人工知能学会研究会資料
SIG-FIN-020

ボラティリティ・クラスタリングが観測される時系列のローソク足同時分布モデル

Candlestick Joint-Distribution Models for Time Series with Clustered Volatility

内木　正隆[1*]　DE BRECHT Matthew[1]　櫻川　貴司[1]
Masataka Naiki[1]　DE BRECHT Matthew[1]　Takashi Sakuragawa[1]

[1] 京都大学 人間・環境学研究科 数理科学講座
[1] Graduate School of Human and Environmental Studies, Kyoto University

Abstract: *Open-high-low-close price* (also OHLC) series have been widely used for the price-movement analysis of financial time series, including to draw *candlestick charts*. Modeling these data is complicated by the fact that such data are often unlikely to be samples of stationary stochastic processes, as can be seen in the well-known phenomenon of *volatility clustering*. In this research, first we try to remedy this matter by using the sequences of differences between high and low prices, which are pointed out to often have higher autocorrelations than the absolute returns of close-price series, and normalize the scales of OHLC by their exponential moving averages. Under our experimental conditions, the *Earth Mover's Distance* (EMD) between normalized S&P500 training and test data is about one-seventh of the EMD between the unnormalized data. Second, we try to model the normalized data by introducing 6 generative models for them. The EMDs between data generated by our learned models and the normalized test data are about one-sixth of the EMD between the normalized test data and the delta distribution located at the barycenter of the normalized training data. However, they are about 5 times larger than the EMD between the normalized test and training data.

1　はじめに

本論文では、金融時系列データ等でよく用いられる、一定期間ごとのいわゆる四本値の一部の情報について、確率過程としての定常性を強める変換の方法と、変換後のデータを学習する生成モデルを提案し、評価を行う。

金融時系列データは多くの回数の売買によって動いた価格変動の履歴全てから構成される。しかしこれらの値動き全体のデータ量は非常に膨大なため、実際には一定期間ごとの最初の値 (始値)、最後の値 (終値)、その期間内での最大値 (高値)、最小値 (安値) の**四本値**、あるいは取引された量 (出来高) を加えた 5 つの値のみがその期間の値動きデータとして利用できる場合も多い。これらは終値のみに比べて与える情報がより大きく、これらを用いると終値のみを用いた場合に比べて良い精度で予測可能になる場合もある [2]。しかし金融関連の時系列分析を行う場合、四本値全てを同時に扱うモデルや研究はあまり多くなく、終値のみのスカラー時系列の予測を行う場合がかなりある。本論文では逆に終値の時系列の情報の部分を除いた他の情報を扱うことを考える。高値や安値の値動きを調べるための方法の一つとして、その期間内の始値に対する終値・高値・安値の相対値を扱うことができる。これら相対値は四本値を同時に時系列として可視化する手法の一つであるローソク足チャートの個々のローソクの形を表す 3 次元データである。本論文ではこれら相対値データのモデル化を目指す。関連研究としては相対的な終値の変化率の予測を試みた例は [10][11] など多くあるものの、高値や安値の動きを終値も含めた同時分布として予測・評価する例は今回見つけられなかった。

モデル化の際に出てきた問題点とそれらの解決を目指す提案を以下に挙げる。まず扱うデータが定常過程のサンプルとは思えず、時期により標準偏差が異なるため定常過程でのモデル化が難しかった。そのため指数移動平均を用いて時系列データの変換をするという方法を提案して非定常性を抑えた。また分布のモデル化の際に、実データがルベーグ絶対連続な同時分布では表現できない分布のサンプルに見えるという問題が

*連絡先：京都市左京区吉田二本松町
E-mail: naiki.masataka.44z@kyoto-u.jp
sakura@i.h.kyoto-u.ac.jp

あった。つまり密度関数で表される分布ではうまく表現できないのである。これについては実際に最大値や最小値を求めるような生成モデルを 3 種類提案し、それでも表現できない部分をそれぞれ別に扱う (後述する間引き処理) ことで合計 6 種類のモデルを提案することで対応した。実際のデータ集合と、それを推定した生成モデルからの推定値集合の類似度を表す指標としては Earth Mover's Distance (EMD) を用いた。

