we encourage future researchers to provide such statistics in order to allay possible criticisms. So far, we are aware of only one study where data snooping tests are conducted for a genetic programming search procedure (Beenstock and Szpiro, 2002).

Our approach to nonlinear model selection is based on the theory of genetic programming. We conduct a systematic stochastically driven search in a ex-ante specified (vast) set of functions for a functional form that best fits the returns-earnings relation. The search is optimal in a very specific sense. In contrast to traditional search algorithms which operate on one point at a time, GP operates on a set of points in parallel. It implicitly divides the search space into subspaces and by evaluating the fit of each candidate solution to the minimization problem, it implicitly evaluates the fitness of numerous sub-expressions representing the defined subspaces. By working in parallel, the GP algorithm is probably the only instance of an algorithm where the curse of dimensionality is an advantage. Optimality of the GP algorithm lies in its rate of exploration of the different subspaces, more precisely, it concentrates the search in the subspaces which have been represented by highly fit functions.

The details of this analysis are explained more fully later in the paper. For derivation of the results, one is referred to Holland (1975, 1992).

Nonlinearity of the earnings-returns relation is a readily observable and empirically supported finding of the finance and accounting literature. A considerable effort has been expended by researchers in the field to identify a better structural alternative, but with little success (see Freeman and Tse, 1992; Beneish and Harvey, 1998, for a discussion of nonlinearity in this empirical context). This is our motivation to apply the GP technique to model selection in this context. The earnings-returns models that are used currently do not capture adequately, in a structural setting, the form of the earnings-returns relation.

The linear earnings-returns model is misspecified. Some of the misspecification has been attributed to omitted variables. Indeed, evidence exists that firm characteristics affect the earning-return relation (Lipe et al., 1998). Authors attempting to deal with this problem in cross-sectional setting have resorted to adding a host of variables as additional regressors attempting to control for factors influencing the relationship. Measurement error has also been cited as an issue, with authors suggesting different measures for the earnings surprises, different normalizing factors, different firm-specific control factors.

The following quote from Beneish and Harvey (1998) illustrates one possible approach to modelling the earnings-returns relation:

We estimate univariate and multivariate versions of various models in time-series and . . . we also estimate the model in cross-section. We augment the usual specification with ex ante proxies for the expected discount rate, risk changes, and the accuracy of forecasts . . . We also pool our data and estimate both the arctan and linear models . . . to assess which model best fits the data.

We advocate a more systematic approach and use genetic programming search together with data snooping tests to conduct a more balanced model selection exercise. The paper bridges three strands of the literature: 1) literature on earnings response coefficients and the nonlinearity of the earnings-returns relation; 2) literature on GP-based model selection; 3) literature on spuriousness and data snooping.

The problem's manifestations extend well beyond the realm of econometrics. Questions about the validity of the simple earnings-return regression have been used in courts to challenge expert testimonies in securities fraud litigation brought under Section 10 of the 1934 Securities Act. Different models of the response of prices to earnings produce different estimates of the abnormal return that should be observed as a result of unexpected earnings. This produces different stock inflation estimates in cases of alleged earnings misstatements and leads to widely different aggregate damages in cases of securities fraud.

Our major findings are that 1) using the described model identification procedure we are able to confirm the existence of a nonlinear model that has quite good performance over the benchmark model; 2) the results are not due to excessive data snooping; 3) the approach presents an advantage over alternative nonparametric techniques as it can be interpreted to give parametric model for use in further empirical work on the earnings-returns relation and in other applications.

The next section reviews the literature on nonlinearity as it concerns the earnings-returns relation. In section 5 we describe our proposed approach, based on an algorithmic technique of model selection. Section 7 shows some simulation results illustrating the effectiveness of the model and discusses its advantages and caveats. Section 8 describes the data we use to perform our analysis. In section 9 we provide the results and provide some discussion. The last section concludes.

2. WHY AND HOW ARE EARNINGS AND RETURNS RELATED?

Quarterly earnings announcements of firms are a key variable that is anticipated by stock analysts and investors alike. In this section we discuss briefly the reasons for and the complexities of the link between quarterly earnings announcements and stock returns. A more extensive discussion regarding the ways this relationship has been modelled by researchers is in the next section.

Every public company on the U.S. stock exchange is required by law to file, on a timely basis, quarterly forms 10-Q (Quarterly Report) and annual form 10-K (Annual Report) with the U.S. Securities and Exchange Commission (SEC). These forms contain a wealth of information, most notably balance sheets (also known as consolidated statement of financial position), profit and loss account (also known as consolidated statement of earnings), and cash flow statement for the past quarter. The key number from these reports is the earnings per share (EPS) which represents net earnings for the quarter divided by the total number of shares outstanding. There are strict guidelines, the Generally Accepted Accounting Principles (GAAP) which dictate how the operations of the company are reported in its financial statements. Therefore, when earnings per share are announced, they provide valuable information about the current state of the company.

Analysts in brokerage houses follow companies and form a link between the company's management and investors. As part of the service they provide to institutional and individual investors, analysts issue EPS forecasts for future quarters. These EPS forecasts are also a key ingredient in analyst stock valuation models which they often use to justify issuing buy/sell recommendations regarding company stock. At any point in time, for a particular company and for a particular upcoming quarterly report, one could assemble EPS forecasts

in some specific way to form a “consensus” forecast. For example, since many EPS estimates are available from different analysts, one could define the “consensus” to be the mean EPS forecast, the median EPS forecast, etc.

Investors keep track of analyst EPS forecasts and when actual quarterly earnings are announced, some time after the end of the relevant quarter, there often is a slight difference between the “consensus” forecast and the actual EPS number reported. This difference is referred to as *earnings surprise*, or *unexpected earnings*, and is promptly reported by newswires, financial internet sites, and newspapers. Sometimes, a significant earnings surprise is news by itself. It prompts analysts to reevaluate their models, form new price targets for the stock, and even issue a different buy/sell recommendation. The numbers from the consolidated quarterly reports themselves are also very informative about the current situation of the company and its future prospects. Thus, when quarterly earnings are announced, there is an instantaneous and powerful signal released into the market which causes investors to reevaluate the true value of the stock and trade based on the new information, thus causing the stock price to move. The stock price reaction to unexpected earnings often occurs on the day after earnings are announced, as companies usually release this information after markets close.

The informational link between the earnings surprise and the resulting stock price movement is not clear as we do not know the valuation model that analysts and investors are using. In addition, the net earnings number may include one-off items which are not expected to repeat in future periods, thus the earnings surprise may be due wholly or in part to a one-off item that bears no consequence for future earnings. Earnings may also include a component that has already been recognized by the market in terms of stock value, but from an accounting point of view, it may not be realized yet. As an example, consider the

case of an asset impairment, i.e. when an asset the company has on its books has real value that is significantly below that on the books. During the quarter, the market may realize that the asset (and therefore the company) is worth less and bid the share price down. However, analysts might not be as quick to adjust their earnings forecasts, partly because of uncertainty about the quarter in which the company will decide to recognize the loss of the asset, the exact amount of the impairment, and other issues. Problems such as this cause the relation between earnings and returns to be very complex.

If we assume that earnings forecasts are efficient, then unexpected earnings are a martingale difference sequence. If we further assume that earnings do not contain transitory components, i.e. the earnings surprise represents a permanent shift in the level of current (and future) earnings, then the “correct” stock price reaction to a dollar of unexpected earnings should equal the present value of a perpetuity of a dollar, namely $1 + 1/r$, where r is the (constant) discount rate for future cash flows.

It is important to bear in mind some features of data, some “stylized facts” about earnings per share data, which make the problem of finding and estimating an earnings-returns model a challenge.

Since EPS numbers are announced once every quarter and law forbids management to give specific estimates in the meantime, for each firm, there are only 4 observations per quarter. For most firms, there are very few observations with which to perform model estimation. Data also show that there is often a difference in price reaction to seemingly identical earnings announcements. Sometimes, a positive one penny surprise has no discernible effect on price, while other one-penny differences are accompanied by large stock price movements. Data also suggest that the magnitude of earnings surprise matters, with very large surprises having a smaller than proportional effect than small surprises. The sign of the earnings

surprise also has been noted to have an effect on stock returns, with plausible explanations advanced for both why positive or negative earnings surprises should be more value-relevant. Lastly, earnings surprises are naturally centered at zero, a factor which present earnings-returns models do not seem to exploit.

There are several desirable features of a good earnings-returns model. First, it should fit the data, i.e. minimize some distance measure with respect to the data. Second, it should account for features of the data, the stylized facts we described above. Third, it is desirable that we find a structural model of the earnings-returns relationship. A structural model that fits the data well would facilitate future research into the value-relevance of earnings and the way in which earnings are used in stock valuation by investors and analysts.

The next section outlines the literature on earnings-returns regressions.

3. PRIOR RESEARCH ON NONLINEARITY IN THE EARNINGS-RETURNS RELATION

The work of Ball and Brown (1968) has prompted a number of researchers to study more closely the relationship between accounting earnings and stock prices. At the heart of empirical models of the earnings-returns relation is the idea that quarterly earnings reports present an updated picture of the company's condition. Consequently, earnings have implications about future cash flows of the company and the value of its stock. It is reasonable to assume that the stock price already reflects the company's present condition and future prospects. This information set also allows analysts to form expectations about future earnings. Thus, when actual earnings figures are announced, the difference between actual and expected quarterly earnings introduces new information regarding the company's condition and future cash flows. This prompts reevaluation of the company's worth in the market and ultimately, a change in the stock price.

Traditionally, studies have used a linear regression model to describe the response of returns to earnings announcements. Common practice among researchers is to employ a linear econometric model similar to the one below:

$$CAR_t = \alpha + \beta UE_t + \epsilon_t,$$

where CAR_t is the cumulative abnormal return of the company's stock around the earnings announcement date t , and UE_t is the unexpected component of earnings.¹ The low explanatory power of this model and lower than theorized estimated coefficient β have also prompted researchers to consider many alternative specifications of this relation. For a review and empirical analysis see Kothari and Zimmerman (1995), who consider price and return specifications of the above model and conclude that price models have less biased coefficient while return models have less serious econometric problems.

In the linear specification of the model above, the coefficient β is called the earnings response coefficient (ERC). It is an important variable as it purports to summarize the value relevance of new information about the company's earnings. If we assume that earnings forecasts are efficient then unexpected earnings are a martingale difference sequence. If we further assume that earnings do not contain transitory components, i.e. the unexpected part of earnings represents a permanent shift in the level of current (and future) earnings, then the "correct" stock price reaction to a dollar of unexpected earnings should equal the present value of a perpetuity of a dollar, namely $1 + 1/r$, where r is the (constant) discount rate for future cash flows. Empirical studies, however, routinely find estimated ERC to be lower than theory would predict.

¹The regression has its cross-sectional variant which suffers from severe econometric problems (see Teets and Wasley, 1996, for a discussion).

Authors have also noted the low explanatory value of such OLS regressions (R^2 is often at or below 5%) and have looked for reasons for the failure of the traditional model. A widely documented feature of the data is an observed “S-shaped” departure from linearity (see Freeman and Tse, 1992; Cheng et al., 1992; Das and Lev, 1994; Lipe et al., 1998) when one plots on a 2-dimensional graph the size of unexpected earnings versus the corresponding abnormal stock price reaction. The stylized S-shape of the earnings-returns relation has been the subject of much scrutiny and refinement. It remains a feature of the data when we remove transitory items from earnings or even if we use unexpected cash flows instead of earnings (see Das and Lev, 1994).

It is by now well known that ERC coefficients vary with both the size and the direction of earnings surprise. We proceed to offer the common justifications proposed for each of these observed departures from linearity.

With regard to the magnitude of the earnings surprise, Freeman and Tse (1992) hypothesize the following explanation. Since proxies for expected earnings are more accurate in predicting the permanent component of earnings, extreme values of earnings surprise reflect mainly transitory components. This is supported by Brooks and Brookmaster (1976) who find that firms experiencing large one-period change in earnings do not sustain this level of earnings in later periods. Since stock prices do not react as strongly to transitory components of earnings, ERC should be lower for earnings surprises of large magnitude. This hypothesis was confirmed in nonlinear cross-sectional studies by Freeman and Tse (1992) and by Das and Lev (1994).

With regard to the direction of the earnings surprise, researchers have documented a difference in the response of stock prices to positive and negative announcements. In particular, Skinner and Sloan (2001) find that for growth stocks abnormal returns around

negative earnings surprises are much larger than abnormal returns for value stocks. This finding adds to the common observation in the financial press that small earnings disappointments have disproportionate price effects. One explanation is that with the numerous ways for CFOs to manage earnings and analysts' earnings expectations, having to miss earnings estimates even by a small margin is a signal of a much larger problem that may affect future cash flows and has immediate implications for the present value of the company.

