遺伝的プログラミングによる市場価格変動の時系列モデルの構築

Constructing Stochastic Model of Financial Markets using Genetic Programming

吉村勇志[1]　陳昱[1]

[1]東京大学大学院新領域創成科学研究科人間環境学専攻

Abstract: 株式や外国為替等の金融市場において、特徴的な統計性質がいくつか知られており、それらの発生メカニズムを解明する為、様々なモデルが考案されてきた。ところが、研究者自身がモデル構築を行う場合、研究者が事前に想定した複雑さの因果関係までしかモデルに含むことが出来ない。そこで、遺伝的プログラミングによって、市場の統計性質を満たすようにモデルを自動生成する。本研究はその第一歩として fat tail と volatility clustering を再現する確率過程モデルの構築を行った。その結果、fat tail は比較的単純な式で再現出来るが、volatility clustering の再現には煩雑な式が要求されることが判明した。両者の同時再現においては要求される式の複雑さや誤差縮小速度の違いにより、ハイパーパラメータの細かな調整が必要である。

1. 序論

株式や外国為替等の金融市場において、fat tail (FT)や volatility clustering (VC)のような特徴的な統計性質がいくつか知られており、それらの発生メカニズムを解明する為、様々なモデルが考案されてきた、例えば ARCH[1]や GARCH[2]は確率過程で価格変動を表現し、Minority Game[3]やスピンモデル[4]のような人工市場は多数のトレーダーの集団的行動に、Maslov[5]や Farmer[6]は limit order book (LOB)のダイナミクスに注目している。これらのモデルは FT や VC の再現に多かれ少なかれ成功している。

ところが、市場に見られる統計性質は FT と VC のみに限られる訳ではない。モデルによって再現されるべき統計性質は他にも多数存在する。そして、再現すべき統計性質が多くなるに従って要求されるモデルの複雑さが増し、構築が難しくなることは言うまでもない。深層学習のような機械学習を用いればそのような複雑なモデルも構築出来る可能性はあるものの、モデルの中身を人間が解釈し、理論を一般化することが出来ない。そこで、遺伝的プログラミング(genetic programming; GP)[7]を用いて、市場の統計性質を再現し、尚且つ人間に可読な金融市場モデルの生成を試みる。

2. 背景

2.1 遺伝的プログラミングの要点

GP 自体は有名な手法であるので記述をしないが、GP の条件設定に関しては十分に知られていないので簡単に要点を列挙する。これらは何れも文献[7]に基づく。

・解候補の多様性が必要であるから、個体数は最低でも 500 以上、可能ならば 1000 以上が望ましい
・進化計算の初期に本質的な改善を見せるので、世代数は 10 から 50 程度で十分
・新世代の生成は交叉で 90%、突然変異が 1%、残りが旧世代からそのまま引継ぎという割合が典型的ではあるが、突然変異が 50%等、全く異なる数値であっても問題ない
・交叉においては 2 つの式それぞれからランダムに 1 点選び、そこで取り換えるという subtree 選択が一般的、その際個数の多い終端子が選ばれ過ぎないよう、例えば関数を 9 割、終端子を 1 割で選ぶようにする
・突然変異において最も簡単な方法は親となる式から 1 箇所を交叉と同様の方法で選び、ランダムに生成した式と取り換える subtree 突然変異である
・交叉や突然変異において親となる式はトーナメント選択によって行うことが多い
・式の初期生成においては ramped half-and-half で深さ上限 2-6 で作るのが一般的で、人口の半分は深さ上限まで成長させた式、残り半分は深さ上限の範囲内で確率的に成長させた式を用いる

