Financial Data Mining with Genetic Programming: a Survey and Look Forward [1]

Nicolas NAVET
LORIA-INRIA
Campus Scientifique B.P. 236
Vandoeuvre 54506, France
E-mail: nnavet@loria.fr

Shu-Heng CHEN
AI-ECON Research Center, NCCU
National Chengchi University
Taipei 11623, Taiwan
E-mail: chchen@nccu.edu.tw

ABSTRACT

Genetic Programming (GP) is an appealing machine-learning technique for tackling financial engineering problems: it belongs to the family of evolutionary algorithms that have proven to be remarkably successful at handling complex optimization problems, and possesses the unique feature of producing solutions under a symbolic form that can be understood and analyzed by humans. Over the last decade, GP has been applied to generate financial trading strategies, forecast stocks and options prices, or grasp some insight into the dynamics of the markets and the behavior of the agents. In this paper, we first provide a brief survey of the existing studies, then highlight fields of investigations that, we believe, should lead to enhance the applicability and efficiency of GP in the financial domain.

1 Relevance of GP for creating trading strategies

Genetic programming (GP) applies the idea of biological evolution to a society of computer programs. Specifically, in financial trading, each computer program represents a trading system - a decision rule - which when applied to the market provides trading recommendations. The society of computer programs evolves over the course of the successive generations until a termination criterion is fulfilled, usually a maximum number of generations or some property of the best individuals (e.g., stagnation for a certain number of generations, a minimum performance threshold is reached). Classical genetic operators, namely mutation, crossover and reproduction, are applied at each generation to a subset of individuals and the selection among the programs is biased towards the individuals that constitute the best solutions to the problem at hand.

In the 80s, economists began to be interested in the idea of evolving populations of decision rules[2] because of the close similarity with the economic agents who are constantly revising - adapting - their own decision rules as they gain experience and as their environment undergo changes. Since then, evolutionary models have proved to be a powerful toolkit for modeling and understanding the behavior of societies of "*imperfectly smart agents exploring their way into an essentially infinite space of possibilities*" (in the words of J. Holland, see (Wal92)). In line with what has just been said, it is clear that evolutionary techniques, such as Genetic Algorithms and Genetic Programming, are relevant

[1] This paper will be presented at the 56th Session of the International Statistical Institute (ISI 2007), Lisboa, August 22-29, 2007. Contact author: Nicolas Navet.

[2] John Holland's work on Genetic Algorithm certainly had a great influence, see (Che01) for a review of evolutionary economics.

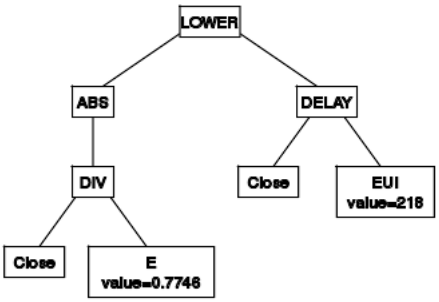

Figure 1: Example of a simple trading rule obtained by GP (NYSE Citigroup Inc. EOD time series - same experimental setup as in (NC07)). It can be noticed that the ABS primitive (*i.e.* absolute value) is extraneous here (an "*intron*" in GP terminology), however it may find its usefulness in descendants of this individual.

to serve as devices to generate financial trading rules, and indeed GP in particular has been already quite often used for that purpose[3]. A simple example of a typical trading rule is given in Figure 1.
The distinguishing trait of GP with regard to almost any other machine-learning tool is that GP does not assume a predefined size and shape for the decision rules: the functional form, along with the value of the parameters, is induced from the training data and the objective function. This is a chance but also a challenge since the search space is of very high dimension, and a crucial question is thus how to design the GP so that the search is likely to be directed towards good solutions. This will be at the heart of the research directions highlighted in Section 3.