Another argument for the asymmetry involves the timeliness of earnings reports. Namely, negative news should induce stronger price response due to the fact that bad news are reported more rarely and on a more timely basis than good news, thus their impact is more localized around earnings announcements. Anticipated gains, on the other hand, are often leaked gradually and thus their effect is spread over a larger time period and is not captured by the short measurement windows around earnings announcements.

The degree of earnings persistence has also been advanced as an important cause for nonlinearity (Penman, 1992; Liu and Thomas, 2000). These studies augment the linear earnings-returns model using variables that relate to earnings persistence, for example, earnings forecast revisions by analysts (Liu and Thomas, 2000). This results in an increase in the explanatory power of the earnings-return regression.

Studies have also considered the effect of firm-specific factors on the earnings response coefficient (Kormendi and Lipe, 1987; Collins and Kothari, 1989; Easton and Zmijewski, 1989). For example, Teets and Wasley (1996) demonstrate how cross-sectional studies lead to downward-biased ERC coefficients and low explanatory power whenever ERC and the variance of unexpected earnings are heterogeneous, and when ERC coefficients are inversely related to the across-firm variance of unexpected earnings. Furthermore, the authors point to empirical evidence that the above two assumptions likely do present a problem with the

cross-sectional regressions employed in practice. As an example, risk is a factor directly related to the variance of earnings and negatively related to ERC (Easton and Zmijewski, 1989). Therefore, one should consider firm-specific estimation of ERC coefficients and weigh the benefits of cross-sectional estimation carefully against the above-mentioned disadvantages.

Earnings response coefficients derived from linear models are biased in both direction and magnitude. Researchers have already begun employing devices that deal with the problem, from transforming regression variables, introducing interactions between variables, using firm-specific and industry factors, and employing nonparametric techniques.² However, most work in this direction has been *ad hoc*, each regression modification or variable transformation being arbitrary, justified by the slightly higher explanatory power of the resulting regression. One is reminded of Merton (1987)'s critique. More importantly, one begins to doubt whether nonlinearity is worth correcting for, as the limited efforts in this direction so far have failed to yield satisfactory results (Beneish and Harvey, 1998).

We argue that it is important to derive a nonlinear structural relation between unexpected earnings and returns since economic theory is firm that earnings influence returns and that there are many good theoretical reasons why the relation should be nonlinear. In that sense, a search for a parametric nonlinear model is desirable, yet such data-mining methods need to be checked for data snooping as they will tend to overexploit the available data but have little predictive power. Research methods should be developed that are able to identify promising nonlinear structural models of this relation. Such methodologies can be used to derive earnings response regressions and aid future research in explaining the residual

²See Kothari and Zimmerman (1995) for discussion of scaling of variables, Beneish and Harvey (1998) for consideration of alternative functional forms, including nonparametric regression.

variation of excess returns that remains after variation due to the unexplained component of earnings has been removed.

4. MODELLING FRAMEWORK

Our objective is to find a parametric model of the earnings-returns relation. However, being limited to a particular functional form, or at best, a set of functions defined by the values of an estimated parameter, as in traditional parametric context, is undesirable. We wish to have more freedom in deciding which function to fit to the data. Nonparametric estimation methods allow us this freedom, but we give up the ability to have as the result of the estimation exercise, a structural model. A technique for estimating a parametric model without specifying too narrow a family of parametric curves is needed. We will describe the method in the next section. Here we state formally the minimization problem and provide some discussion.

Suppose we are given data $\{x_n, y_n\}_{i=1}^N$ where the data generating process is $y_n = f(x_n) + \epsilon_n$, where ϵ_n are zero-mean uncorrelated random variables with a constant variance σ^2 and $f(x_n)$ are the values of some unknown function $f(\cdot)$ evaluated at the points x_n . Let $f(\cdot)$ belong to some family of functions Ξ .

The minimization problem can be written as:

$$(4.0.1) \quad \min_{\hat{f} \in \Xi} \left(y_n - \hat{f}(x_n) \right)^2 .$$

Determining a suitable inferential methodology for (4.0.1) depends on what assumptions we make about $f(\cdot)$. If the form of $f(\cdot)$ is known up to finitely many unknown parameters, we can employ parametric regression techniques to estimate $f(\cdot)$. Such techniques have desirable properties. For example, when properly specified, the corresponding inferential

method has good efficiency properties. (For most parametric estimators the expected sum of square errors decays to zero at rate n^{-1} .) Furthermore, parameters estimated from a parametric model have some meaning which makes them interpretable and of interest. However, as we stated before, this assumption of specific functional form can be too restrictive when there are no good reasons to choose a particular functional form.

A nonparametric regression approach only assumes that $f(\cdot)$ belongs to some infinite dimensional collection of functions. Usually, some mild conditions of differentiability are imposed on $f(\cdot)$. The expected sum of squared errors for most nonparametric estimators decays to zero at the rate $n^{-\delta}$ where $\delta \in (0, 1)$ depends on the smoothness conditions imposed on $f(\cdot)$. For example, when $f(\cdot)$ is twice differentiable, $\delta = 4/5$. However, nonparametric estimators do not obtain a specific functional form for $f(\cdot)$.

If we are interested in obtaining a particular structural representation of the conditional mean $E[y_n|x_n]$, a criterion of how well our model does is the degree to which the resulting structural representation resembles the true data generating process (DGP). For example, if the data is generated using $f(x_n) = x_n^2 - 1$ we would like a model that has in its representation expressions such as x_n^2 and even $x_n - 1$ (which can be multiplied by $(x_n + 1)$ to give the true data generating process). We would also like to discourage expressions such as $1/x_n^7$ which do not seem part of the true DGP structure.

Our proposed method is essentially a parametric estimation algorithm within a much wider set of functions. At issue is the way the algorithm searches through this wider set of parametric functions. Optimality refers to the way this search is performed, given a limit on our computational resource and the vastness of the search space.

The GP algorithm is a minimization algorithm. Specifically, we define a (large) set of functions Ξ and minimize the criterion (4.0.1 on the page before) over this set. The result

SEARCH FOR A STRUCTURAL SPECIFICATION OF THE EARNINGS-RETURNS RELATION¹⁷ is a parametric model. This parametric model is the result of an optimal search over Ξ , i.e., the algorithm divides the search space into sub-spaces and then as the search progresses, it concentrates on the sub-spaces which contain functions with high fit to the data.

In the next section we describe in detail the optimization algorithm and present results which show the optimal way in which it searches through Ξ for a functional form of the true data generating process.

5. THE GP ESTIMATION METHODOLOGY

We use genetic programming (GP) to find nonlinear functional forms for the returns-earnings relation. In the subsections that follow, we explain in detail the workings of the algorithm, outline the theory behind GP, and demonstrate its model selection ability.

5.1. Genetic Programming. Genetic programming is an adaptive search technique, and as with all other adaptive search techniques there is a structure within the model that undergoes adaptation. In GP this structure is a collection of points in the search space, as opposed to a single point. This collection of points is called *population*. Thus, at any moment during its search, GP simultaneously handles one population of plausible solutions to the minimization problem in equation 4.0.1 on page 15.

In order to discover better functional representations of the true data generating process, genetic programming manipulates the points in the population. The points in the population represent functions with an internal tree representation which is used by the GP algorithm to adapt them. We will call these representations *tree structures*. As an example of a simple tree structure, we present the GP tree representation of the expression $x + 2$ in figure 1 on the next page.

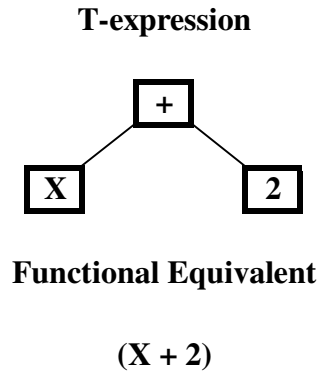


FIGURE 1. The tree structure $(x+2)$

Tree structures have two types of nodes: internal and external. Both internal and external nodes contain a value which specifies the value they return. External nodes are terminal nodes; they do not link to any other nodes besides the node above them. Thus, external nodes form the boundary of the tree structure. When an external node is executed, it simply returns its value. Internal nodes, on the other hand, have other nodes below them; internal nodes take the nodes immediately below them and perform the operation specified by the operator value that is in the internal node. In our simple example above, the internal node containing the operator $[+]$ takes the two nodes immediately below it and sums their values. This is, by definition, the function of an internal node containing $[+]$. The nodes containing $[x]$ and $[2]$ are external nodes. They evaluate to $[x]$ and $[2]$, respectively. Thus, the whole tree structure evaluates to the linear function $f(x) = x + 2$.

Note that simple tree structures like the one above can be combined with each other to form more complex tree structures. Evaluation of each such expression proceeds similarly in a recursive fashion. An example of the recursive process of tree structure evaluation is presented in figure 2 on the facing page.

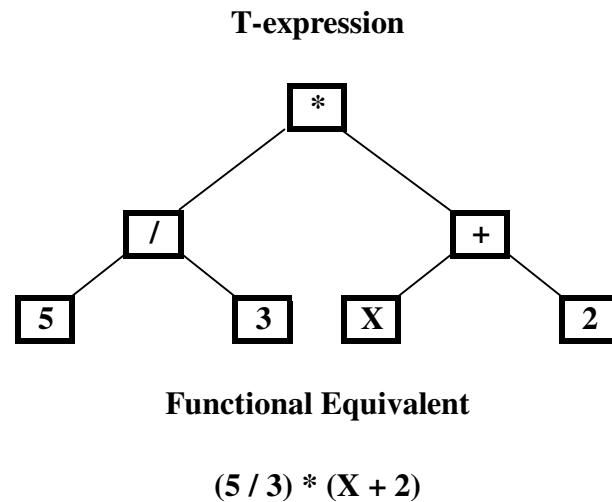


FIGURE 2. The tree structure $(5/3)*(x+2)$

First, the internal (root) node $[*]$ is evaluated and since it is an internal node, the evaluation proceeds to evaluate its two sub-nodes by recursion. The first sub-node is $[/]$, also an internal node which takes two arguments, so the recursive step is to traverse one level down and evaluate its arguments. The two arguments are external nodes and evaluate to $[5]$ and $[3]$, respectively, thus the recursive step is complete at this level and the overall internal node $[/]$ now evaluates to $5/3$. The second sub-node, $[+]$ also takes two arguments and each of them is evaluated recursively to produce $[x]$ and $[2]$, respectively. These values are passed to the parent node which applies the operation $[+]$ to them to produce the value $[x+2]$. The recursion returns to the original root node where the operator $[*]$ applied to the two, now evaluated, internal nodes below it yields the final functional expression for this tree structure: $f(x) = (5/3) * (x + 2)$.

Define a path along the tree as a traversal of the tree starting from the root and ending at an external node, each time going down one level by choosing exactly one of the available

internal or external nodes. The length of such path is the number of nodes traversed, including the root node (beginning) and the external node (end).

Definition 5.1.1. The *depth* of a tree structure is defined as the maximum length of all paths for the given tree.

Tree structures can be very flexible in the types of functional expressions they represent. The internal and external nodes contain values which define the domain of functions Ξ where GP searches for a solution to the maximization problem. Once the allowable values for internal and external nodes are specified and a maximum tree structure depth is set, tree structures define the domain for the GP search. The search space is all the possible tree structures that can be created by the recursive composition of the allowed values for internal and external nodes.

When specifying the allowable set of internal and external node values, we need to ensure the properties of closure and sufficiency are satisfied.

Definition 5.1.2. The allowable internal node values obey the **closure** property when each internal node value is able to accept as its arguments the value of any external node or any value which is returned from any internal node.

Closure is also an important property as it allows GP to modify the tree structures in such a way that they continue to be valid tree structures.

Definition 5.1.3. The allowable internal and external node values obey the **sufficiency** property when there exists a tree structure formed by these values which is a solution to the maximization problem in equation 4.0.1 on page 15.

We need to specify the allowable sets of external and internal node values. The choice influences both the eventual success of GP in solving the optimization problem and the form of the solutions that emerge.

For our particular application we choose the following set of values for the internal nodes: $F = \{[+], [-], [*], [%]\}$, where [%] is the protected division operator, defined to return a large fixed number with the appropriate sign of the numerator, when the denominator is below a certain small threshold value.

The external nodes can take two possible values in the set: $L = \{[x], [r]\}$. The value [x] stands for the input argument of the tree structure, thus tree structures resemble a simple function of one argument. The external node [r] is often described as a constant. Its value is the realization of a standard uniform random variable R , i.e. r is distributed uniformly on the interval $[0, 1]$. Once r , which is a realization of R , is seeded into a particular external node as its value, it remains a constant in that node. Different external nodes may hold different realizations of the random variable R .