2.2 生成対象の選択

GP は関数や終端子の取り得る範囲内において任意の数式、条件式を生成することが出来、またほぼ

全ての金融市場のモデルは複数の数式と条件式の組み合わせで記述可能である為、金融市場のモデルを対象に GP を実行することは不可能ではない筈である。

例えば、①エージェントが指値注文と成行注文のどちらを選択するかの条件式②指値注文の場合に注文する価格を算出する計算式③指値注文の数量を算出する計算式④成行注文の数量を計算する計算式、という 1 つの条件式と 3 つの数式を与えれば、それで 1 つの人工市場モデルを記述することが出来る。これだけでは売買方向を指定出来ないように見えるが、売買方向をランダムにする、他の数式の計算結果に依存して自動で決定する（指値注文であれば、高い価格は売り、低い価格は買いであると自動的に決め、成行注文の数量が正なら買い、負なら売りとする）等の方法を予め考え、用意しておけばシミュレーションを行うにあたって問題はない。売買方向を独立に決定したければ、モデルを記述する条件式を増やせば良い。キャンセル注文の導入も、同じく条件式や数式の追加で行うことが出来る。これら複数の式からなるモデルに対し、そのシミュレーション結果と実際の市場の統計性質の誤差を小さくするよう式を進化させる。

尤も、複数の式からなるモデルに GP を適用するには、問題及び GP の性質を深く理解し、それに応じた GP のテクニックの起用が必要となる。そこで、先ずは最も単純なモデルとして、1 つの数式のみによって記述される確率過程モデルの生成を行う。

今回対象とする確率過程モデルの大枠を式(1,2)のように設定する。

$$\sigma(t) = f(r(t-1), r(t-2), \dots) \qquad (1)$$

$$r(t) = \sigma(t)\varepsilon \qquad (2)$$

式(1)は過去のリターンから標準偏差を計算する式で、(2)は計算された標準偏差の正規分布から次の時刻のリターンを計算する式で、εは正規乱数である。式(1)の右辺を$a + br^2(t-1)$とすれば ARCH(1)に近い形となる。モデルを実質的に記述するのは(1)の右辺のみである為、非常に簡潔である。

このモデルを GP の対象として選ぶ利点は単純性の他にも存在する。価格変動に関する有名な統計性質には FT と VC の他にリターンの無相関性があるが、これに関しては式(2)により、目的関数に含めずとも自動的に満たされる筈である。モデルの構造故に、GP で生成する数式が満たすべき条件が緩和されている。また、GP の探索可能な範囲に ARCH に近いモデルが含まれている為、GP によって優れた確率過程モデルが生成されなかった場合に、GP の条件設定やテクニックのみに原因を求めることが出来、モデル内の GP によって進化を行わない部分の不備によりどのような式を GP で生成しようとも理想的なモデルを決して作れないということがない。その上探索が上手くいけば、ARCH と異なり過去のリターンの高次の項が式中に現れ、従来知られていなかった複雑な時間相関を発見出来る可能性もある。

3 確率過程モデルの生成

3.1 GP1：VC 目的標準 GP

3.1.1 条件設定

先ずは最も簡単な条件下で GP による確率過程モデル生成を行った。条件は以下を除いて 2.1 と同じである。

・関数として加減乗除の 4 つを採用(除算において/0 は 1 を返すものとする)
・終端子は 1-5 ステップ前までの過去のリターン及び定数から選択し、定数は-5 から+5 までの範囲から 0.1 刻みでランダムに選ばれる。終端子が定数である確率が 5 割、過去リターンである確率が 5 割
・適合度はリターンの絶対値の自己相関を 100 ステップまで、日経平均のものとの誤差によって評価する。誤差は単純に差分の絶対値の総和とする。データの期間は 1991/1/4 から 2015/12/30 であり、日次の終値を使用した
・パラメータはトーナメント選択で競わせる解候補の数が 10、ランダムに式を成長させる時に関数になる確率が 5 割、終端子が 5 割。個体数は 2000
・式のランダム生成における最大の深さを 5 とする
・終端子が定数のみで構成され、定数しか返さない式が生成されないよう、何らかの操作で定数式が生じたら操作をもう一度やり直す