評価時の対象データとしては、アメリカ株の代表的なインデックス指数の一つである S&P500[1]、期間は 2001 年から 2016 年までの期間を対象とした。2013 年までを訓練データ Train、2014 年以降をテストデータ Test として分割して、指数移動平均のパラメータの学習は標準偏差のばらつきを最小化することで行った。結果は指数移動平均によるある種の正規化を行った後の訓練データとテストデータの EMD が 0.0571 で、変換前の訓練データとテストデータの EMD が 0.4236 だった。

変換後の訓練データの 6 種の生成モデルでの学習は、訓練データとモデルが生成するサンプルの EMD を最小にするパラメータを求めることで行った。変換後のテストデータと学習後の生成モデルが生成するサンプルの EMD は 0.3 程度、変換後の訓練データと生成サンプルの EMD は 0.18-0.23 程度であった。

論文の構成は以下の通りである。2 節で本論文で考察対象とする事象について述べ、関連用語を説明する。3 節で時系列の定常性の定義、ボラティリティ・クラスタリングが起きている場合に用いる既存のデータ処理方法と、データの定常性を増加させる為に本論文で導入した手法を述べる。これはある種の正規化を行なっていることになる。4 節では分布間の類似性を表す指標として今回採用した Earth Mover's Distance について説明する。5 節では正規化後の分布を学習する為導入した 6 種の生成モデルと、それらのパラメータの学習 (最適化) 方法、S&P500 のデータへの適用結果について述べ、6 節では考察を行う。

2 時系列とローソク足

時系列の期間毎の四本値を図 1,2 のように表示したものは**ローソク足**と呼ばれる。これは高値と安値の幅を棒の長さ、終値と始値の幅を箱の高さ、終値と始値の大小を箱の色で示したもので、こういった情報を同時に表現することができる。

以降全取引期間中の (通常一定幅の) 各期間を $I_t, t = [0, 1, .., T]$ で指し、I_t における始値:$(Open)_t$・高値:$(High)_t$・安値:$(Low)_t$・終値: $(Close)_t$ について考える。例えば 2017 年 3 月の市場が開場している日の日足 (各期

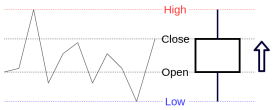

図 1: 始値 < 終値の場合

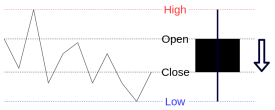

図 2: 始値 > 終値の場合

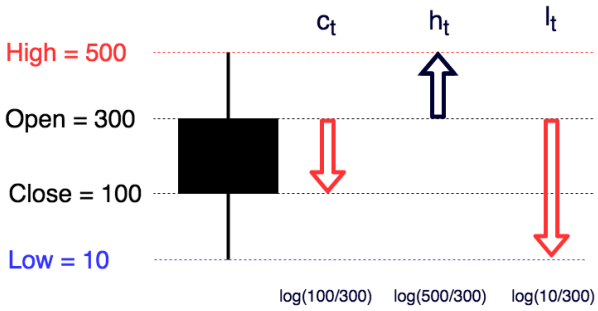

図 3: ローソク足と c_t, h_t, l_t の関係性

間を各営業日の取引時間としたもの) のインデックスは、各日付 (3/1,3/2,3/3,3/6,...) を各インターバル区間 $(I_0, I_1, I_2, I_3, ...)$ に対応させる。

今回は終値の相対値だけではなく、始値と比較した高値・安値の相対的な値についても関心があるため、それらの同時分布を扱い、(1) 式で c_t, h_t, l_t を定義する。図 3 でもわかるように、これらから成る 3 つ組はローソク足の形に対応している。これらの値は常に $l_t \leq c_t \leq h_t, l_t \leq 0 \leq h_t$ を満たす。

$$\begin{aligned} c_t &\stackrel{\text{def}}{=} \log((Close)_t) - \log((Open)_t) \\ h_t &\stackrel{\text{def}}{=} \log((High)_t) - \log((Open)_t) \quad \geq 0 \\ l_t &\stackrel{\text{def}}{=} \log((Low)_t) - \log((Open)_t) \quad \leq 0 \end{aligned} \tag{1}$$