2 Financial knowledge discovery with GP

Prominent examples of GP used to discover knowledge can be found in the work of John Koza who, for instance, employed GP to rediscover some basic physical laws from experimental data, in particular Kepler's third law (Koz92). In that experiment, GP not only manages to rediscover Kepler's third law but, along the evolution process, it also rediscovered an earlier conjecture. GP was thus demonstrated as a tool that is helpful to discover knowledge.
However, economics in general, and finance in particular, does not obey time-unvariant deterministic laws, such as Kepler's laws of planetary motions, and the discovery of the rules as well as the interpretation of the results can be expected to be more involved. Indeed, in the financial literature, there are few clear-cut positive outcomes as the aforementioned Koza's result. A more thorough review of the applications of GP to knowledge discovery is given in (CK03a), we should mention here only a few results. In particular (NW99) where, by examining the structure of the trading rules, the authors highlight that the *interest differential* is the most important input to the trading rules in the foreign exchange markets. In (CY96), the authors apply genetic programming to rediscover the *efficient market hypothesis* (EMH), then, in (CY97), they provide an explicit measure of predictability expressed in terms of search intensity that provides an alternative formulation of the EMH.
The list given here is clearly not exhaustive but the results in the literature are indeed scarce, and this can not be explained alone by the difficulty of the task, but mainly because GP has raised much more interests as a tool to generate profitable trading strategies than as a tool to discover knowledge. In actual fact, the results of applying GP for market-timing decisions are typically not very convincing,

[3]The reader may for instance refer to (CKH07, NWD97) for GP applied to trading in foreign exchange markets, (AK99, CKH07, PSV04) in stock markets, (Wan00) in future markets and (Keb99, CYL98) for GP used for pricing options.

and other techniques may possibly be better suited in that regard. However, as pointed out in (Kei02), GP has a major interest in scientific discovery, which is "*its ability to generate a large number of different, yet meaningful hypotheses in a very short amount of time*" and propose solutions "*that are non-intuitive and sometimes provocative*". In our view, GP has not been yet used at the fullest of its potential in knowledge discovery in the financial domain and one should expect many more applications of GP in this line of research. For instance, we believe that GP could be successfully used to get insight into the practice of investors, in the line of (WCFW98), to study the changing characteristics of the markets (LPJ+06), or the specific effects of some regulations rules as the "*uptick rule*".

3 Improvements ahead of us

GP has been applied to the financial domain for the last ten years but it turns out that the number of studies published is still rather limited[4] and many questions are left unanswered. In this section, we identify several lines of research, inspired from what has been done in other machine learning fields or aimed at better addressing the specificities of the financial domain, which, we believe, may improve the efficiency of GP as a tool to find trading strategies.

3.1 Selecting the right instruments

When GP is applied to the financial domain, there are two main reasons why it may be unsuccessful at producing good results: either the design of GP is wrong (e.g., bad choices for the set of terminals, insufficient search intensity), or there might be no way to take advantage of the training set to come up with good solutions, simply because the market is efficient. This latter problem could be overcome by selecting instruments whose price time series are evidenced to embed temporal dependencies, and are thus, to some extent, potentially predictable. Numerous metrics[5], emerging from the fields of information theory, the study of dynamical systems and algorithmic complexity or statistics, have been devised to quantify the predictability of a system observed by the data it produces. One can mention the Lyapunov exponent, which is a measure of the rate of divergence of nearby trajectories and thus an indication of the short-term predictability, the entropy rate which measures the uncertainty that remains in the next information produced given complete knowledge of its past or the Grassberger-Crutchfield-Young statistical complexity which informs us of the amount of information which is relevant to the system's dynamic.

The correlation between the predictability of a time series and the profitability of GP induced rules, and more generally of any trading strategies, is an intriguing and still open question, whose answer constitutes, in our view, a major step towards efficient market timing decision tools. A first step in that direction is proposed in (NC07) where an estimate of the entropy rate is used to evaluate the predictability of the price time series of the stocks composing the NYSE US 100 index. As the left-hand distribution in Figure 2 shows, the price time series of NYSE U.S. 100 stocks do not all have equal entropies. Furthermore, surrogate testing with shuffled time series (the corresponding distribution of the entropy rate is shown in the right-hand graphic of Figure 2), suggests to us that there are temporal dependencies in the time series.

However, if a predictability test tells us about the existence of temporal patterns, it does not give further information on how easy or difficult it is to discover the patterns. In addition, as the abundant literature on the subject suggests, predictability may have a multi-dimensional description, and a single

[4] At the time of writing, the Genetic Programming Bibliography, located at url `http://liinwww.ira.uka.de/bibliography/Ai/genetic.programming.html` returns 67 documents with a search on the keyword "finance", out of more than 5500 references in the database.

[5] The reader interested in predictability measures can refer to (BCFV02) and (Sha06) for comprehensive surveys.