In the discussion that follows, it will be helpful to refer to the flowchart in figure 6 on page 32

5.1.1. *Generating The Initial Population.* The GP algorithm works with a set of tree structures at a time. Such a set of tree structures is called a *population*. The initial population of tree structures is created one tree structure at a time, as follows. GP first selects an element from the set of allowable internal and external values using a uniform distribution (each element of the set F has a probability of $1/|F|$ of being selected, where $|F|$ is the number of values in the set F). Call this initial value f . It is placed as the trunk of the tree structure. Depending on the exact element $f \in F$ that is selected, GP then selects

the necessary number of elements from the combined set of internal and external values $C = F \cup L$. Each element has an equal chance of being chosen. For each of these elements, it selects again and operator from the combined set C . The particular branch of the tree structure terminates when an external node is chosen at random from the set C . During the recursion process, the GP algorithm also checks whether the depth of the tree structure has reached a predetermined value specified by the researcher. In this case, GP chooses an external node at random from the set of external node values L .

There are three methods of creating the initial population of tree structures. The “grow” method has just been described above. The “full” method is to create all tree structures with maximal allowed length, i.e. to restrict the choice of node at each level of the tree structure to the set of internal nodes F and at the last level of the structure only choose from the set of external nodes L . Key to the success of GP is the variety of tree structures that it is working with, thus a third method, the “ramped half-and-half” has been suggested and used widely in the literature. It involves generating an equal number of tree structures with tree depth parameter ranging from 2 to the researcher-specified maximum. For each value of the depth parameter, half of the tree structures are created using the “grow” method and half according to the “full” method. We have used the ramped half-and-half method in the present application.

A key element of the GP algorithm is the assignment of a fitness value to each tree structure. This fitness value is specific for each expression and determines how well the expression fares in comparison to other tree structures in its generation. It also needs to correspond to the criterion we are minimizing. Fitness also drives the stochastic search path of the algorithm, as we will describe later.

Fitness can be measured in many ways. In our particular application we are interested in the degree to which the tree structure can describe a nonlinear functional relation. As mentioned before, the tree structure, when evaluated, is designed to resemble a function. We can think of forming a measure by sampling this function at several points and evaluating a SSE-type value. For example, given the set of n points, $(x_i, y_i)_i^n$, for each tree structure j , we can define the measure

$$\mu_j = \sum_{i=1}^n (y_i - T_j(x_i))^2$$

where $T_j(x_i)$ is the value that the tree structure j evaluates to when the input value x_i is used. Many variants of the above raw fitness expression are possible and the proper fitness depends on the particular application.

/iffalse

The fitness measure μ_j is a function that is small for better-performing tree structures and large for worse performers. Thus, the GP is designed to search for a tree structure that minimizes this value. As the value of μ_j gets smaller, it might be desirable to exaggerate the size of the fitness difference between the best tree structures. Thus, one often uses as measure a function of μ_j , for example,

$$m_j = \frac{1}{1 + \mu_j}$$

Note that As often happens with GP, as it searches through the parameter space it encounters more and more fit tree structures. This modified fitness measure has been used successfully by Koza (1992) to emphasize the difference between a good tree structure and a very good one. It is extremely potent when with prior knowledge the researcher has chosen the set C so that an exact solution (with $\mu_j = 0$) can be found.

/fi

Normalized fitness is an essential ingredient of GP as it determines how it will search through the space of allowable tree structures. Normalized fitness is defined for each tree structure as follows:

$$n_j = \frac{\mu_j}{\sum_{j=1}^J \mu_j}$$

where j indexes all the tree structures in the present population.

This normalized fitness has the following important properties by construction:

- (1) $0 < n_j < 1$
- (2) $\mu_j > \mu_k \Rightarrow n_j < n_k$
- (3) $\sum_{j=1}^J n_j = 1$

The normalized fitness governs how GP samples the points in the current population to produce points in the new population. The sampling technique is termed “fitness-proportionate sampling” and will be described in the subsection that follows.

To summarize, in the initial stage, the GP algorithm begins by generating a population of J tree structures according to the procedure outlined above. The parameter J as well as the tree depth and the sets F and L are supplied by the researcher. The GP algorithm then computes the fitness measure for each tree structure. Storing the population and the fitness measures is the only memory requirement of the GP algorithm, besides storing the input data itself. Only one population of functions is stored at any one time.

5.1.2. Modifying the Population Using Operations. After creating the initial population the algorithm applies three operations that act on the current population by selecting tree structures in the population to produce new tree structures which form a new population set. Each of the operations is defined below.

The *reproduction* operation selects a tree structure from the current population according to fitness, i.e. each tree structure j has a probability n_j of being selected. The tree structure is then placed, unaltered, in the new population. The reproduction is fitness-proportionate, meaning, the chances of any one tree structure being selected for the reproduction operation is given by its normalized fitness

$$n_j = \frac{\mu_j}{\sum_{j=1}^J \mu_j}$$

Fitness-proportionate reproduction ensures that the better tree structures have a greater chance of being represented in latter populations.

The *crossover* operation works on two tree structures at a time. It takes two tree structures using proportional selection from the present population, i.e. using the same selection method as the reproduction operation. For each tree structure, the following procedure is performed to identify a “break point”. Denote by k the number of nodes and leaves in the tree structure. The k nodes and leaves are numbered consecutively using the integers from 1 through k , going down-first, and left-to-right. A random variable is drawn from a uniform discrete distribution over the integers $\{1, 2, \dots, k\}$. After a break point is selected for each of the two tree structures, the tree structures are disassembled at the break-point. Figure 3 on the next page presents an example of two tree structures selected for a crossover operation.

Their nodes have been numbered and a random node has been chosen according to the rule above. The break point has been shaded in the figure and the tree structures have been disassembled. The crossover operation then swaps the lower disassembled parts of the tree structures so that two new tree structures are formed. Figure 4 on page 27 shows these new structures in our example.

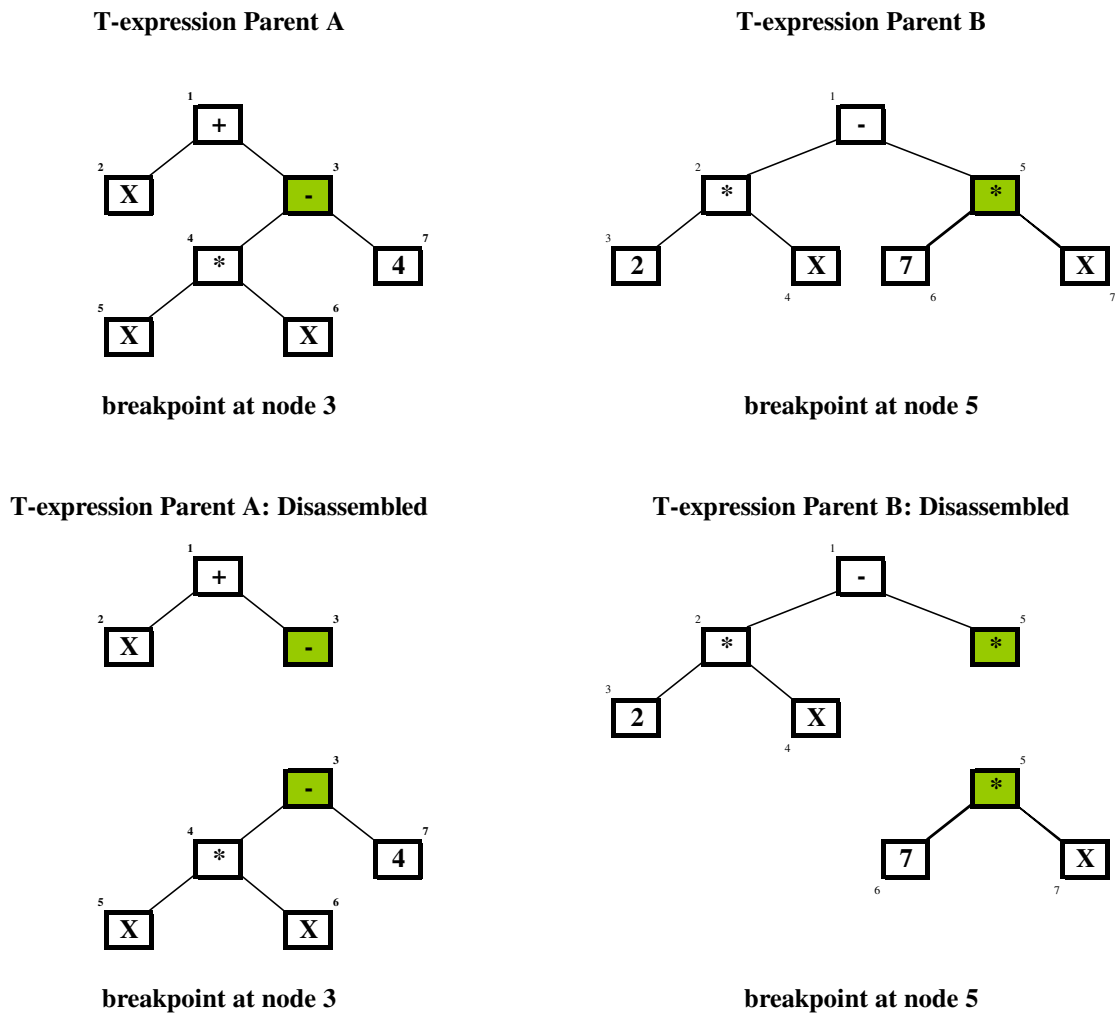


FIGURE 3. Crossover

The closure property ensures that the resulting trees are valid tree structures. A check is made that the resulting tree structures are not too big, i.e. that the tree depth does not exceed some limit pre-specified by the researcher. If the maximal allowed depth is exceeded, the offending tree structure is discarded in favor of one of the original two tree structures (chosen with equal probability). Note that the other new tree structure may still be valid.

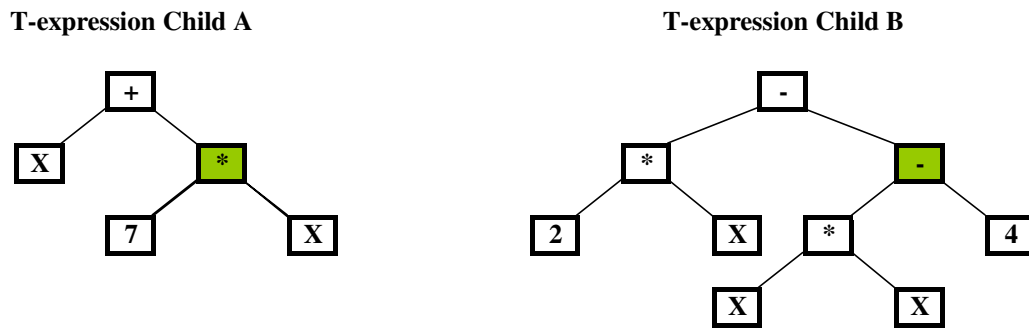


FIGURE 4. Crossover

Once two valid tree structures are produced, the crossover operation places them into the new population set.

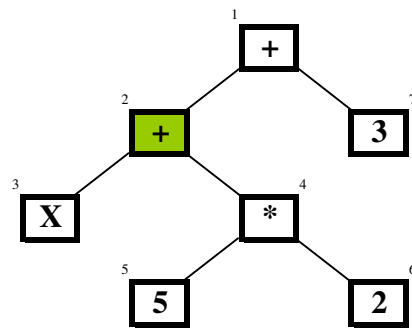
The *mutation* operation is used to reintroduce diversity in the population of tree structures. Inherently, the crossover and replication operations result in the convergence of the population to several fit tree structures due to proportional sampling. Mutation ensures that the GP algorithm continues to explore new points in the parameter space and reduces the chance that it gets stuck at a local maximum. The mutation operation chooses one tree structure from the current population, proportional to fitness. It then identifies a break point of this tree structure in the same way this was done by the crossover operation. The tree structure is separated at the break point and its lower part is deleted. Then, starting from the breakpoint, a new tree structure is generated via the “grow” method used for constructing the tree structures in the very first population. The resulting mutated tree structure is added to the new population.

Figure 5 on page 29 displays the process of mutation applied to a sample tree structure. The original tree structure corresponds to the expression $(x(5*2)) + 3$. The seven nodes of the tree structure are numbered top-down, left-to-right, and an integer is selected from

a uniform distribution over the integers from 1 to 7. In this example, the realization of the uniform random variable is 2, thus the second node is selected as the breakpoint. The subtree below the breakpoint is discarded. With it is discarded the underlying structure, which may have been the result of many generations of crossover. Starting from this second node as its trunk, a new subtree is generated to replace the already discarded one. The new subtree is randomly generated thus it has the effect of introducing new building blocks into the tree structure. This is similar to applying a random disturbance to a candidate-solution to a maximization problem, except that this disturbance is not necessarily local. The purpose of the disturbance is to avoid identifying a false local maximum as the global maximum.