金融時系列の時間あたり変化率の分布は、裾の重い分布とよばれる性質を持つことが多いことが知られている [4]。ここで裾の重い分布とは、確率分布の裾がガウス分布のように指数減衰的ではなく、それよりも緩やかな分布のこととする。

以上の 3 次元データを発生させる生成モデルを考える場合、例えば $h_t = 0$ や $c_t = h_t$ となる割合が実データで 0 でない場合がかなりあることが問題となる。これは密度関数等で表されるような**ルベーグ絶対連続**な確率分布ではうまく表現できないデータであることを意味しており、それを考慮に入れた生成モデルを考える必要がある。

[1] YahooFinance (https://finance.yahoo.com/quote/%5EGSPC/) のデータを用いた。

3 変動率の時系列モデル

確率過程の性質として**定常性**がある。これは基本的な統計量が時刻に関わらず一定である性質のことである。

定義 3.1 確率過程 $\{Y_t\}_{t\in\mathbb{Z}}$ が (弱) 定常性をもつとは任意の $t, h \in \mathbb{Z}$ において期待値・共分散が定義できて、かつ以下が成立することをいう[2]。このとき $\{Y_t\}$ は定常過程とよばれる。

$$\mathbb{E}[Y_t] = \mu(\text{定数})$$
$$\mathrm{Cov}[Y_t, Y_{t-h}] = \gamma_h(h \text{ の関数})$$

モデルを作る対象の時系列がこうした定常性をもった過程から生成されていると仮定できればモデル化が容易なことが多い。しかし一般に金融時系列では一定期間ごとの変化率の絶対値の分布が必ずしも時刻に対して一定ではなく、しかも連続して似た値になりやすい傾向が知られている。こういった現象はボラティリティ・クラスタリング、あるいは分散不均一性と呼ばれる [7]。図 4 の左側を見ると、今回用いた S&P500 のデータでも経済が順調だった 2006 年の $\mathrm{std}[c_t]$ [3]と比べて、リーマン・ショックが起きた 2008 年の $\mathrm{std}[c_t]$ の方が約 4 倍である。この場合に 2006 年と 2008 年を同じ過程から発生した値と仮定してモデル化するのは適当ではないと思われる。

こうしたデータの変化率の絶対値のばらつきを揃えるため、本論文では $\{u_t \stackrel{\text{def}}{=} \log((High)_t/(Low)_t)\}$ のパラメータ α の指数移動平均 $\bar{u}_t$ を用いることを提案する。実際の変換は $c_{t+1}, l_{t+1}, h_{t+1}$ を、パラメータ $0 < \alpha < 1$ による指数移動平均 $\bar{u}_t$ で割ることで行う。$\bar{u}_t$ は (2) で定義される。

$$\bar{u}_t = \begin{cases} \alpha u_t & (t = 0) \\ \alpha u_t + (1-\alpha)\bar{u}_{t-1} & \text{otherwise.} \end{cases} \quad (2)$$

元の変化率データである c_t, h_t, l_t を $\bar{u}_t$ を用いて $c'_t \stackrel{\text{def}}{=} c_t/\bar{u}_{t-1}, h'_t \stackrel{\text{def}}{=} h_t/\bar{u}_{t-1}, l'_t \stackrel{\text{def}}{=} l_t/\bar{u}_{t-1}$ と変換する。指数移動平均のパラメータ α は訓練データの期間ごとのばらつきをできるだけ一定とするような、つまり短い期間毎の標準偏差集合の標準偏差最小化によって求める。訓練データ学習期間中の i 番目の月 (2001 年 1 月,2001 年 2 月,...,2013 年 12 月) を m_i として、(3) 式で与えられるデータのばらつき具合を最小化する α を選択する。

$$\sum_{r\in\{c',h',l'\}} \mathrm{std}[\{\mathrm{std}[\{r_t\}_{t\in m_i}]\}_i] \quad (3)$$