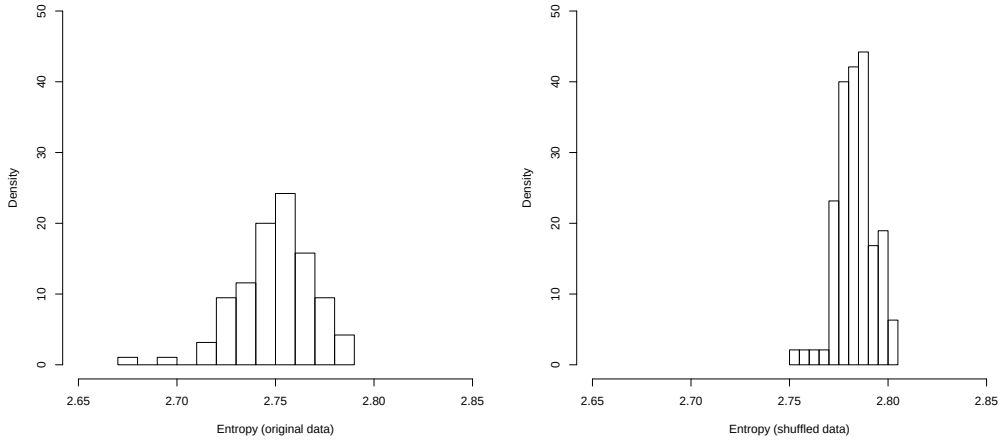

Figure 2: Distribution of entropy rates of the price time series of the NYSE US 100 stocks (left-hand graphics) and shuffled price time series (right-hand graphics - 1000 shuffled time series for each stock). The time series are processed so that the data points are the log ratios between consecutive daily closing prices: $r_t = \ln(p_t/p_{t-1})$ and points are then further discretized into 8 distinct states (the maximum theoretical entropy is 3). Log x-axis ranges from 2.65 to 2.85 on both graphics.

measure of predictability may not be enough to capture all of its attributes, that is why further studies about the relation between predictability and profitability should probably not rest upon only a single predictability measure.

3.2 Rigorous assessment of the GP outcomes

Most studies on GP select a risk-free investment (e.g., treasury bills) or, most often, the buy-and-hold strategy as the benchmark to which the GP's outcomes are compared. As highlighted in (CKH07), the conclusion that "*GP performs better than buy-and-hold in a bearish market and worse in a bullish market*" is very often found in the literature. However, nothing different can be expected since buy-and-hold is the worst possible strategy in a steadily decreasing market and the best possible strategy in a steadily increasing market. This shows the limits of choosing buy-and-hold as a benchmark, especially in trendy markets.

Typically, one observes in the literature that the result of applying GP for market-timing decisions is not very convincing, but the investigators always suggest the possibility of further improvements without really convincing that there is something to learn from past data (*i.e.* that market is not fully efficient) and that GP is suitable for this task. In (CN06), the problem is addressed by proposing a series of pretests aimed at giving more clear-cut answers as to whether GP can be effective with the training data at hand. Precisely, pretesting allows to distinguish between a failure due to the market being efficient or due to GP being inefficient. The basic idea is to compare, using statistical tests, the outcomes of GP with the outcomes of several variants of random searches (*"zero-intelligence strategies"*) and random trading behaviors (*"lottery trading"*) having well-defined characteristics. In particular, if the outcomes of the pretests reveal no statistical evidence that GP possesses a predictive ability superior to a random search or a random trading behavior, then this suggests that there is no point in investing further resources in GP.

The study published in (CN06) is a first step towards establishing well-defined statistical techniques for analyzing the GP outcomes. More broadly, sound experimental research methodologies in the vein of (BB03) are needed to improve the assessment, the understanding and the comparability of GP-based

studies.

3.3 Reducing variability of the results

Everyone having done experiments with GP has noticed that the outcomes of GP are very variable from run to run. In our experience, this is something that happens, in a more or less acute manner, whatever the problem at hand (see for instance the experiments on various problems in (GSPT06)). The high variability of the results constitutes a severe hindrance to the use of GP, especially in the financial domain where controlling the risk is of primary importance. Improvements in the direction of more predictable results are crucially needed and, although to our best knowledge no general solution is known yet, several techniques can be envisaged to alleviate this problem.

A first plausible explanation is that variability might simply be caused by an insufficient search intensity. Usually the GP population is made of a few hundred individuals evolving during at most 100 generations; given the huge search space, this might be insufficient, as some results published in (CK03b) suggest. Increasing the population size, the number of generations, and having possibly several populations that evolve in parallel (*Island* model) may lead to improvements.