3.1.2 結果

誤差は第 1 世代では 3.89 であったが、15 世代では 1.55 に低下した。進化計算によってモデルが改良されていることが分かる。GP が生成したモデルと日経平均の比較を図 1 に示す。時間遅れが 3-5 ステップ目の相関が強く出てしまっていることを除けば良好な結果である。尚、ランダム式生成における最大深さを 4 とすると誤差が 2.20 と大きくなり、再現が良好とは言えないことから、ボラティリティの時間相関にはある程度複雑な構造が必要だということが分かる。実際、生成された数式も本モデルにおいては

(((PrevR(2)/3.8)-(((PrevR(3)/(((PrevR(3)-PrevR(4))+PrevR(3))/-0.4))+(4.4/((-0.1*PrevR(3))-((PrevR(4)+PrevR(3))*PrevR(3)))))--3))-(((((0.6+(((PrevR(5)+1.5)/(PrevR(4)+PrevR(3)))+0.9))*(-4.3+(PrevR(2)-PrevR(1))))+PrevR(1))/PrevR(2))+(((PrevR(3)-PrevR(4))*(0.6/-1.3))+-3.8)))-(((((4.9+((1.9/PrevR(2))--2))-0.6)+1.5)/(PrevR(4)+PrevR(3)))+((4.3*PrevR(5))/((PrevR(5)-3.4)-PrevR(5))))であるが、最大深さ 4 の場合には((-2.4-(PrevR(1)-(0.5/(1.5/(0.7*PrevR(5))))))*(0.5/((5-2.1)+(0.6*(PrevR(4)/-2.6)))))-((-0.8*PrevR(5))-(((-1.2+PrevR(3))+(-4.9/4.1))/2.1))であり、比較的単純な式であった。

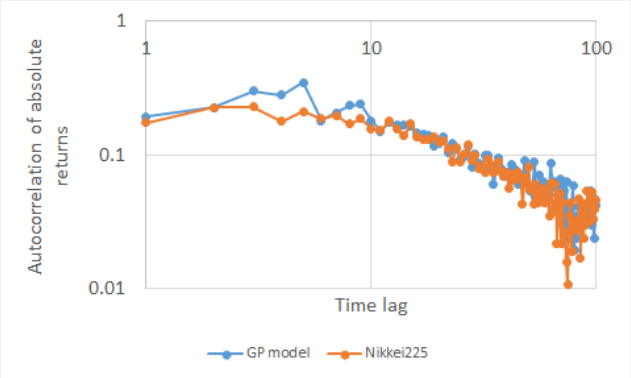

図 1　VC のみを目的関数とした標準的な GP が生成したモデルにおけるボラティリティの自己相関

3.2 GP2：FT 目的標準 GP

3.2.1 条件設定

次に目的関数を VC ではなく FT に変更する。使用する日経平均のデータにおいてリターンのサンプルサイズが 6135 であるから、シミュレーションの時間ステップ数も調整し、同じく 6135 個のリターンが観測されるようにし、リターンの絶対値の順に並び替え、シミュレーションと日経平均の誤差の総和を誤差関数として採用する。また、式のランダム生成における最大の深さを 4 とする。

3.2.2 結果

誤差は第 1 世代では 6.41 だったのが、第 12 世代では 0.769 に低下した。FT に関しては VC よりも GP によるモデルの改善が良好であり、図 2 に示す通りリターンの絶対値の累積分布は現実の市場と極めて近い形状をしている。生成された式は((0.6/-4.5)-(PrevR(2)+PrevR(5)))*(PrevR(5)+(PrevR(5)+0.1))、整理すると $2r_{-5}^2 + 2r_{-5}r_{-2} + 0.367r_{-5} + 0.1r_{-2} + 0.0133$ となり、VC の再現と比べて FT の再現は遥かに単純な式で可能であるということが分かる。第 1 項は ARCH に近い形であるが、このモデルにおいてボラティリティの自己相関は弱かった。