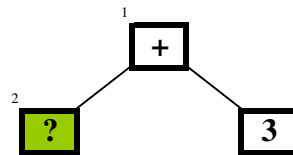
The central loop of the GP algorithm involves taking the current population and, using the operations above, producing a new population with the same number of tree structures in it. The exact frequency of use of each of the operations is governed by parameters specified by the researcher. The researcher specifies constant probabilities for performing reproduction, crossover, and mutation, denoted by, $p(R)$, $p(C)$, $c(M)$ respectively, such that $p(R) + p(C) + c(M) = 1$ and each of $p(R)$, $p(C)$, $c(M)$ is between 0 and 1. Starting with the initial population, the GP method chooses an operation $E \in \{R, C, M\}$ with probability $p(E)$, and performs the operation, thus adding tree structures to the new population. Once the new population reaches maximal size, equal to the size of the current population, the old population is deleted from memory. The new population is evaluated, the fitness of each tree structure is calculated and this concludes the present iteration and allows the GP algorithm to perform a new one.

Original T-structure



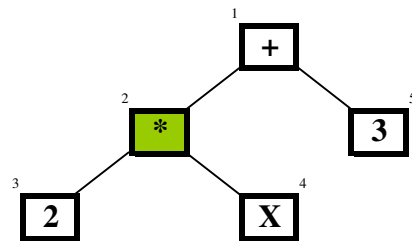
$$(X + (5 * 2)) + 3$$

Selected Subtree Is Discarded



$$? + 3$$

New Mutated T-structure



$$(2 * X) + 3$$

FIGURE 5. Mutation

As each new population is created, and the fitness of its tree structures evaluated, one structure from that population is designated as the best tree structure. It is the structure that achieves the highest fitness value among all tree structures, namely,

Definition 5.1.4. The best tree structure T_i from a given population is the structure which satisfies

$$i = \arg \max_j \left\{ \{m_j\}_{j=1}^J \cup T_{best} \right\}$$

Where T_{best} is the best tree structure encountered by the algorithm in all previous populations. It is cached by the algorithm to ensure that good performance, once achieved, is not lost due to mutation and crossover.

When more than one tree structure has a maximal fitness larger than the fitness of T_{best} , one of the tree structures with maximal fitness is arbitrarily chosen by the algorithm as the best tree structure for that generation.

5.1.3. *Condition for Termination and Best Result Designation.* The iterative step above terminates when an ex-ante termination criterion is fulfilled. It is common in GP applications to have a three-pronged termination criterion: the GP algorithm stops when either the best tree structure from the last population has achieved a certain threshold fitness value (specified by the researcher); or the GP algorithm has performed a given number of iterative steps; or the fitness value for the best tree structure has not improved by a pre-specified margin over the last several generations (the number of generations is again specified by the researcher). We have chosen to use only the stopping criterion based on reaching the maximum number of iterative steps. This does not influence our results, because if the algorithm fulfills one of the other possible criteria prematurely, we can only improve the fitness of the best tree structure by going a few iterations further.

5.2. **GP Control.** The control of the GP algorithm is performed via the specification of several parameters.

Firstly, the researcher needs to specify the data, i.e., the sequence $\{y_i, x_i\}_{i=1}^n$. Note that more than one dependent variable is possible. The researcher also specifies the search space Ξ . This is accomplished by 1) deciding on the allowable values for the set of internal and external nodes, 2) setting the maximal tree depth allowed during the search.

The trajectory of the search is governed by several parameters: 1) the probabilities of replication, crossover, and mutation, $p(R), p(C), P(M)$, respectively, 2) the size of the population (i.e. how many tree structures are contained by the population at any one time) 3) the number of iterations (since each iteration is referred to as a “generation”, this parameter is also called the maximal number of generations), 4) the initial tree depth (used to perform the initialization of the GP algorithm and generate the tree structures in the initial population).

We present a table of the key researcher-specified GP control parameters and some typical values.

Name	Variable	Typical Value
Population Size	M	3000
Maximum number of populations	G	100
Reproduction probability	p(R)	0.10
Crossover probability	p(C)	0.85
Mutation probability	p(M)	0.05
Maximal Initial tree depth	D_i	3
Maximal tree depth after crossover	D_c	10

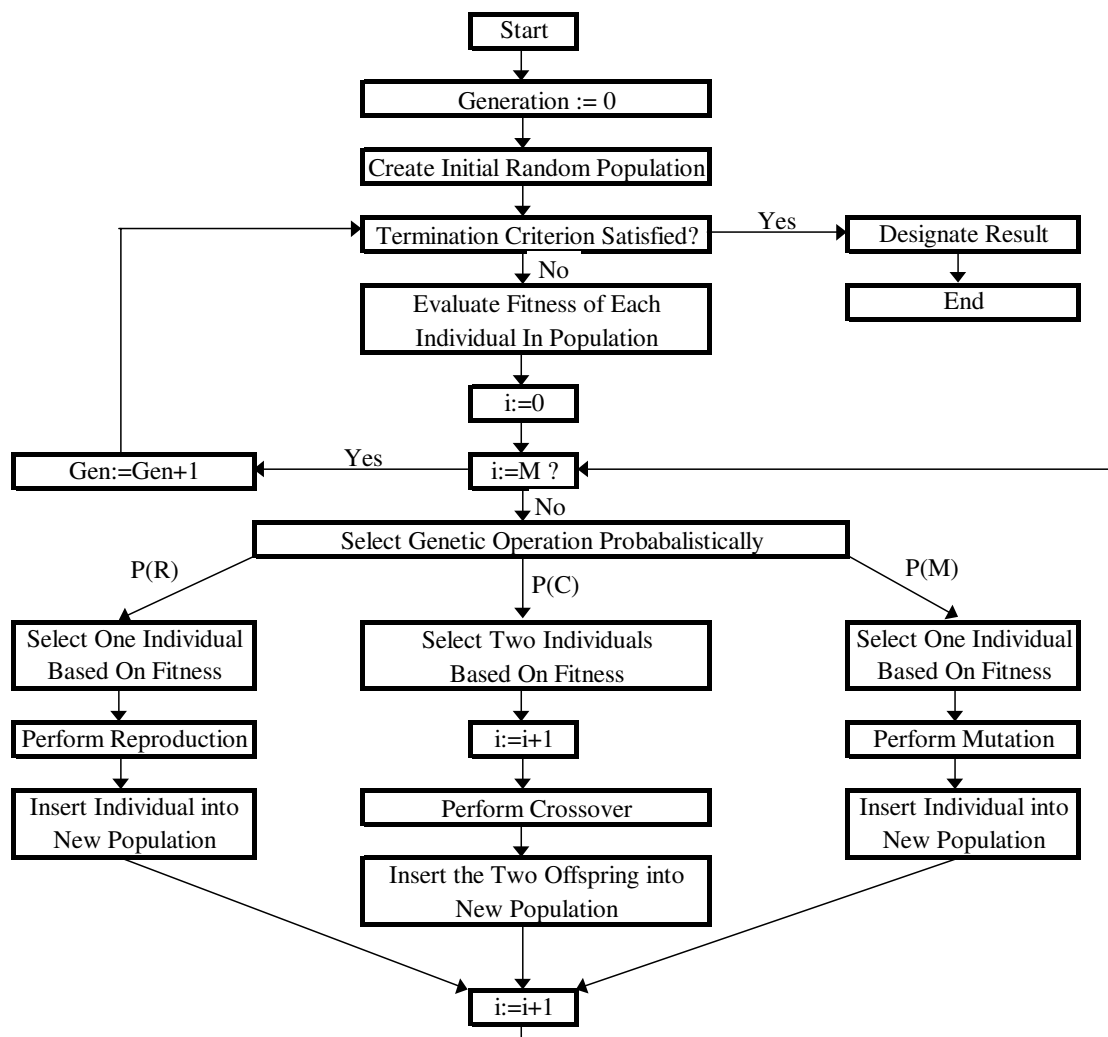


FIGURE 6. GP Algorithm

The flow of the GP algorithm is illustrated by the flowchart in figure 6 on the following page.

6. OPTIMALITY OF THE GP ALGORITHM

This section makes explicit the notion of optimality of the GP search procedure. Note that even with large population sizes, the GP algorithm samples only a fraction of the

possible points in the search space. The key is how information gathered from evaluating the fitness of this limited number of tree structures is processed by the algorithm.

We start by discussing the two-armed bandit problem, the multi-armed bandit problem, and their optimal solution. We will then draw the parallel between the solution to these problems and the search strategy used by the GP algorithm.

6.1. Two-Armed Bandit Problem. Suppose we are given two random variables, ξ_i and ξ_j , and we know the two mean-variance pairs (μ_1, σ_1^2) and (μ_2, σ_2^2) . The researcher can distinguish between ξ_i and ξ_j only by their letter index but does not know which of the two mean-variance pairs describes the distribution of ξ_i and which pair describes the distribution of ξ_j . When we sample from one of the distributions, the particular realization we obtain is called the payoff from this drawing. The researcher is allowed to draw a total of N realizations from the two distributions. The two-armed bandit problem concerns the optimal allocation of these N trials between the two random variables in order to maximize the total payoff.

It is not possible to know with certainty which of the two random variables has a higher expected payoff $\max\{\mu_1, \mu_2\}$. We can only allocate some trials to each variable and then continue to sample each variable in such a way as to maximize expected payoff. The problem can be restated in terms of minimizing the loss associated with sampling the wrong distribution. Suppose of the N trials, n are allocated to sampling ξ_j and $N - n$ trials are allocated to sampling ξ_i . Without loss of generality, after N draws, let $\xi_i(N)$ be the random variable with larger average payoff per trial. There are two possible sources for loss from erroneous sampling: if the observed best variable (i.e. the one that gave higher average payoff per trial after N trials) ξ_i is indeed the variable with the higher mean, then we have

“wasted” n trials by sampling ξ_j realizing an average loss of $|\mu_i - \mu_j|$ per trial. Denote the probability of this by $1 - q(N - n, n)$. However, if the observed best variable is not the real best variable, we have “wasted” the $N - n$ draws we expended on ξ_i , losing $|\mu_i - \mu_j|$ per draw. This occurs with probability $q(N - n, n)$. Thus, we can write the total expected loss as

$$L(N - n, n) = [q(N - n, n)(N - n) + (1 - q(N - n, n))n]|\mu_1 - \mu_2|$$

Holland (1975) shows that there exists an $n^*(N)$ such that it minimizes the loss function $L(N - n, n)$. Furthermore, the following result establishes an upper bound on the number of trials n that optimally should be allocated to sampling ξ_j :

Theorem 6.1.1 (Holland (1975) Theorem 5.1). *Given N trials to be allocated to two random variables, with means $\mu_1 > \mu_2$ and variances σ_1, σ_2 respectively, the minimum expected loss results when the number of trials allocated to the variable with the lower revealed expected payoff after the N trials is*

$$n \leq n^* \sim b^2 \ln[N^2 / (8\pi b^4 \ln N^2)]$$

where $b = \sigma_1 / (\mu_1 - \mu_2)$. Furthermore, if initially each variable is equally likely to be the one with the higher payoff, then $n = n^*$ and the expected loss becomes

$$L^*(N) \sim b^2(\mu_1 - \mu_2)[2 + \ln[N^2 / (8\pi b^4 \ln N^2)]],$$

where $Y(t) \sim Z(t)$ denotes $\lim_{t \rightarrow \infty} (Y(t)/Z(t)) = 1$ for any two real-valued functions $Y(t)$ and $Z(t)$.

The theorem allows us to establish the following result:

Corollary 6.1.2. *Denote by $N^* = N - n^*$ the number of trials allocated to the variable with higher expected payoff. Then*

$$N^* \sim N \sim \sqrt{8\pi b^4 \ln N^2 \exp(n^*/2b^2)}$$

regardless of the distribution of ξ_1 and ξ_2 .

The above corollary shows that in order to minimize the expected loss function, the number of trials allocated to the better variable ξ_i should grow slightly faster than the exponential function of the number of trials allocated to ξ_j . Holland showed that the GP algorithm using fitness-proportional crossover and mutation asymptotically samples each schemata using this optimal rate. We will explain this in greater detail in the next subsection.

Note that the above discussion was based on the assumption that we know the variable ξ_i that will display higher expected payoff after N draws. Holland (1975) (Corollary 5.2) shows that the following sampling implementation produces a loss function which approaches the minimal loss $L^*(N)$: calculate n^* and allocate n^* trials to each of the two random variables. The remaining $N - 2n^*$ trials should be allocated to the variable displaying higher expected payoff after the initial $2n^*$ draws.