The usefulness of validation[6], which has been widely used in the financial domain as a device to fight overfitting (NWD97, AK99), is still an open question. Some studies shows that validation is beneficial in terms of average performance (CKH07), others demonstrate that it helps to reduce the variability (GSPT06), while (CK03b) wonders whether validation is really needed since GP would tend to suffer more from underfitting than overfitting. In the financial domain, something that has to be taken into account is that market characteristics are evolving over time, more or less quickly. It can for instance happen that the strategies created on the training interval might not be suited anymore when used out-of-sample, and the existence of a validation period can aggravate the problem. On the other hand, one may imagine that in more stable markets validation can be helpful. The question of the usefulness of the validation could be revisited in the light of these observations.

3.4 Re-thinking the data-division scheme

There are numerous evidences in the literature (see for instance (CKH07) and (NC07)) that GP is most generally not efficient when the training interval exhibits a time series pattern which is significantly different from the out-of-sample period (e.g., "bull" versus "bear", "sideways" versus "bull", etc). This is not surprising per se since GP is a learning algorithm and it cannot be expected to come up with strategies that are profitable in market conditions that are substantially different from the ones experienced during the training period. The way data are divided, and the re-learning scheme, are thus crucial settings of the GP experimental design, and certainly deserves further studies.

A solution, already widely explored in conjunction with other learning techniques (Lan99), is to re-learn from updated training data if the current performance level is below a given threshold. In the financial domain, a natural choice for the performance metric would be the equity curve: if the current equity diverges too much from an expected equity curve, then a re-learning mechanism would be automatically triggered[7]. The abundant literature on active-learning and incremental-learning should provide us with a good starting point or how to design the mechanisms.

[6] Validation means that the best rules induced on the training interval are further selected on unseen data, i.e., the validation period. The best individual on the validation period is then applied on the testing period.

[7] This is what is called "*trading the equity curve*".

3.5 Preprocessing the data: still an open issue

Data preprocessing serves the purpose of "smoothing" the raw data and removing what is not essential before the machine learning algorithm is applied. It is widely accepted that preprocessing is usually beneficial and, indeed, most studies using GP classically transform the original time series by dividing each day's price by a 250-day moving average ((NWD97, AK99, CKH07)). This way of preprocessing the data is shown to have positive effects in (CKH07) but the general problem of how to best preprocess the data is wide open.
Intuitively, the preprocessing should depend on the market characteristics. In particular, if the market is volatile, one would tend to think that the influence from the past should be limited, which means, for instance, a moving average having a small length. Besides moving averages, there are many other transforms that could be meaningful: log ratio between consecutive values, FFT, wavelets, etc. which one to select and how to define the parameter values is something that has not been investigated yet.

3.6 Re-thinking fitness functions

In (LP02), Langdon and Poli experimentally show that, on some problems, GP is only marginally better than plain random search, and they analyze the underlying reasons. One of the explanations lies in the shape of the fitness landscapes of these problems: they possess characteristics rendering their exploration difficult for GP. Langdon and Poli suggest that one way to alleviate the problem is to re-define the objective - the fitness function - so as to possibly obtain a more "GP-friendly" fitness landscape.
Typically, for financial trading, the performance metric that is used is the rate of return. This may not be the best choice. On the one hand, it might lead to a difficult fitness landscape for GP, and, on the other hand, risk-adjusted metrics could be better since a few lucky trades alone can produce an outstanding rate of returns. The latter problem is particularly acute since the trading frequency of GP-induced rules is typically quite low (e.g., in (CKH07), the trading frequency ranges from 1 to 9 round-trip transactions every two years).
Another device that may prove to be effective is the use of *sensitivity adjusted fitness functions*[8]: that is, adjusting the fitness of an individual depending on where the individual is located in the fitness landscape. If an individual is on a peak (*i.e.* very similar individuals can possess very different fitnesses), its fitness is artificially reduced (for instance, by averaging with the fitness of its neighbors) because there is a good chance that the solution is the result of overfitting the training data and will not be robust when used out-of-sample. With sensitivity adjusted fitness, individuals located on plateaus of the fitness landscape are selected preferentially.

3.7 Embedding more domain specific knowledge

As illustrated by the experiments in (GS04) and (Nav06), the choice of the function set used in GP has a large influence on the quality of the outcomes. Several problems may arise. If unnecessary functions are included, then the size of the search space increases uselessly and computing power is wasted, leading to results of lower quality (CKS02). On the other hand, if necessary functions are not available then much computing power is consumed to create the missing primitives from existing ones, and there are cases where this task may simply be out of reach of GP[9]. In addition, in the process of

[8]Sensitivity adjusted fitness is already implemented in the IO optimizer (Ton07), which is a software for optimizing the parameters of trading strategies.