S&P500 からの訓練データ Train で α の最適化を行った結果 $\alpha = 0.29072$ となった。またこのとき $\sigma_{c'} = 0.77795$ だった。このように時刻 u_t の指数移動平均を時刻 u_{t+1} のある種の予測値的な値として用いることは、金融時系列で変動の幅を推定するモデルとして一般的な **GARCH モデル** [2] と類似している。ただし GARCH モデルでは予測値と実際の値との誤差などによって係数を最適化するのに対し、提案手法では (3) の値を最小化しているところに新規性があると考えられる。

図 4 の右側は指数移動平均による変換後の c'_t, h'_t, l'_t のグラフで、左側の変換前に比べ各年ごとの標準偏差の大きさのばらつきが抑えられているのが判る。さらに訓練期間中の t に関して $c''_t \stackrel{\text{def}}{=} c'_t/\sigma_{c'}, h''_t \stackrel{\text{def}}{=} h'_t/\sigma_{c'}, l''_t \stackrel{\text{def}}{=} l'_t/\sigma_{c'}$ とする。ただし $\sigma_{c'} \stackrel{\text{def}}{=} \mathrm{std}[c'_t]$。

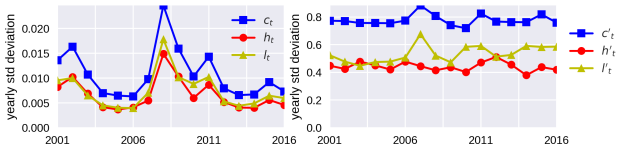

図 4: 2001 年から 2016 年まで各年内での S&P500 についての標準偏差。左は $\mathrm{std}[\{c_t\}_{t\in m_i}], \mathrm{std}[\{h_t\}_{t\in m_i}], \mathrm{std}[\{l_t\}_{t\in m_i}]$、右は指数移動平均による変換を行った $\mathrm{std}[\{c'_t\}_{t\in m_i}], \mathrm{std}[\{h'_t\}_{t\in m_i}], \mathrm{std}[\{l'_t\}_{t\in m_i}]$, のグラフ。

4 Earth Mover's Distance

本論文では正規化の結果の評価と生成モデルの学習のために、**Earth Mover's Distance** (EMD) を用いて分布間の類似性を測る手法を提案する。学習の場合には実際の同時分布と、それを推定した生成モデルからの出力の分布間の類似度を最大化することになる。EMD は一般に 2 つの分布の類似度を表す指標で、値が少ないほど類似性が高く、値の計算の基礎となる後述の $d(-,-)$ について $d(x,x) = 0$ であれば同じ分布同士の EMD は 0 である。EMD を直感的に説明するなら、分布を砂山と考えて片方の砂山からもう片方の砂山へ砂を移動させるための最小コストと説明できる。

計算機関連ではもともと色・テクスチャなどを考慮して定義される画像間の距離を与えるために使われ、画像修復や画風変換に用いられる [5]。定義の形式こそ異なるものの、EMD に相当する距離が Wasserstein Distance や最適輸送距離という名前でも知られており [9]、近年では深層学習における生成モデルで最適化の目的関数としても注目されるようになっている [3]。

以下の有限の場合の説明は [8] を参考にしている。このとき EMD は以下の LP 問題の最小値によって与えられる。各点に重み付けを持つ分布 $P \stackrel{\text{def}}{=} \{(p_i, w_{p_i})\}_{i=1}^m$,

[2] $\mathbb{E}$ は期待値、Cov は共分散をとる記号

[3] $\mathrm{std}[\{r_t\}_{t\in X}] \stackrel{\text{def}}{=} \sqrt{\frac{1}{|X|}\sum_{t\in X}(r_t - \bar{r})^2}$.