The theorem above has its multivariable equivalent. It can be stated as follows:

Theorem 6.1.3 (Holland (1975) Theorem 5.3). *Given N trials to be allocated among r random variables, with means $\mu_1 > \mu_2 > \dots > \mu_r$ and variances $\sigma_1, \sigma_2, \dots, \sigma_r$ respectively, the minimum expected loss results when the number of trials allocated to the variable with the lower revealed expected payoff after the N trials must exceed*

$$(r - 1)(\mu_1 - \mu_2)b^2[2 + \ln[N^2/(8\pi(r - 1)b^4 \ln N^2)]],$$

where $b = \sigma_1 / (\mu_1 - \mu_r)$

6.1.1. *Optimal Sampling of Schemata by the GP Algorithm.* In order to formalize the GP algorithm and show the way its search is optimal, we first discuss the way GP searches through the parameter space. We use the notion of a schema:

Definition 6.1.4. A schema (pl. schemata) is defined by a given subtree. For a given subtree, a schema is the set of all individual trees that contain the specified subtree. Thus, each tree belongs to many schemata and each schemata contains many trees.

Thus, each tree structure represents an infinite number of schemata, as it can be a subtree of an infinite number of possible trees. However, when the depth of the trees is fixed, each tree structure belongs to a finite (still very large) number of schemata.

An important insight made by Holland (1975) is that one can view each schema as a random variable ξ with an unknown mean and variance. When a fitness value is calculated for a tree structure, one can also view this fitness value as a drawing from the distribution characterizing the random variable ξ . Thus, when the GP algorithm evaluates the fitness of a particular tree structure, it implicitly derives information about all the possible schemata to which this tree structure belongs. This is the heart of the argument showing the intrinsic parallelism of the GP algorithm.

One can then define the fitness of a schemata H at generation t , denoted by $f(H, t)$, as the average fitness of all tree structures belonging to this schema that have been sampled.

$$f(H, t) = 1/k \sum_1^k f(T_H)$$

Holland (1975) showed that for genetic algorithms with reproduction and crossover which operate proportionally to fitness, the expected number of occurrences of the schema in

generation t , denoted by $m(H, t)$, is bounded below by a factor proportionate to its observed relative fitness.

$$m(H, t + 1) \geq \frac{f(H, t)}{\bar{f}(t)} m(H, t) (1 - \epsilon)$$

where $\bar{f}(t)$ is the average fitness of the population t and ϵ is a small factor that is proportionate to the length of the schema H . For short, compact schema, ϵ is smallest and thus the sampling rate of such schema is close to the exponentially optimal sampling rate in the multi-armed bandit problem.

Thus, Holland's result show the efficiency of the algorithm in exploring the vast space of possible solutions Ξ by concentrating the search on subspaces with high fit to the data. Given the limited amount of the computational resource, the algorithm optimally samples the different subspaces to minimize the aggregate sampling loss associated with 1) sampling these subspaces that does not represent a solution, and 2) failing to sample only subspaces which represent the solution.

7. SIMULATION

In our present application, we are interested in identifying the functional relationship between two variables. In order to illustrate the effectiveness of the GP algorithm in this task, we present several simulation results. The first set of simulations deal with the ability of GP to discover the functional relationship when no noise is present. We show that in this case, GP is a highly effective tool. We then allow for a limited amount of noise to the system and show that GP still performs well but as the signal-to-noise level increases, GP is prone to "overfitting" i.e. fitting a spurious function to the noise level. This has also been verified by Kaboudan (2000), although his concern was with the statistical properties of the errors.

We simulate a data series $\{x_n, y_n\}_n^N$ where $y_n = f(x_n)$. For our choice of $f(\cdot)$ we use the following two specifications which have been suggested by previous earnings-returns research as plausible alternatives to the linear specification. Each of these models has the stylized S-shape characteristic of plots of the earnings-returns relation.

Modified-quadratic relation (Das and Lev, 1994):

$$y_n = \alpha_1 x_n + \alpha_2 \text{sign}(x_n) x_n^2$$

Arctan relation (Freeman and Tse, 1992; Das and Lev, 1994; Beneish and Harvey, 1998):

$$y_n = f(x_n) = \arctan(x_n)$$

Note that the arctan function cannot be expressed as a tree structure using the operations $\{+, -, *, \%\}$. However, the response function in the arctan case is $1/(1 + x^2)$ which is expressible as a tree structure.

7.1. GP performance without noise. We simulate 10 time series each using the two alternative specifications for $f(\cdot)$. We then run the GP algorithm on these 20 data sets and stop the GP search after 100 generations. The parameters of the GP model can be found in the table. The GP algorithm is programmed to produce output suitable for import into the Mathematica computer algebra system. We use mathematica to simplify each of the 20 resulting best expressions in order to see whether GP was able to discover the true data generating process. The results from this experiment were very successful.

Name	Variable	Value
Population Size	M	3000
Maximum number of populations	G	100
Reproduction probability	p(R)	0.10
Crossover probability	p(C)	0.85
Mutation probability	p(M)	0.05
Maximal Initial tree depth	D_i	3
Maximal tree depth after crossover	D_c	10

We simulated 100 pairs of $\{x_n, y_n\}$ according to the deterministic relation: $x_n = n - 50$ and $y_n = 0.4x_n + 0.8\text{sign}(n)x_n^2$. The data follows the well-known modified quadratic relation used in nonlinear earnings-returns models.

We performed a single run of the GP algorithm, stopping the algorithm after 100 generations. The resulting expression had a non-parsimonious form. We programmed the algorithm to produce output ready for import into the Mathematica software algebra system, for simplification. The simplified version of the expression is

$$G(X) = 0.353X + 0.80027\text{sign}(X)X^2$$

Recall, that the function being estimated is

$$F(X) = 0.4X + 0.8\text{sign}(X)X^2$$

We present a lot of the percentage difference between the two functions, namely, we plot $F(X) - G(X)/G(X)$. As can be seen from figure 7 on the next page, the two functions are very close numerically. The spike on the graph corresponds to the fact that the true function is zero at $X = 0$.

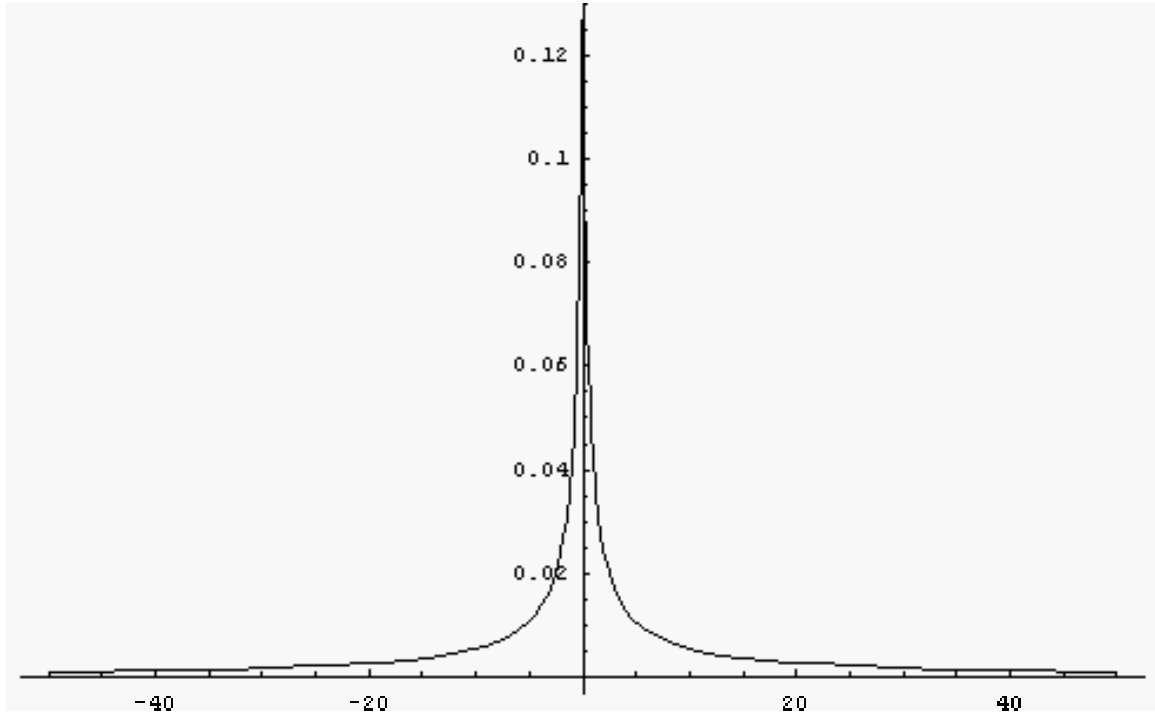


FIGURE 7. GP Algorithm

We also found that in the case without noise, the GP algorithm gravitated either to the correct function, evidenced by a high fitness of the best-of-generation tree structure, or a value very far away from it. In this latter case, several warning signs of a false maximum were present. Overall, the tree structures were short, allowing the GP algorithm to evaluate quickly the fitness of each population. This problem, due to poor initial seeding, resulted in the termination of the GP algorithm in half or even a third of the time required for well-seeded initial populations. We are confident that this false convergence problem can be easily spotted by the researcher.

7.2. GP performance with noise. We simulate ten time series each using the two alternative specifications for $f(\cdot)$. In addition, we add noise to the dependent variable, namely,

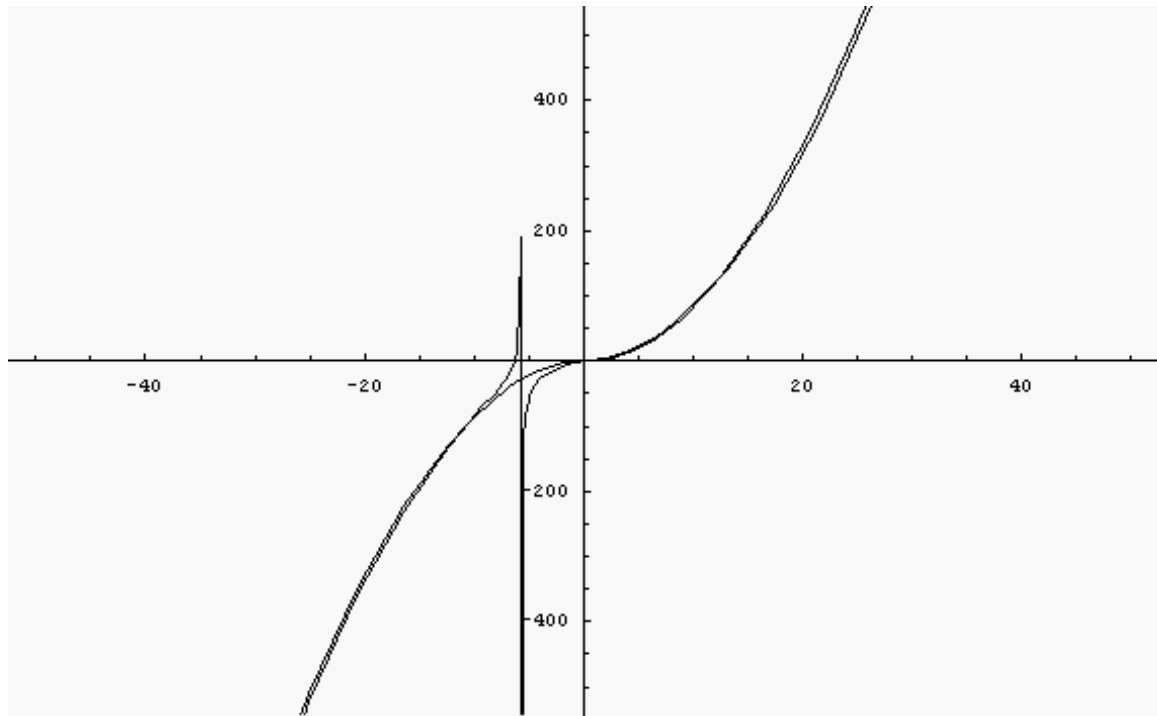


FIGURE 8. Modified Quadratic and GP-generated Function

we construct $y_n = f(x_n) + \epsilon_n$ where $\epsilon_n \equiv \text{iid}N(0, \sigma_\epsilon)$. We report the results of the GP estimation using two values for σ_ϵ , representing different noise level in the system.

We find that when the noise-to-signal ratio is small, GP performs very well in uncovering the true data generating process. However, at higher noise-to-signal ratios, the GP search is not able to discover the true process within 100 generations. This result is consistent with the findings by Kaboudan (2001).