[9]Let us consider the case of $\sin(x)$, if no other trigonometric functions are available, $\sin(x)$ can be approximated by its Taylor series $\sin(x) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \frac{1}{5040}x^7 + ...$ but coming up with a polynomial leading to a good accuracy, while solving the problem at hand, is obviously no easy task for GP.

creating higher-level primitives, the well documented "code bloat" phenomenon introduces redundant or noisy elements, which may further slow down the evolution process.
Unfortunately, there are no guidelines on how to best select the primitive sets for GP for the problem at hand. However, as the historical development of computational intelligence consistently teaches us, achieving high levels of performance necessitates *extensive domain-specific knowledge* (Fei03). This is a route that has not been taken yet by existing works, the functions used are very primitive[10] and far away from what traders or quantitative analysts employ. In our view, embedding more domain specific knowledge is a very promising and necessary line of investigation. For instance, the terminal set could be enriched with the volume time series, values of some indexes, the bid/ask spread, while the function set could be complemented with technical analysis functions, measures of cross-correlation between instruments, time series predictability estimates, etc. Of course, this would lead to a larger search space and extensive experiments will be needed to figure out which functions and terminals are really beneficial and which ones are "extraneous".

References

[AK99] F. Allen and R. Karjalainen. Using genetic algorithms to find technical trading rules. *Journal of Financial Economics*, 51:245–271, 1999.

[BB03] T. Bartz-Beielstein. Experimental analysis of evolution strategies: Overview and comprehensive introduction. Technical Report Reihe CI 157/03, SFB 531, Universität Dortmund, Dortmund, Germany, 2003.

[BCFV02] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani. Predictability: a way to characterize complexity. *Physics Reports*, 356:367, 2002.

[Che01] S.-H. Chen. *Evolutionary Controversy in Economics Towards a New Method in Preference of Trans-discipline*, chapter On the Relevance of Genetic Programming to Evolutionary Economics. Springer Verlag Tokyo, 2001. ISBN:4-431-70303-9.

[CK03a] S.-H. Chen and T.-W. Kuo. Discovering hidden patterns with genetic programming. In S.-H. Chen and P. P. Wang, editors, *Computational Intelligence in Economics and Finance*. Springer-Verlag, 2003.

[CK03b] S.-H. Chen and T.-Z. Kuo. Overfitting or poor learning: A critique of current financial applications of GP. In C. Ryan, T. Soule, M. Keijzer, E. Tsang, R. Poli, and E. Costa, editors, *Proceedings of the Sixth European Conference on Genetic Programming (EuroGP-2003)*, volume 2610 of *LNCS*, pages 34–46, Essex, 14-16 April 2003. Springer-Verlag.

[CKH07] S.-H. Chen, T.-W. Kuo, and K.-M. Hoi. Genetic programming and financial trading: How much about "what we know". In C. Zopounidis, M. Doumpos, and P. M. Pardalos, editors, *Handbook of Financial Engineering*. Springer, 2007. Forthcoming.

[CKS02] S.-H. Chen, T.-W. Kuo, and Y.-P. Shieh. Genetic programming: A tutorial with the software simple gp. In S.-H. Chen, editor, *Genetic Algorithms and Genetic Programming in Computational Finance*, pages 55–77. Kluwer, 2002.

[CN06] S.-H. Chen and N. Navet. Pretests for genetic-programming evolved trading programs: zero-intelligence strategies and lottery trading. In Irwin King, Jun Wang, Laiwan Chan, and DeLiang L. Wang, editors, *Neural Information Processing, 13th International Conference, ICONIP 2006, Proceedings, Part III*, volume 4234 of *Lecture Notes in Computer Science*, pages 450–460, Hong Kong, China, October 3-6 2006. Springer.

[CY96] S.-H. Chen and C.-H. Yeh. Genetic programming and the efficient market hypothesis. In *Genetic Programming 1996: Proceedings of the First Annual Conference*, pages 45–53, Stanford University, CA, USA, 28–31 1996. MIT Press.

[10] Typically, the function set is comprised of +, -, {*}, /, norm, moving_average, max, min, lag, and, or, not, >, <, if-then-else, true, false.

[CY97] S.-H. Chen and C.-H. Yeh. Toward a computable approach to the efficient market hypothesis: An application of genetic programming. *Journal of Economic Dynamics and Control*, 21:1043–1063, 1997.