$Q \stackrel{\text{def}}{=} \{(q_j, w_{q_j})\}_{j=1}^{n}$ があるとする。距離行列 $D \in \mathbb{R}^{m\times n}$ を、p_i と q_j の間の輸送コストである $d_{ij} = d(p_i, q_j) \geq 0$ を要素に持つ行列として定義する。ただし $d(-,-)$ はユークリッド距離などベクトル同士の何らかの性質の差を表す指標である。d が距離の公理を満たせば EMD も距離の公理を満たす。$F \in \mathbb{R}^{m\times n}$ は輸送フローを表し、p_i から q_j へのフローをあらわす $f_{ij} \geq 0$ を要素に持つ行列である。最適化の目的関数はコストとフローの対応する要素同士の積の和で、これが分布間の移動の仕事量をあらわす。また制約条件によって、フローは 0 以上で p_i, q_j へのフロー流入量の最大値は決まっており、全フローの合計フローを一定に定めている。

$$\begin{aligned}&\underset{F}{\text{minimize}} \sum_i \sum_j d_{ij} f_{ij} \\ &\text{subject to } f_{ij} \geq 0, \sum_j f_{ij} \leq w_{p_i}, \sum_i f_{ij} \leq w_{q_j}, \\ &\qquad \sum_i \sum_j f_{ij} = \min\left(\sum_i w_{p_i}, \sum_j w_{q_j}\right). \quad (4)\end{aligned}$$

これによって求められた最小値をそのときのパラメータ $\hat{F}$ で正規化したのが P, Q の間の EMD である (5)。以上の定義は P, Q のそれぞれの重みの合計値が 1 であることを仮定していない。それぞれが分布、つまり全ての i, j において $w_{p_i} = 1/m, w_{q_j} = 1/n$ である場合も多く、こういった場合には正規化しなくても分子がそのまま $\text{EMD}(P, Q)$ となる。本論文ではこの意味で EMD を用いる。

$$\text{EMD}(P, Q) \stackrel{\text{def}}{=} \frac{\sum_i \sum_j d_{ij} \hat{f}_{ij}}{\sum_i \sum_j \hat{f}_{ij}} \quad (5)$$

(4) の最適化問題は線形計画問題の中の最適輸送問題のソルバーを用いて解かれる。これの計算コストが $\mathcal{O}(n^3 \log n)$ あるため大きな要素数に対しては計算コストが高くなる [1]。また (5) の値は一意に定まるが $\hat{F}$ は一意とは限らない。こういった問題の解決目的で様々な EMD の変種が存在するがここでは割愛する。

5 生成モデルの構成と結果

本論文で導入した計 6 つの生成モデルの具体的構成方法・評価方法・実験結果についてまとめる。

5.1 モデルの構成

本論文では実際に価格の変動率を左右する何らかの定常な分布が存在すると仮定して、全ての各インターバル区間 I_t の間に一定回数 (K 回) の価格の変動が、ある分布に従って独立に発生するモデルを考える。また、ルベーグ絶対連続でない分布をうまく表す方法として、後述の「間引き」処理を試みる。各生成モデルは分布モデル、同時分布計算モデル、(もしあれば) 間引き処理、スケールの調整からこの順で構成される。間引きを行う方法を (b)、行わない方法を (a) とする。今回は分布の形について (i, ii, iii) の 3 通り、間引きを行うかどうかで (a,b) の 2 通り、これらを組み合わせて (i)-(a),(i)-(b), (ii)-(a), (ii)-(b), (iii)-(a), (iii)-(b) の計 6 通りのモデルを実験対象とする。

分布モデル　変化率の元となる分布モデル p_θ の候補は多い。本論文では p_θ をガウス分布とする方法 (i)、ジョンソン S_U 分布システム [6] とする方法 (ii)、分散が或るパレート分布に従うガウス分布とする方法 (iii)、の 3 通りを試みた。