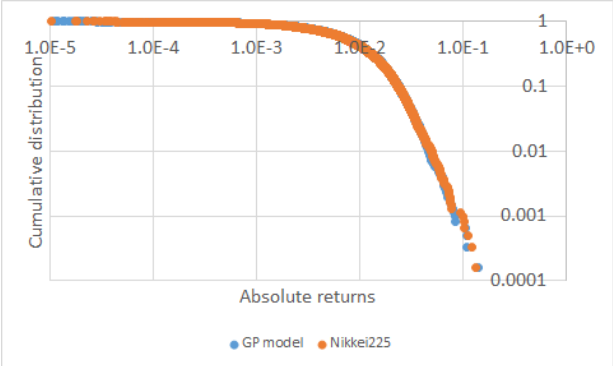

図 2　FT のみを目的とした標準的な GP が生成したモデルにおけるリターンの絶対値の累積分布

3.3 GP3：FTVC 両目的標準 GP

3.3.1 条件設定

誤差関数を式(3)のように、累積分布の誤差と自己相関の誤差の線形和で表現する。

$$E_{total} = E_{CDF} + cE_{ACF} \qquad (3)$$

ここで、係数$c = 8.9$とする。ランダム式生成における最大深さは VC 再現の為、5 とする。

3.2.2 結果

GP の 15 世代目のモデルのシミュレーション結果を図 3、4 に示す。

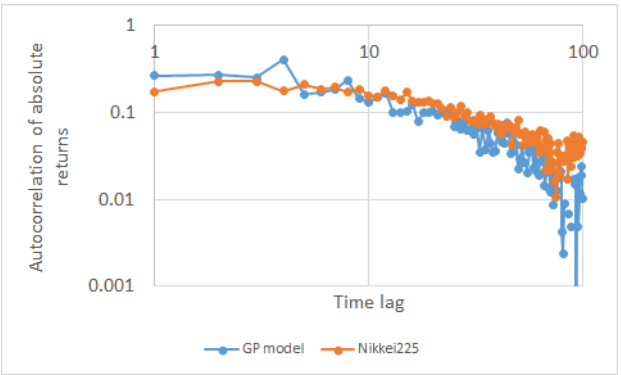

図 3　VC と FT を目的関数とした標準的 GP が生成したモデル及び日経平均におけるボラティリティの自己相関

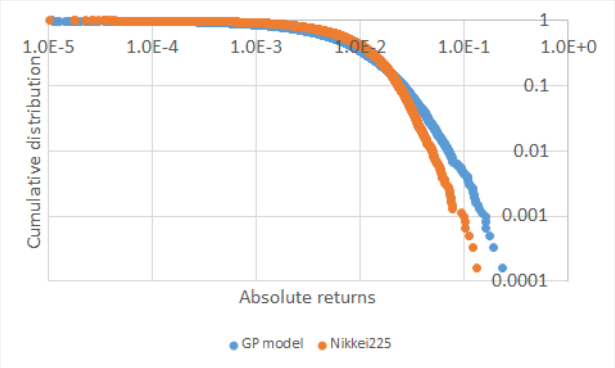

図 3　VC と FT を目的関数とした標準的 GP が生成したモデル及び日経平均におけるリターンの絶対値の累積分布

誤差の重み付け係数にも依存するが、VC か FT の一方の再現は、その片方のみを目的とした GP より劣り、両方を上手く再現する為には c の値を細かく調整する必要がある。今回進化計算で生成された式は
(((((PrevR(3)+PrevR(4))+(PrevR(4)+PrevR(2)))*4.1)/(-3.6+(PrevR(3)--0.2)))-(((PrevR(2)+PrevR(1))--0.2)*(0.8*(PrevR(3)+(0.8*-0.3)))))/(((PrevR(3)+PrevR(4))+(((((-0.8+PrevR(1))+1.6)-PrevR(3))*(PrevR(5)/(-4.5/(0.8*((PrevR(4)+(PrevR(3)*PrevR(1)))+PrevR(2)))))+PrevR(2)))-((((((PrevR(3)+(PrevR(2)+PrevR(1)))+(PrevR(4)+PrevR(2)))*PrevR(1))+(-3.8-(PrevR(3)*(PrevR(3)--0.2))))*PrevR(1))+(-3.8-PrevR(5))))であった。