First, we use the GP algorithm on data generated by the modified quadratic process above and σ_ϵ set such that the signal-to-noise ratio, $\text{var}(y_n)/\text{var}(\epsilon_n)$ is 80.

The graphs 8 and 9 on the next page illustrate the closeness of fit of the best tree structures from two GP runs when noise is present.

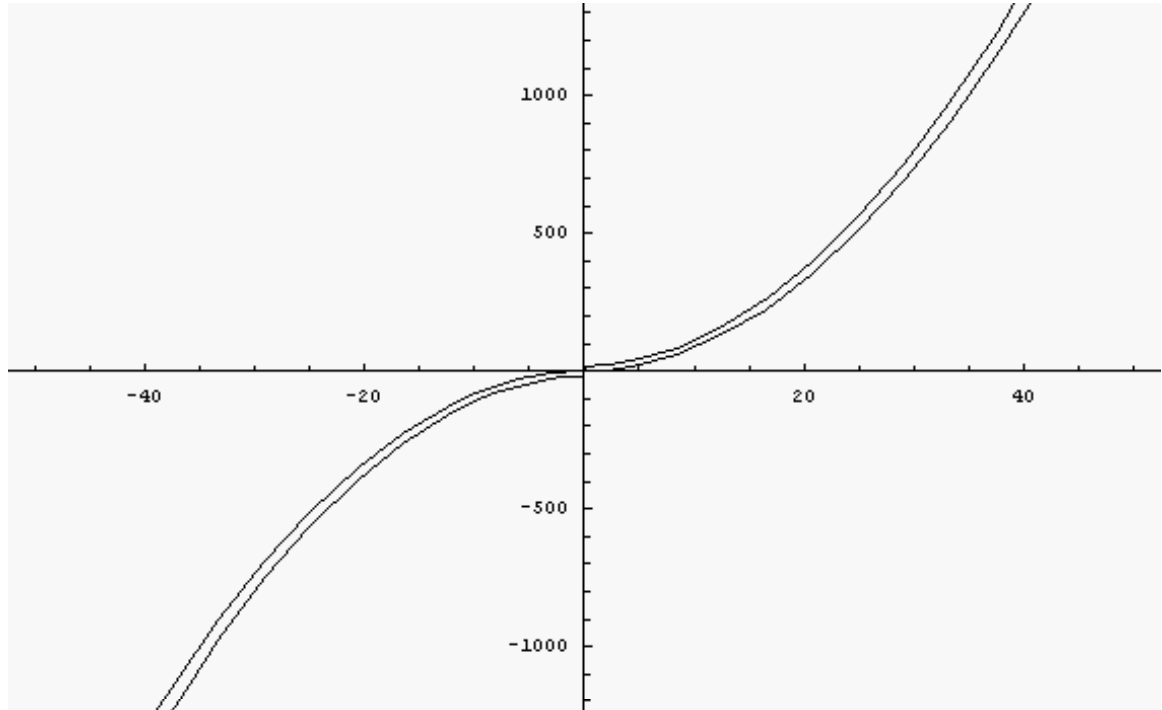


FIGURE 9. Modified Quadratic and GP-generated Function

We see that the fit is good and GP is able to uncover the fundamental shape of the underlying relation. Due to the premium put on parsimony, the resulting GP expressions are cumbersome and involve extra terms which do not have much weight compared to the key underlying relation, unless one is very close to one of their poles.

7.3. Consistency. We perform simulations seeking to determine the rate at which the sum of squared errors from a GP algorithm decays to zero as sample size increases. This is of interest when comparing the performance of the GP algorithm to the performance of parametric and nonparametric estimators. It is inevitable that the extended search space will come at a cost. We are interested in the rate of convergence δ of the sum of squared

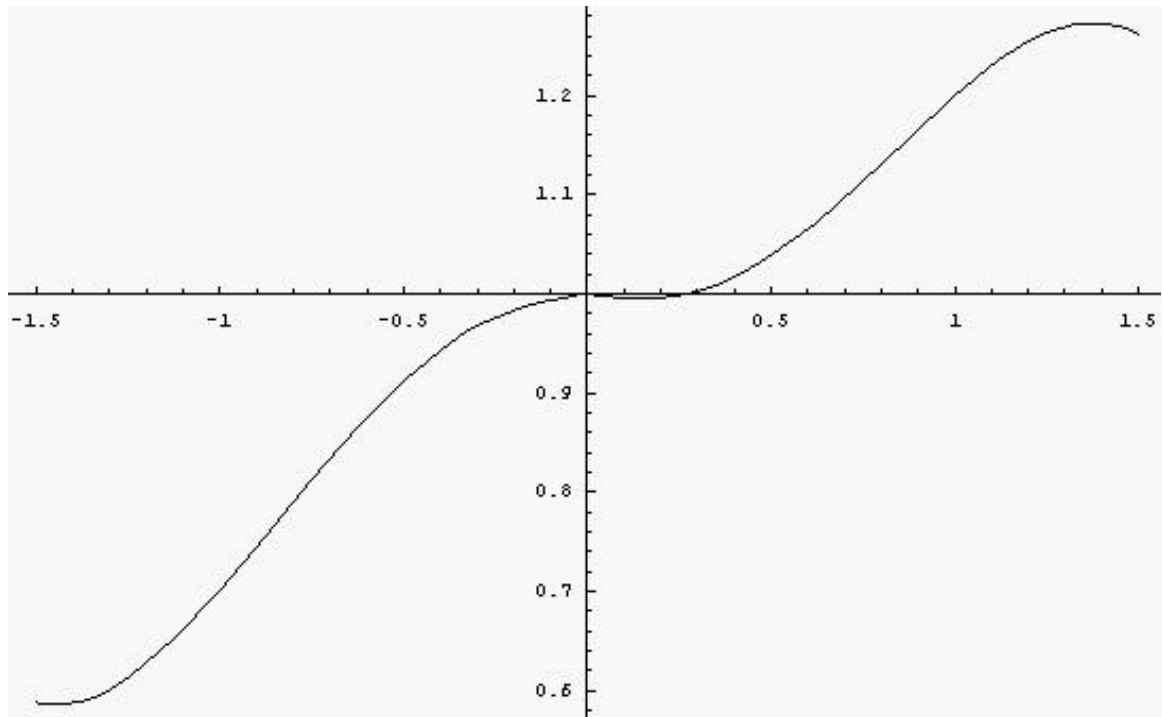


FIGURE 10. Data Generating Process Consistent with the Stylized Earnings-Returns Relation

errors toward zero, as n increases,

$$\lim_{n \rightarrow \infty} SSE(\hat{f}) \sim kn^\delta$$

We defined the following function:

$$1 - (x(0.5x - \text{Sign}(x)) / (x + \text{Sign}[x])(x^2) + 0.05x\text{Sign}(x))$$

The function has the characteristic stylized S-shape of the earnings-returns relation, as shown in figure 10

We simulated 10,000 earnings-returns observations from this data generating process. Then, we performed 250 runs of the GP algorithm on this data set and calculated the SSE for each \hat{f} . Due to the inherent randomness of the initial seeding and the fact that GP is

stochastically driven, this high number of runs was to ensure that the maximum we found is not due to poor initial seeding. We then computed the value $\frac{\ln(1/nSSE(\hat{f}))}{\ln(n)}$ for the model with the lowest SSE. The value is a proxy for the rate of decay of $1/nSSE(\hat{f})$ as $n \rightarrow \infty$, up to a term $\ln(k)/\ln(n)$, where k is a constant. We obtained the value $\delta = 0.938$.

Recall that for most parametric estimators the expected sum of squared error decays to zero at rate $\delta = -1$ when the correct parametric model is used. For nonparametric regression, δ will depend on the conditions imposed on the search space Ξ , and if we assume that \hat{f} is twice-differentiable, for example, $\delta = -4/5$. Thus, we see that the loss of efficiency by GP in this particular illustration is rather small.

7.3.1. *GP Estimation and Closeness to the True DGP.* Lastly, we present an illustration which underscores the difference between GP and the traditional classical and Bayesian modelling frameworks.

We will perform a simulation to match the illustrative examples in Ploberger and Phillips (2003) and for ease of comparison we adopt their notation throughout this section.

Data is generated using the linear regression model

$$y_t = \theta x_t + u_t$$

where $u_t \equiv \text{iid}N(0, 1)$ for $t = 1, 2, \dots, n$. The regressor x_t is generated using each of the following five processes: 1) stationary autoregression $x_t = \alpha x_{t-1} + \epsilon_t, \epsilon_t \equiv \text{iid}N(0, 1), \alpha = 0.5$, 2) Gaussian random walk $x_t = x_{t-1} + \epsilon_t, \epsilon_t \equiv \text{iid}N(0, 1), x_0 = 0.5$; 3, 5) deterministic trends $x_t = t, t^3$.

We use the GP algorithm to estimate the models for sample sizes $n = 10, 11, \dots, 100$ and compute the GP forecast $\hat{y}_n + 1$. That is, we fix n , use GP to search for a tree structure that best describes the data, and using this tree structure we forecast the remaining observations.

Note that the models described by the tree structure need not be linear. We denote the optimal forecast by $\tilde{y}_{n+1} = \theta x_{t+1}$ and compute the forecast divergence

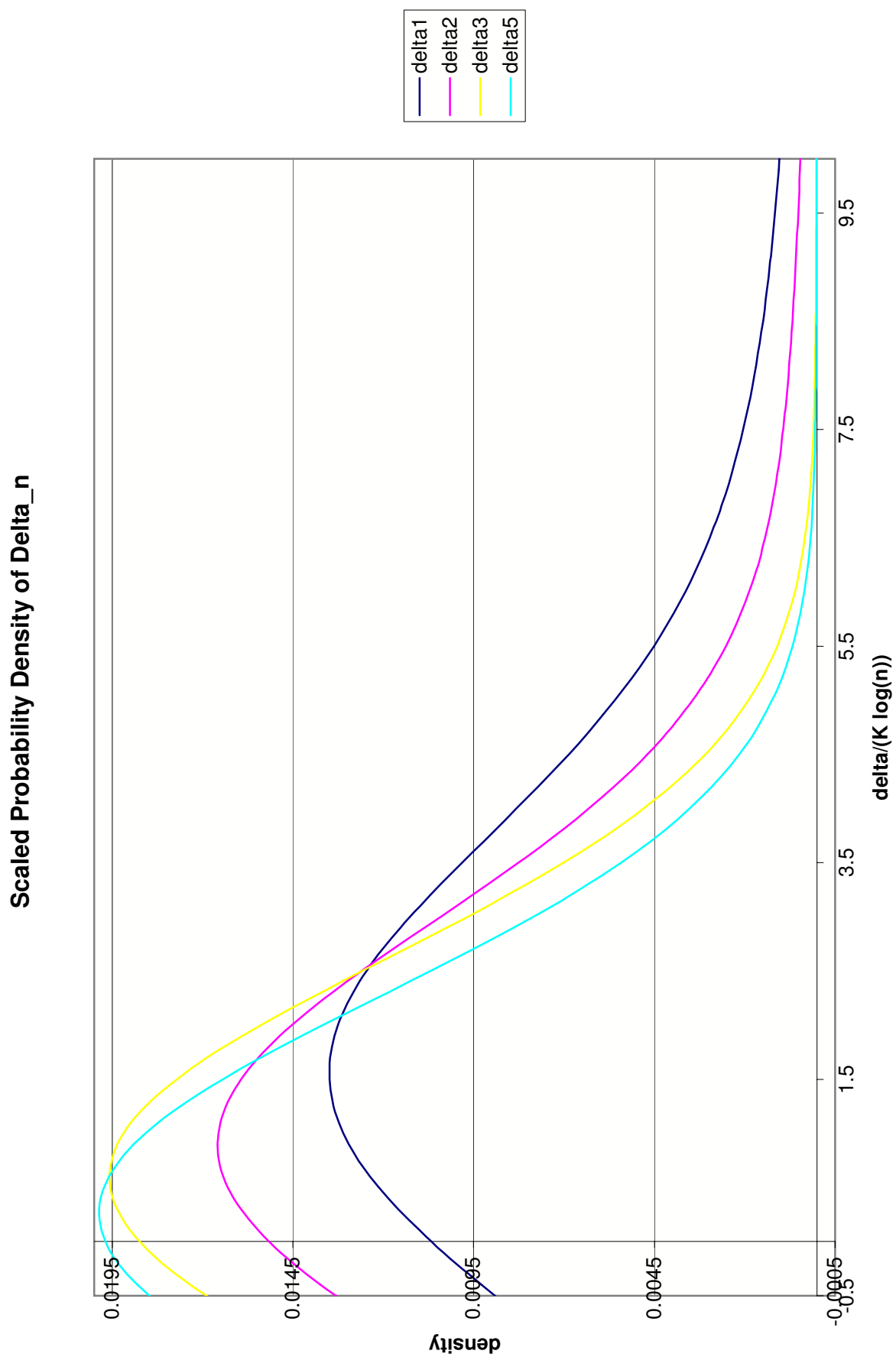
$$\Delta_n = \sum_{t=n_0}^n \{(y_t - \hat{y}_t)^2 - (y_t - \tilde{y}_t)^2\}$$

starting from $n_0 = 10$.