(i) は (ii), (iii) など p_θ が裾の重い分布 (確率分布の裾がガウス分布のように指数減衰的ではなく、それよりも緩やかな分布 [4]) を表現可能な場合と比較するために採用する。スケールに関しては学習パラメータに含めないため $\theta = (\mu \in \mathbb{R})$ となる。$r_{t,k} = q_{t,k} + \mu, q_{t,k} \sim N(0,1)$ として $r_{t,k}$ を計算する。独立なガウス分布からのサンプルの和はガウス分布に従うことが知られており、p_θ が $N(\mu, \sigma^2)$ であるとき $\hat{c}_t = \sum_{k=1}^{K} r_{t,k} \sim N(K\mu, K\sigma^2)$ となる。

(ii) は $\theta = (\mu' \in \mathbb{R}, \sigma' \in \mathbb{R}_{>0}, \mu \in \mathbb{R})$ で $r_{t,k} = \mu + \sinh(\sigma' q_{t,k} + \mu'), q_{t,k} \sim N(0,1)$ とする。σ' によって裾の重さを調節可能で、μ, μ' の組み合わせによって左右非対称な分布も表現可能となっている。

(iii) は $\theta = (\mu \in \mathbb{R}, c \in \mathbb{R}> 0, d \in \mathbb{R}_{>0})$ で $r_{t,k} = \mu + \sigma_{t,k} q_{t,k}, q_{t,k} \sim N(0,1), \sigma_{t,k} \sim \text{pareto}(c, d)$ となる。c は $\sigma_{t,i}$ の下限を意味するパラメータ、d は裾の重さを調節するパラメータとなっている。

同時分布計算モデル　$k = 1, 2, .., K$ として各変動率 $r_{t,k} \in \mathbb{R}$ を分布モデルから独立同分布にサンプリングする。$r_{t,k}$ から長さ $K+1$ の時系列 $\{s_{t,k} \stackrel{\text{def}}{=} \sum_{k'=1}^{k} r_{t,k'}\}_{k=0}^{K}$ を計算する。$s_{t,0} = 0$ である。今回のモデルでは $s_{t,k}$ はインターバル区間 t の k 回目の変動後の価格と見なす。よってインターバル区間 t の始値を 0 としたときの最大値・最小値・終値の値を計算することで、サンプル $\hat{c}_t, \hat{h}_t, \hat{l}_t$ の同時分布を得ることができる (図 5)。つまり $\hat{c}_t \stackrel{\text{def}}{=} s_{t,K}, \hat{h}_t \stackrel{\text{def}}{=} \max_k s_{t,k}, \hat{l}_t \stackrel{\text{def}}{=} \min_k s_{t,k}$ とする。常に $\hat{l}_t \leq 0 \leq \hat{h}_t, \hat{l}_t \leq \hat{c}_t \leq \hat{h}_t$ である。

このようにモデル化することで、直感的には変動の積み重ねから結果が生じている実際の取引データの発生の仕組みをある程度近似しているのみならず、有限回の累積で高値・安値を発生させているため、ルベー

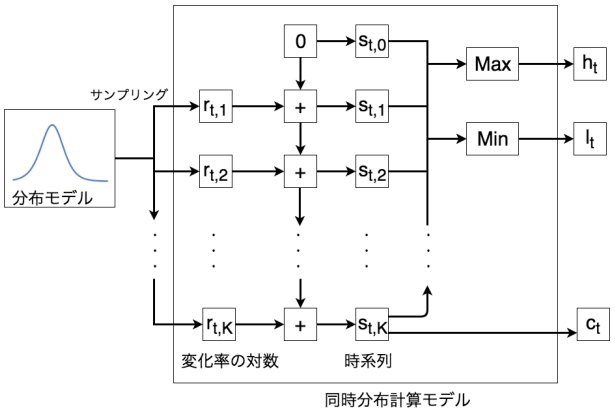

図 5: 同時分布計算モデル

グ絶対連続でない 3 次元分布を表すモデルとなっている。ただし少なくとも後述のように分布の比率を合わせるための処理が必要と考えられる。

間引き処理　今回用いるデータでは分布中のルベーグ絶対連続ではないと思われる部分の比率が同時分布生成モデルのそれらの比率と一致しない場合が多い。例えば今回用いた S&P500 のデータの場合、高値と始値が等しいデータが 10 %以上もあり、それを表せる間引きなしのモデルは限られてしまう。訓練データと、モデルが生成するサンプルの高値＝終値、高値＝始値、安値＝終値、安値＝始値の 4 つの場合の組み合わせの比率を一致させるように生成データを独立な乱数により一定の割合で間引く。各分布モデルについて、間引きする場合・しない場合の両方の生成モデルを考える。