[CYL98] S.-H. Chen, C.-H. Yeh, and W.-C. Lee. Option pricing with genetic programming. In John R. Koza, Wolfgang Banzhaf, Kumar Chellapilla, Kalyanmoy Deb, Marco Dorigo, David B. Fogel, Max H. Garzon, David E. Goldberg, Hitoshi Iba, and Rick Riolo, editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 32–37, University of Wisconsin, Madison, Wisconsin, USA, 1998. Morgan Kaufmann.

[Fei03] E. A. Feigenbaum. Some challenges and grand challenges for computational intelligence. *J. ACM*, 50(1):32–40, 2003.

[GS04] W. Gang and T. Soule. How to choose appropriate function sets for GP. In Maarten Keijzer, Una-May O'Reilly, Simon M. Lucas, Ernesto Costa, and Terence Soule, editors, *Genetic Programming 7th European Conference, EuroGP 2004, Proceedings*, volume 3003 of *LNCS*, pages 198–207, Coimbra, Portugal, 5-7 April 2004. Springer-Verlag.

[GSPT06] C. Gagné, M. Schoenauer, M. Parizeau, and M. Tomassini. Genetic programming, validation sets, and parsimony pressure. In P. Collet, M. Tomassini, M. Ebner, S. Gustafson, and A. Ekárt, editors, *Proceedings of the 9th European Conference on Genetic Programming (EuroGP-2006)*, volume 3905 of *Lecture Notes in Computer Science*, pages 109–120, Budapest, Hungary, 10 - 12 April 2006. Springer.

[Keb99] C. Keber. Option pricing with the genetic programming approach. *Journal of Computational Intelligence*, 7(6):26–36, 1999.

[Kei02] M. Keijzer. *Scientific Discovery Using Genetic Programming*. PhD thesis, Danish Technical University, Lyngby, Denmark, March 2002.

[Koz92] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[Lan99] C. Lanquillon. Dynamic aspects in neural classification. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 8(4):281–296, 1999.

[LP02] W. B. Langdon and R. Poli. *Foundations of Genetic Programming*. Springer-Verlag, 2002.

[LPJ+06] J. W. Lee, J. B. Park, H.-H. Jo, J.-S. Yang, and H.-T. Moon. Complexity and entropy density analysis of the Korean stock market. In *Proceedings of the 5th International Conference on Computational Intelligence in Economics and Finance (CIEF2006)*, 2006.

[Nav06] N. Navet. Genetic programming for financial trading: a tutorial. Tutorial at the 5th International Conference on Computational Intelligence in Economics and Finance (CIEF2006), 2006. Slides available at url `http://www.loria.fr/~nnavet`.

[NC07] N. Navet and S.-H. Chen. Entropy rate and profitability of technical analysis: experiments on the NYSE US 100 stocks. In *Submitted to the 6th International Conference On Computational Intelligence in Economics and Finance (CIEF2007)*, 2007.

[NW99] C. Neely and P. Weller. Technical trading rules in the European monetary system. *Journal of International Money and Finance*, 18(3):429–458, 1999. available at http://ideas.repec.org/a/eee/jimfin/v18y1999i3p429-458.html.

[NWD97] C. Neely, P. Weller, and R. Dittmar. Is technical analysis in the foreign exchange market profitable? a genetic programming approach. *Journal of Financial and Quantitative Analysis*, 32(4):405–427, 1997.

[PSV04] J.-Y. Potvin, P. Soriano, and M. Vallée. Generating trading rules on the stock markets with genetic programming. *Comput. Oper. Res.*, 31(7):1033–1047, 2004.

[Sha06] C. R. Shalizi. Methods and techniques of complex systems science: An overview. In T.S. Deisboeck and K.J. Yasha, editors, *Complex Systems Science in Biomedicine*, pages 33–114. Springer Verlag, New-York, 2006.

[Ton07] F. Tonetti. *IO: an Intelligent Optimiser*, 2007. Available at `http://finance.groups.yahoo.com/group/amibroker/files/`.

[Wal92] M. Waldrop. *Complexity: The Emerging Science at the Edge of Order and Chaos.* Simon & Schuster, January 1992.

[Wan00] J. Wang. Trading and hedging in S&P 500 spot and futures markets using genetic programming. *Journal of Futures Markets*, 20(10):911–942, 2000.

[WCFW98] A.S. Weigend, F. Chen, S. Figlewski, and S.R. Waterhouse. Discovering technical traders in the T-bond futures market. In *Knowledge Discovery and Data Mining*, pages 354–358, 1998.