We compute $R = 1,000$ replications of Δ_n and compute kernel estimates of the density function of Δ_n for each of the five generating mechanisms for x_t described above.

Ploberger and Phillips (2003) shows that in the classical estimation case, the closeness that an empirical model can achieve to the true DGP depends on the nature of the regressor, in particular, whether it is a time trend, a stochastic trend, or stationary. More precisely, the bound for the minimal information loss has asymptotic form $(K/2) \log n$, where $K = \sum a_i$ where $a_i = 1$ for stationary regressors, intercepts and dummy variables, $a_i = 2$ for stochastic regressors, and $a = 3$ for linear time trends. Thus, K depends on the nature of the regressor, as well as the number of the regressors. We perform the standardization of Δ_n using the factor $K \log n$ and produce the kernel density graph of $\text{pdf}(\Delta_n/(K/\log n))$. As can be seen, the standardization overcompensates in the case of trending regressors.

Figure 11 on the next page of $\text{pdf}(\Delta_n/(K/\log n))$ in the four cases for x_i reveals that when using the GP estimation methodology, the densities do not suffer from a marked standardization problem, as was the case in Ploberger and Phillips (2003). Thus, it seems that GP forecasting is more forgiving than traditional forecasting methods in the case of trending series.

FIGURE 11. Scaled pdf(δ_n)

8. DATA SAMPLE

We obtain data on analysts' earnings per share (EPS) forecasts for the universe of US firms followed by the Institutional Brokers Estimate Systems (I/B/E/S) between 1974 and 2003. As we discussed earlier, there are some econometric problems associated with estimating earnings regressions from pooled data. Furthermore,

For our analysis we chose to use data on International Business Machines Corporation (NYSE: IBM) from 1984 to 2003. We selected all available EPS median estimates for IBM between 1984 and September 2003. We also required that actual EPS figures were announced as of September 30, 2003. This gave us 75 quarterly median EPS estimates. The median estimates were used as they are less prone to outliers and typographical errors, although I/B/E/S does review for possible errors all individual analyst estimates that are reported to it.

At the time of writing, IBM Corp. is a premier international computer manufacturer and services company with over \$20 billion in quarterly revenue and is actively followed by 22 brokerage houses. We chose the stock because IBM's is a well-known stock trading on NYSE in an efficient market with wide analyst coverage. We can count on the presence of an immediate and consistent link between unexpected earnings and excess stock price returns. Stock price data for IBM was obtained from Bloomberg Systems. In order to calculate abnormal returns we used the NASDAQ Composite Index (Bloomberg: CCMP), which is a valuation-weighted index of all common stocks listed on NASDAQ.

We calculate abnormal returns around each earnings announcement date, using a 5-trading-day window starting one day before the announcement (day -1) and ending in the

third day after announcement (day +3). Stock and Nasdaq returns are calculated as log-returns, $r_t = \log(P_{t+3}/P_{t-1})$, where P_t is the last price on the day before earnings were announced. Abnormal returns are calculated as buy-and hold returns from holding IBM stock in excess of what an investor would have earned if he held the NASDAQ portfolio of companies.

Our data on quarterly earnings forecasts come from the I/B/E/S historical database. I/B/E/S collects EPS forecast data through the first half of the final month of the quarter when a firm announces its quarterly results (Form 10-Q). Thus, we need to be careful to avoid the possibility that estimates contain information from pre-announcements. As Skinner and Sloan (2001) note, more than 75% of EPS pre-announcements are made within the last 12 trading days of the final month of the quarter.

Unexpected earnings are calculated as the difference between actual announced earnings, minus the mean earnings estimate produced during the one-month period before the earnings announcement, as reported by I/B/E/S. Since our goal is to explore nonlinearity in the unexpected earnings variable, we do not use any additional regressors, although the logical next step, after finding a suitable functional form for the earnings-returns relation, would be to explore whether additional variables can be incorporated in the analysis.

We also construct the variable POSSUP which takes the value of 1 if the earnings surprise was positive, -1 if it was negative, and 0 if the earnings surprise was zero. This variable allows GP to construct tree structures allowing for asymmetric reaction of returns to positive and negative surprises—a key stylized feature of the data on the earnings-returns relation.

In order to mitigate the effect of pre-announcements, we compute several measures of abnormal return, closely mirroring previous studies. The first (longest) interval measures

abnormal returns from two days after the announcement of earnings for the previous quarter to two days after the announcement of earnings for the current quarter. Quarterly announcement dates are given by I/B/E/S. This measure of abnormal return we call the full abnormal return. It is a noisy measure of the reaction to earnings. The full abnormal return period is divided into two equal parts. The first sub-interval starts two days after the prior-quarter earnings announcement and ends thirteen trading days before the end of the current quarter. The second sub-interval begins twelve trading days before the current fiscal quarter ends and extends two days after the current earnings are announced. This second sub-interval is likely to capture best the stock reaction to both positive and negative earnings announcements.

We also construct a very short-window measure of abnormal return, designed to capture only reaction to the earnings announcement. The window begins one day before the earnings announcement and ends two days after the announcement. This abnormal return we call immediate return. It is likely to exclude any pre-announcements (usually negative) and thus underestimate the reaction to negative EPS surprises.

The data on unexpected earnings and returns of IBM stock is summarized in figure 12 on the following page.

9. RESULTS AND DISCUSSION

9.1. Data on a single company: IBM. We perform 10 runs of the GP algorithm using data on abnormal earnings and returns constructed from IBM's stock price and median analyst EPS estimates. Similar to Beenstock and Szpiro (2002) we found that GP performed better when we used a constant factor to scale the two variables, cumulative abnormal returns (CAR_t) and unexpected earnings (UE_t). Each run uses the data on abnormal

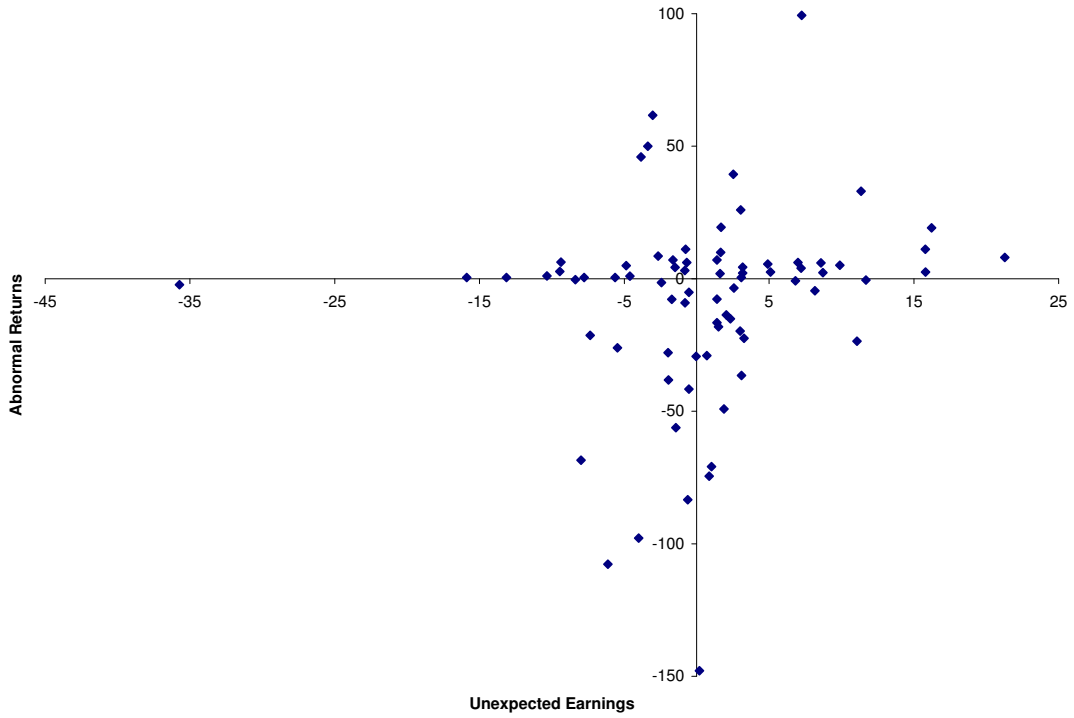


FIGURE 12. Earnings>Returns for IBM Stock

returns and unexpected earnings to search through the space of tree structures for a likely generating process that reflects the data. The output of each run is, as expected, highly parsimonious, reflecting the model's design that puts parsimony at a premium.

We have enabled the program to produce output suitable for immediate \LaTeX inclusion and also for importing within the Mathematica computer algebra system. The latter allows us to simplify the functions produced by the GP algorithm.

The function with best fit, produced by the GP algorithm, from among 100 runs of the algorithm (after simplification in Mathematica) is as follows:

$$(9.1.1) \quad CAR_t(UE_t) = 0.703 - \frac{0.519}{0.468 + 0.159UE_t} + 2 * \text{sign}UE_t$$

A plot of this function appears in figure 13 on the next page.

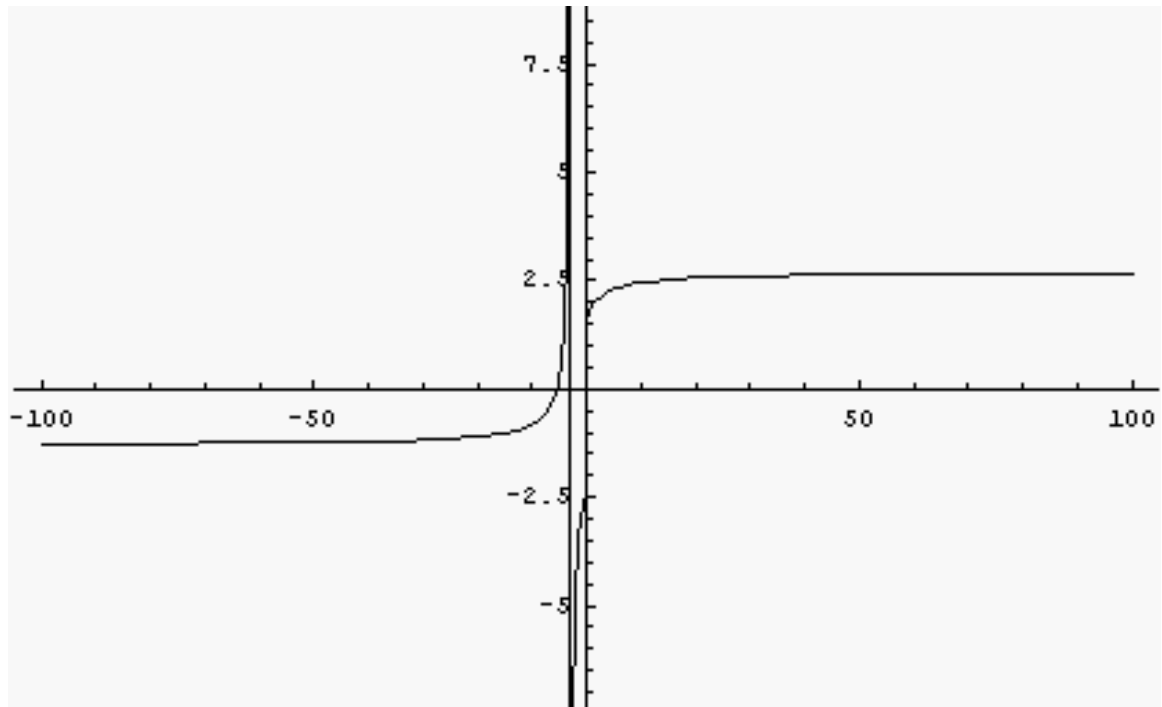


FIGURE 13. GP-Generated Function of the Earnings-Returns Relation

Its fit against the actual data is plotted in figure 14 on the following page.