3.1 及び 3.2 で示した通り VC の再現は複雑な式を用いないと困難で FT は比較的単純な式で可能であるが、この結果より FT は複雑な式においても再現が阻害されることはないということが分かった。

3.4 GP4：定数項分離型 GP

通常の GP は人間から見て明らかに解がないと分かるような解候補も探索している為に十分な結果を出しにくいと考え、本問題の構造に合わせ以下のように GP の条件設定を変更した。

・1 つのモデルの記述に 1 つの数式ではなく、1 つの係数と 1 つの数式を用いて表現する
・GP により生成する数式の終端子は定数と過去のリターンを別個に扱うのではなく、定数と過去のリターンの積を 1 つの終端子として扱う
・数式とは別個にモデルが持つ定数は交叉が行われる際に誤差に基づく重み付き平均により計算された値が新しい個体に与えられる。突然変異及び引継ぎにおいては親と同じ値を与える

このモデル改変の意図としては、この問題において GP で真に探索したいのは過去のリターン同士の複雑な関係であり、定数しか返さない定数同士の演算ではないので、定数倍の過去のリターン同士でのみ演算が起きるようにすることである。定数項に関しては何らかの最良の数値が存在すると仮定し、一旦最適解を得た後は変化しないように誤差による重み付き平均で更新することとした。

しかしながら、改変した GP は標準的な GP に勝る結果を出すことは出来なかった。ハイパーパラメータを式のランダム生成における最大深さを 4、終端子選択率を 20%、新世代生成において交叉 90%、突然変異 1%とした場合の標準的な GP よりかは高い性能を示したが、3.1 から 3.3 に示した結果とは明らかに劣る。

4 GP の経験的考察

GP は様々な問題点の発見と改良が行われているが、GP はハイパーパラメータに依存して出力する解が大きく変動する為、GP 改良のテクニックの導入の必要性は GP で扱う問題毎に検討されるべきである。

例えば、GP においては bloat（世代数が増えるにつれて数式のサイズが拡大し構造が複雑になっていくにも関わらず適合度の上昇が伴わない）が発生するということが知られており、その対策として Parsimony 係数(式のサイズに対しペナルティを科す)や Size fair 交叉や Shrink 突然変異等が提案されている。ところが、3.1.2 で示した通り、ランダム式生成における最大深さが 1 異なるだけで進化計算の結果得られる式のサイズが大きく異なるので、bloat の対策は特殊な技法を導入する前に GP のハイパーパラメータ調整で十分でないかを確かめる必要がある。

また、式のランダム生成における最大深さを 4、終端子選択率を 20%とした場合の結果の悪さについては、初期の解候補の多様性が少ないのが原因であると考えられる。最大深さが小さく終端子選択率が低い(ランダムに生成した式が最大深さまで伸びやすい)程、最大深さまで多くの枝を伸ばした式ばかりが初期解として生成されてしまう。交叉や突然変異は現在の解候補の部分から新たな解候補を作る過程であるから、解候補は初期解から進化計算によりすぐに変化していくにも関わらず、その離れ方の不均一性が不足し、探索が上手くいかなかったのであろう。

5 GP の人工市場への応用の考察

3.2.2 において GP が 2 ステップ前と 5 ステップ前のリターンのみを利用して FT を再現する確率過程モデルを構築したように、GP が作り出すモデルにおいて使われる変数の選択は人間のそれとは大きく異なっている。人間であれば、1 ステップ前のリターンを差し置いて 2、5 ステップ前のリターンを使ってモデルを作ることはないであろう。