5.2　モデルのパラメータ最適化手法と結果

生成モデルのパラメータ最適化の際には 4 節で説明した EMD を用いる。各要素間の指標 $d(-,-)$ としてはユークリッド距離の 2 乗を用いた。今回は毎回生成する (間引き後の) サンプル数も訓練データの個数と同じとした。前節であげた 6 つのモデルから成る集合を M、各モデルを $m \in M$、モデル m のパラメータを $\theta^{(m)}$ と表記する。以降各インターバル内の変動回数が K 回で、パラメータが $\theta^{(m)}$ の、モデル m から発生させた N 個の値を、$\mathrm{Predicted}(\langle m, \theta^{(m)}, K\rangle, N)$ と表記する。

$\theta^{(m)}, K$ の探索についてはある種のランダムサーチを用いた。ある分布に従って実際に $\theta^{(m)}, K$ をサンプリングした中で、最も $\mathrm{EMD}(\mathrm{Train}'', \mathrm{Predicted}(\langle m, \theta^{(m)}, K\rangle, N))$ の小さかった $\theta^{(m)}, K$ を選択する。ここで Train'' は指数移動平均による変換後に標準偏差を 1 に正規化したデータである。なお、指数移動平均の係数 α と標準偏差正規化の係数 $\sigma_{c'}$ は Test データに対しても同じ値:$\alpha = 0.29072, \sigma_{c'} = 0.77795$ を用いた[4]。結果のデータを Test'' と表記する。各モデル m ごとに K は 3 から 200 の整数の中から 100 個サンプリング、各ペア $\langle m, K\rangle$ ごとに $\theta^{(m)}$ のサンプリングを行う。$\theta^{(m)}$ のサンプリングの設定は表 1 の通り。

モデル	個数	パラメータ	下限	上限
(i)	30	μ	$-0.02/K$	$0.02/K$
(ii)	200	μ	$-0.02/K$	$0.02/K$
		μ'	$-0.02/K$	$0.02/K$
		σ'	0	3.0
(iii)	200	μ	$-0.02/K$	$0.02/K$
		c	0	0.01
		d	0	5.0

表 1: 各モデルにおけるランダムサンプリングの対象とする分布とサンプリング個数

6 通りの各モデルの生成値に対する $\mathrm{Train}'', \mathrm{Test}'', \mathrm{Mean}''$ との EMD の値を表 2 に記載した。ここで Mean'' は Train'' の重心に位置する δ 分布である。

EMD	Train''	Test''	Mean''
(i)-(a)	0.2317	0.2966	1.768
(ii)-(a)	0.2324	0.2938	1.781
(iii)-(a)	0.2314	0.2953	1.780
(i)-(b)	0.2318	0.2984	1.768
(ii)-(b)	0.1769	0.2957	1.774
(iii)-(b)	0.1797	0.2935	1.770
Train''		0.05712	1.803

表 2: 生成サンプル,$\mathrm{Train}'', \mathrm{Test}'', \mathrm{Mean}''$ 同士の EMD

6　まとめ・考察

本論文では従来それほど研究されていないと考えられる、時系列データの各インターバルでの高値・安値・終値の始値に対する相対値の生成モデルを研究対象とした。まず、このような場合に分布の近似度を測る方法として EMD を用いることを提案した。また、定常性を強める変換と、生成モデルを提案し、評価実験を行った。