An inspection of the graph reveals that the function does follow the stylized S-shape characteristic of the functions used to estimate the earnings-returns relation. Generally, the second derivative of the function is negative for large positive values of the earnings surprise and positive for large negative earnings surprise. This is consistent with the theory that extreme values of earnings reflect transitory components and as such do not have as profound influence on the value of the stock. Freeman and Tse (1992)

We also note that the function has an asymptote for small negative value of unexpected earnings. Graphically, this allows the function to “cover” a lot of vertical ground within a small variation of the abscissa variable. This was expected of the GP algorithm since for small negative values of unexpected earnings, data shows highly irregular response in

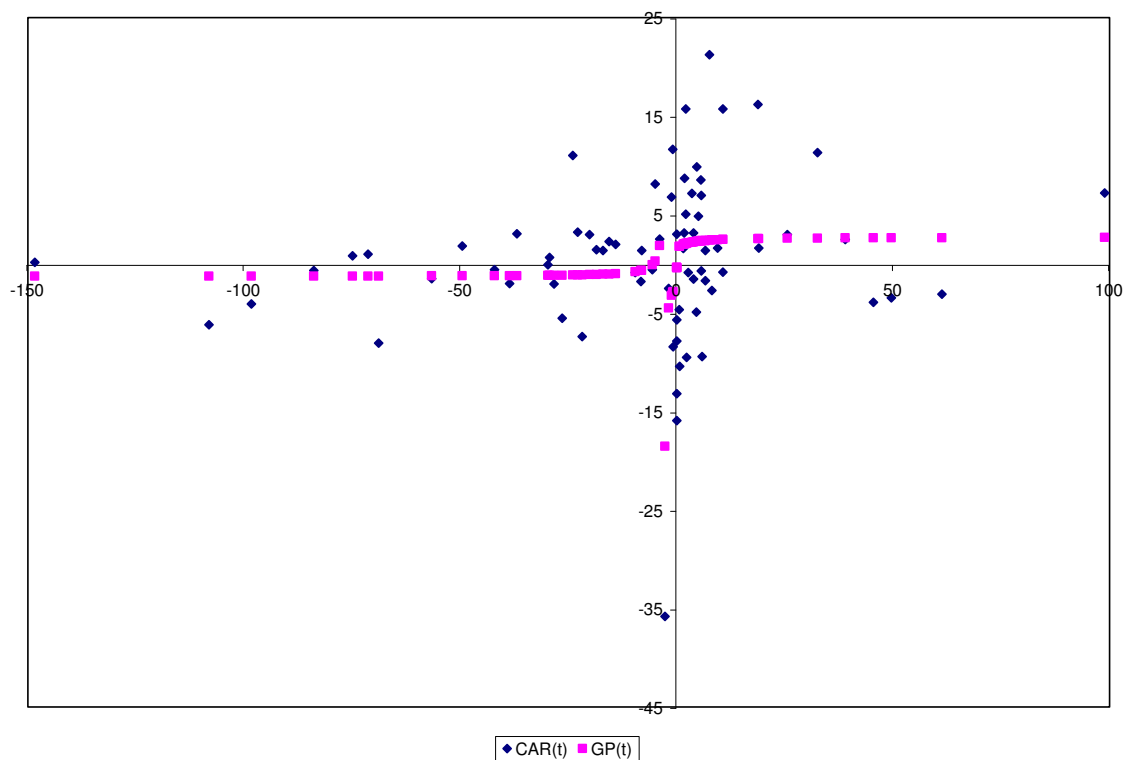


FIGURE 14. GP-Generated Function of the Earnings>Returns Relation

abnormal earnings, exhibiting great variation. This is consistent with the idea that CFO's are usually able to manage earnings expectations through pre-announcements and are also able to manipulate earnings numbers so that they meet or beat expectations. With this in mind, then expectations are not met, the market is unclear whether the small shortfall is a symptom of a much larger problem, thus the variation in responses to a slight shortfall in earnings is much greater than for larger negative earnings surprises.

Perhaps a surprising feature of the GP-generated function is the term $2 * \text{sign} U E_t$, which implies a stronger market response for positive earnings surprises than for negative ones. Two observations are interesting here: one is the fact that the difference in response does not depend on the size of the earnings announcement (as was the case in the modified quadratic

SEARCH FOR A STRUCTURAL SPECIFICATION OF THE EARNINGS-RETURNS RELATION⁵³
functional form used by Freeman and Tse (1992)). The other interesting observation is that this extra term implies that a positive earnings surprise elicit stronger price response than negative earnings surprise of the same magnitude. This latter observation is surprising, and we would be interested to see if it persists in our cross-sectional studies of the earnings-returns relation.

10. MODEL ANALYSIS

Using the functional form given by the GP algorithm, we estimate model parameters. We can generalize the functional form in equation 9.1.1 on page 50 as

$$y_i = \beta_1 + \frac{1}{\beta_2 + \beta_3 * x_i} + \beta_4 \text{sign}(x)$$

Table 1 shows the estimated parameters, together with their standard errors.

TABLE 1. GP Specification

Variable	GP value	Estimated Value	Std. Error	z-Statistic
β_1	0.703	1.036	0.122	8.51
β_2	0.902	0.906	0.144	6.27
β_3	0.306	0.319	0.052	6.08
β_4	2.000	0.899	0.126	7.16

The traditional linear specification is given in table 2 on the next page:

TABLE 2. OLS Specification

Variable	Estimated Value	Std. Error	z-Statistic
α_1	0.726	0.947	0.76
α_2	0.032	0.025	1.27

The arctan specification

$$y_i = \delta_1 + \delta_2 \arctan(\delta_3 x_i) + \epsilon_i$$

used in Freeman and Tse (1992) is estimated in table 3:

TABLE 3. Arctan Specification

Variable	Estimated Value	Std. Error	z-Statistic
δ_1	0.803	0.118	6.79
δ_2	0.584	0.099	5.91
δ_3	1.772	2.675	0.66

The modified quadratic used in Beneish and Harvey (1998) and in earlier drafts of Freeman and Tse (1992) is as follows:

$$y_i = \gamma_1 + \gamma_2 x_i + \gamma_3 * \text{sign}(x_i) * x_i^2 + \epsilon_i$$

and is estimated in the table 4 on the next page.

TABLE 4. Modified Quadratic Specification

Variable	Estimated Value	Std. Error	z-Statistic
γ_1	0.727	0.952	0.76
γ_2	0.059	0.061	0.96
γ_3	-0.0003	0.0006	-0.47

We now present F-statistics and the associated p-values for testing the null hypothesis that the linear OLS specification 2 on the facing page is not statistically different from the equation 1 on page 53 identified using the GP algorithm. The use of the Fstatistic was suggested by Cleveland and Devlin (1988) and relies on a Chi-squared approximation. We base the test on the difference between the sum of squared errors of the alternative specification (GP) and the original specification (OLS). The statistic has as numerator the reduction in residual sum of squares due to the alternative specification and as the denominator the residual sum of squares of the original specification. The test is analogous to ANOVA in a parametric context.

The F-statistic is reported here for three cases for three null hypotheses 1) Null of OLS specification; 2) Null of modified quadratic, 3) Null of Arctan.

We now perform a J-test, as proposed by Davidson and MacKinnon (1993). The test provides a method for choosing between alternative, nonnested models. The idea is that if one of the models is the correct one, then the fitted residuals from the other model should not have any explanatory power when estimating the true model. We have thus the null hypothesis H_0 : linear model is the correct specification, H_1 : GP specification.

We calculate the fitted values from the GP model estimated above and use them as an additional regressor in the linear model. The table 5 presents the regression results. The coefficient on the fitted values of the GP model is highly significant and close to 1, implying that the GP specification contributes information that is not captured within the linear specification. Notably, the GP model does not contain a linear term. Thus, it is important to verify if a linear term can add information not already contained in the GP model. We calculate the fitted values from the linear model 2 on page 54 and use them as an additional regressor in the GP specification. The results are in table 6 on the facing page. As can be seen, the linear model does add information not contained in the GP specification. However, the coefficient on the fitted values is only 1/2 implying that the quantitative contribution of the linear specification is small compared to the quantitative contribution of the GP model.

TABLE 5. Incremental information
of GP model in the Linear specification

Variable	Estimated Value	Std. Error	z-Statistic
α_1	0.176	0.032	5.50
α_2	-0.002	0.001	-1.91
\hat{y}_{GP}	0.994	0.004	253.48

TABLE 6. Incremental information
of linear model in the GP specification

Variable	Estimated Value	Std. Error	z-Statistic
β_1	0.768	0.133	5.74
β_2	1.050	0.194	5.40
β_3	0.371	0.070	5.26
β_4	0.500	0.163	3.06
\hat{y}_{linear}	0.533	0.130	4.10

11. CONCLUSION

We have documented the existence of a nonlinear functional form which is able to explain the reaction of stock prices to earnings surprise better than the traditionally used linear model.

We also confirm that the linear model does indeed contain information not wholly captured by the nonlinear specification. Thus, we advocate the extensive further testing of nonlinear model specifications of the form identified in our analysis.

REFERENCES

- Ball, R. and P. Brown, 1968. An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research*, 6: 159–178. Supplement.
- Beenstock, Michael and George Szpiro, 2002. Specification Search in Nonlinear Time Series Models Using the Genetic Algorithm. *Journal of Economic Dynamics and Control*, 26: 811–835.
- Beneish, Messoud D. and Campbell R. Harvey, 1998. Measurement Error and Nonlinearity in the Earnings>Returns Relation. *Review of Quantitative Finance and Accounting*, 11: 219–247.
- Brooks, L. D. and D. A. Brookmaster, 1976. Further Evidence of the Time Series Properties of Accounting Income. *Journal of Finance*, 31: 1359–73.
- Cheng, C. S., W. S. Hopwood, and J. C. McKeown, 1992. Nonlinearity and Specification Problems in Unexpected Earnings Response Regression Model. *The Accounting Review*, 67: 579–598.
- Cleveland, William S. and Susan J. Devlin, 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403): 596–610.
- Collins, D. W. and S. P. Kothari, 1989. An Analysis of the Inter-Temporal and Cross-Sectional Determinants of Earnings Response Coefficients. *Journal of Accounting and Economics*, 11: 143–181.
- Das, Somnath and Baruch Lev, 1994. Nonlinearity in the Returns-Earnings Relation: Tests of Alternative Specifications and Explanations. *Contemporary Accounting Research*, 11: 353–379.

- Davidson, Russell and James G. MacKinnon, 1993. *Estimation and Inference in Econometrics*. Oxford University Press.
- Easton, P. and M. Zmijewski, 1989. Cross-Sectional Variation in the Stock Market Response to the Announcement of Accounting Earnings. *Journal of Accounting and Economics*, 11: 117–141.
- Freeman, Robert N. and Senyo Y. Tse, 1992. A Nonlinear Model of Security Price Responses to Unexpected Earnings. *Journal of accounting research*, 30: 185–209.
- Holland, John H., 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press.
- , 1992. *Adaptation in Natural and Artificial Systems*. MIT Press.
- Kaboudan, M. A., 2000. Genetic Programming Prediction of Stock Prices. *Computational Economics*, 16: 207–236.
- , 2001. Genetically Evolved Models and Normality of their Fitted Residuals. *Journal of Economic Dynamics and Control*, 25: 1719–1749.
- Kormendi, R. and R. Lipe, 1987. Earnings Innovations, Earnings Persistence, and Stock Returns. *Journal of Business*, 60: 323–345.
- Kothari, S. P. and Jerold L. Zimmerman, 1995. Price and Return Models. *Journal of Accounting and Economics*, 20: 155–192.
- Koza, John R., 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press.
- Lipe, Robert C., Lisa Bryant, and Sally K. Widener, 1998. Do Nonlinearity, Firm-Specific Coefficients, and Losses Represent Distinct Factors in the Relation Between Stock Returns and Accounting Earnings? *Journal of Accounting and Economics*, 25: 195–214.

- Liu, Jing and Jacob Thomas, 2000. Stock Returns and Accounting Earnings. *Journal of Accounting Research*, 38: 71–102.
- Merton, R., 1987. On the State of the Efficient Market Hypothesis in Financial Economics. In R. Dornbush, S. Fischer, and J. Bossons, eds., *Macroeconomics and Finance: Essays in Honor of Franco Modigliani*. Cambridge: MIT Press.
- Penman, S. H., 1992. Financial Statement Information and the Pricing of Earnings Changes. *The Accounting Review*: 563–577.
- Ploberger, Werner and Peter C. B. Phillips, 2003. Empirical Limits for Time Series Econometric Models. *Econometrica*, 71(2): 627–673.
- Skinner, Douglas and Richard Sloan, 2001. Earnings Surprises, Growth Expectations and Stock Returns or Don't Let an Earnings Torpedo Sink Your Portfolio. University of Michigan Business School mimeo.
- Teets, Walter R. and Charles E. Wasley, 1996. Estimating Earnings Response Coefficients: Pooled versus Firm-Specific Models. *Journal of Accounting and Economics*, 21(3): 279–95.
- White, H., 2000. A Reality Check for Data Snooping. *Econometrica*, 68: 1097–1126.