このことは GP が人間ならば見落とすであろう市場の性質を捉える可能性だけでなく、表面的には正しそうに見えても人間から見れば明らかに間違っているモデルを生成する可能性も示唆している。例えば、エージェントが板情報を用いて取引を行う人工市場を構築した時、最良気配値と第二気配値の挙動が仮に似ていたとして、裁量気配値を用いずに第二気配値を用いて意思決定を行うモデルが構築されたとしたら、それは望ましいものとは言えない。これには 3 つの解決法が考えられる。

1 つ目として、入力変数の独立性が高い条件下で GP を使うことである。例えば温度、圧力、経過時間から物質の物性値を予測するような問題であれば、温度の代わりに圧力が使われてしまうといった問題は起きないであろう。だが、このような制限された設定で構築出来る金融市場のモデルは数少ないが、一応存在し得ると考えている。例えば数学的に 1 人のトレーダーの最適行動を記述したモデルで、何らかの分布等、少数個の関数形さえ仮定すれば多数のエージェントで取引を行わせコンピューターシミュレーションを行えるようなものがあった時、その少数個の関数形を GP で生成するというのは可能であると予想する。

2 つ目として、再現する目標となる統計性質の数を増やすということが挙げられる。これはモデルが満たすべき条件を厳しくすれば小さな痂疲も見逃されなくなるだろうという考えに基づくが、多数の目標を立てると同時にそれぞれを満たすのが難しくなり、そのようなモデル自体構築出来ない可能性がある。この方法のみではあまり現実的ではない。

3 つ目として、突然変異を subtree 突然変異のみにせず、式の構造を変えず終端子のみを僅かに変えるタイプの突然変異を導入することである。例えば、過去 2 ステップ前のリターンを参照する部分を、過去のリターンを参照するという部分は変えずに、それを 3 ステップ前に変える等がこれに該当する。これによって人間から見てより適切なモデルが調査から漏れるということをある程度防ぐことが出来ると期待される。もしこのような操作を経ても不自然なモデルの方が優れていた場合には、不自然なモデルこそより優れていたと主張出来、理論の発展に貢献する可能性まであるかもしれない。

また、人工市場モデルの構築においてエージェントの行動規則を GP ではなく他の技術により生成するということも考えられる。確率過程モデルの場合、容易に数式が読めるのは FT 再現のみを目的とした場合だけであり、VC 再現が入ると数式の整理、読解に手間がかかる。そこで、例えば fast function extraction (FFX) [8]のように、関数及びそれらの一度の合成の線形和で数式を作るということも考えられる。この場合、一定以上に複雑な式が出てこず、理論への拡張性も高いと期待される。FFX は不要な項は係数推定と同時に自動的に削除されるアルゴリズムになっておりその点でも極めて利便性が高いが、FFX はモデルの生成に直接そのままの形では使えないので、改変が必要となる。

GP、改変型 FFX の何れを使うにせよ、四則演算のみならず指数関数や対数関数等多数の関数をモデル生成に使うようになれば、シミュレーションという入力される数値の範囲が不確定な中で数値の発散等を如何に防ぐかというのも問題となってくる。

参考文献

[1] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987-1007.

[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307-327.

[3] Challet, D., Marsili, M., & Zhang, Y. C. (2000). Modeling market mechanism with minority game. *Physica A: Statistical Mechanics and its Applications*,*276*(1), 284-315.

[4] Iori, G. (1999). Avalanche dynamics and trading friction effects on stock market returns. *International Journal of Modern Physics C*, *10*(06), 1149-1162.

[5] Maslov, S. (2000). Simple model of a limit order-driven market. *Physica A: Statistical Mechanics and its Applications*, *278*(3), 571-578.

[6] Smith, E., Farmer, J. D., Gillemot, L. S., & Krishnamurthy, S. (2003). Statistical theory of the continuous double auction. *Quantitative finance*, *3*(6), 481-514.

[7] Poli, R., Langdon, W. B., McPhee, N. F., & Koza, J. R. (2008). *A field guide to genetic programming*. Lulu. com.

[8] McConaghy, T. (2011). FFX: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX* (pp. 235-260). Springer, New York, NY.