具体的には、定常性が必ずしもなく、ボラティリティ・クラスタリングが生じるような確率過程が発生する時系列について、高値と安値の差の指数移動平均の値によりデータを変換することで定常性を増やすことを意図した方法を提案し、S&P500 のデータを使用した評価を行った。標準偏差を正規化した Train と Test の EMD

[4] それゆえ Test'' の標準偏差は必ずしも 1 にならない。

に比べて、変換後に標準偏差を正規化した Train″ と Test″ の EMD は約 1/7 であった。

また、変換後の Train″ を学習する生成モデルとして、累積和を取ってから最大値・最小値を求め、間引きを行うことでルベーグ絶対連続でない同時分布を表せるモデルを 6 つ提案し、やはり S&P500 のデータにより評価実験を行った。しかしながらこれらの異なるモデルを考えたものの、各生成モデルが発生したサンプルと Test″ の EMD は全て 0.3 程度であった。一方、Train″ の重心に位置する δ 分布 Mean″ と Test″ の EMD が約 1.8 であった。ただし 0.3 という値は Train″, Test″ の EMD の値約 0.06 に比べると大きいため、今回採用した生成モデルの表現力は実験データにはそれほどマッチしていなかったと考えられる。また、裾の重い分布への対応を意図した (ii),(iii) の分布モデルや間引きを行うモデル (b) が優れていることを期待していたものの、今回の実験結果ではそういった傾向は認められなかった。(i)-(iii) にそれほど違いが見られなかった理由としては、元の分布モデルのこの程度の違いでは、累積和を取るとどれも正規分布に近くなってしまい、結果に影響を及ぼさなかった、ということが考えられる。

発展としては異なる性質を持つ為替・債券といった金融時系列を用いた実験や生成モデルの表現能力を上げることが考えられる。例えば同時分布計算モデルの出力に対してジョンソン S_U 分布システムのような変換を行うことで Test″ が裾の広い分布となっている場合にさらに適応することを目指すことができる。分布モデルが発生するサンプルは今回各回で独立としたが、独立でない場合をモデル化する方向性も考えられる。また EMD を用いた非定常性を抑えるための変換として、指数移動平均以外の方法も考えられるし、今回のように標準偏差の標準偏差を最小化するのではなく、訓練データの範囲で一定期間ごとに分割されたデータペアの EMD の平均を最小化することで、最適なパラメータ (今回の場合には α) を求める方法もある。売買アルゴリズムによっては生成サンプルに基づく売買シミュレーションを行うことが可能なので、それを行うことが考えられる。今回の研究は終値の時系列そのものを予測対象としたものでないが、それを行った研究があるので、それらと組み合わせて四本値の確率予測モデル的なものを作ることや、それらを用いた売買シミュレーションを行ってみることも考えられる。

謝辞

京都大学人間・環境学研究科数理科学教室の渡部崇氏との討論は大変有益でした。また、日頃お世話になっている同教室の先生方、学生の方々にも感謝します。

参考文献

[1] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Chapter iv network flows. *Handbooks in operations research and management science*, Vol. 1, pp. 211–369, 1989.

[2] Sassan Alizadeh, Michael W Brandt, and Francis X Diebold. Range-based estimation of stochastic volatility models. *The Journal of Finance*, Vol. 57, No. 3, pp. 1047–1091, 2002.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Søren Asmussen. *Applied probability and queues*, Vol. 51. Springer Science & Business Media, 2008.

[5] Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference on*, Vol. 30, p. 158, 2011.

[6] N. L. Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, Vol. 36, No. 1/2, pp. 149–176, 1949.

[7] Thomas Lux and Michele Marchesi. Volatility clustering in financial markets: a microsimulation of interacting agents. *International journal of theoretical and applied finance*, Vol. 3, No. 04, pp. 675–702, 2000.

[8] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, Vol. 40, No. 2, pp. 99–121, 2000.

[9] Cédric Villani. *Optimal transport: old and new*, Vol. 338. Springer Science & Business Media, 2008.

[10] Halbert White. Economic prediction using neural networks: The case of ibm daily stock returns. 1988.

[11] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, Vol. 50, pp. 159–175, 2003.