機械学習を用いた共和分ペア・トレード戦略

A machine-learning approach to pairs trading strategy

今村　光良 [1,3*]　中川　慧 [1,2]　吉田 健一 [2]
Mitusyoshi Imamura[1,3], Kei Nakagawa[1,2], Kenichi Yoshida[2]

[1] 日興グローバルラップ株式会社
[1] Nikko Global Wrap Ltd.
[2] 筑波大学 大学院 ビジネス科学研究科
[2] University of Tsukuba Graduate School of Business Sciences
[3] 筑波大学 大学院 システム情報工学研究科
[3] University of Tsukuba Graduate School of Business Sciences

Abstract: 機械学習を用いた株価予測については近年多く研究されている．しかしながら，その多くが株価ないし株価指数そのものをそのまま予測対象としている．一般に株価は非常に複雑な振る舞いを示し，予測が難しい．一方で，同業種同規模などの似通った銘柄間では価格差 (スプレッド) が平均回帰すること (所謂共和分性) が知られている．そこで本研究では，株価そのものではなく，共和分性を満たす株価ペアのスプレッドを機械学習で予測する手法を提案する．具体的には，LSTM を用いて，定常性を満足するモデルで最も代表的な AR(1) 過程に従う人工的な時系列データを事前学習する．学習した LSTM を用いてペア・トレード戦略に適用した結果，単純な AR(1) 過程や実データを直接学習させた LSTM よりも良好な結果が得られた．

1　はじめに

機械学習を用いた株価予測については近年多く研究されている．しかしながら，その多くが株価ないし株価指数そのものをそのまま予測対象としている．一般に株価は非常に複雑な振る舞いを示し，予測が難しい．一方で，同業種同規模などの似通った銘柄間では価格差 (スプレッド) が平均回帰することが知られている．計量経済学の文脈ではこれを共和分性 (co-integration) として様々な研究が行われている [2].

ペア・トレード戦略はこのような価格変動が似通った銘柄を見つけ，当該ペアの価格差が均衡水準の周りを推移すると仮定する．共和分性を満たすペアのスプレッドは定常過程となるため，平均が時点に依らず一定，すなわちある均衡水準への平均回帰性を持つ．そして，スプレッドが均衡水準から乖離したとき，将来その乖離が修正されるだろうという平均回帰に賭けて，相対的に割高な方を売り，割安な方を買うことで収益獲得を狙う戦略である．

共和分性を利用したペアトレードの実証研究として，[3] の研究がある．彼らは 1962 年から 2002 年までの米国株式市場において共和分関係にあるペアに注目し，ボリンジャーバンドを用いた実証分析を行った．具体的には，ペアのスプレッドが均衡水準から ±2 標準偏差以上乖離したときをシグナルとしてポジションを構築し，均衡水準に平均回帰したときにポジションを解消するという方法を用いた．このような共和分ペアに対してボリンジャーバンドを用いてポジションの構築を行う手法は，ペアトレード戦略の有効性を検証する上で実務及び実証分析のスタンダードとなっている．また，先行研究 [7] では，スプレッドの均衡水準からの乖離が小幅である場合と大幅である場合での回帰スピードの非線形性を指摘し，スプレッドが TAR モデルに基づく非線形共和分関係 [1] を満たすと仮定した．そして回帰スピードが変化する閾値をトレードのシグナルとして使用し，その有意性を確認した．上述の通り，共和分性を満たすスプレッドは定常過程となり，株価そのものよりも予測しやすいことが想定される．

そこで本研究でも，株価そのものではなく，共和分性を満たす株価ペアのスプレッドを機械学習で予測する手法を検討する．従来研究で一定とされていた共和分の性質を決めるパラメータが時間により変化する事を LSTM を使ってモデル化する事が提案手法の特徴である．

*連絡先： 日興グローバルラップ株式会社
〒 103-0016 東京都中央区日本橋小網町 9-2
E-mail: ic140tg528@gmail.com

2 提案手法

本研究では，共和分性を満たす株価ペアは定常性を満たすという特性を利用したペア・トレード戦略のためのスプレッド予測方法を提案する．まず共和分性を満たす株価ペアを見つけ，スプレッドを作成する．次に作成したスプレッドに適合すると推定される共和分の性質を決めるパラメータを使い AR(1) 過程に従う人工データを生成し，LSTM を用いて学習を行う．

Step 1:
共和分性を満たす株価ペアからスプレッドを作成し，AR(1) 過程を特徴づけるパラメータを推定する．

Step 2:
推定したパラメータを使い AR(1) 過程に従う人工データを生成し，LSTM で学習する．

Step 3:
学習した LSTM で実際のスプレッドの予測を行う．

実際の株価データから作成したスプレッドを直接学習するのではなく，一旦共和分の性質を決めるパラメータ (詳細は後述する) を推定し，推定したパラメータで生成した人工データを LSTM で学習させる事が提案手法の特徴である．

以下，提案手法の設計意図について述べる．

2.1 共和分性

共和分性は非定常な時系列データの線形結合が定常過程となる時系列的性質であり，[2] によって提唱され，長期的な均衡関係を記述するものとして経済，ファイナンスの様々な実証分析において利用されてきた．定常過程は，時系列の平均が時間に依らず一定であるため，平均回帰という扱いやすい性質を持っている[1]．

一方で通常，株価はランダムウォークであると言われている．ランダムウォークとは非定常な時系列の代表例であり，単位根過程とも言われる．一般にランダムウォークのような非定常な時系列を線形に組み合わせても，同様に非定常である．しかし，組み合わせ方をうまく選べば，定常過程となる場合がある．このとき共和分の関係があるという．具体的にはランダムウォークする 2 つの株価のペア $\{X_t, Y_t\}$ に対して，ある定数 β が存在し，以下の (見せかけの) 回帰式 $Y_t = \beta X_t + \epsilon_t$ における誤差項 $\{\epsilon_t\}$ が定常となることが共和分性の満たす条件である．

[1] 正確には過程の期待値と自己共分散が時間を通じて一定である確率過程を指す．

2.2 LSTM

LSTM は，データ間の依存関係を学習できるニューラル・ネットワークであるリカレント・ニューラル・ネットワーク (RNN) の一つである [4]．RNN は，AR 過程と同様に前の出力を次の入力に追加するモデルであり，時間方向に展開すると静的なニューラル・ネットワークと見ることができる．RNN は，時間方向に展開したネットワーク上で誤差逆伝播を用いて学習を行うが，系列が長くなると，勾配が消失してしまう．したがって，長期依存を学習できないという問題が生じる．これに対し，LSTM では，重みを掛けずに誤差を逆伝播させることによって，長期依存を学習できなくなる問題を解消している．

ここで，本研究では従来研究で一定とされていた共和分の性質を決めるパラメータが時間により変化する事を LSTM を使ってモデル化する．単純な AR 過程のモデル化では難しかったパラメータの時間変化を扱う事が LSTM を用いる事の意図である．

なお、本研究では，オープンソースの深層学習フレームワークである Chainer[6] のライブラリにて提供されている LSTM によりこれを実装した．また，ネットワークの構成は，同ライブラリにて提供されている LSTM のサンプルコード[2] と同様の構成とした．ただじ，入力層は全結合層に変更し，ユニット数を 1，LSTM 層についてはユニット数を 200，出力層はユニット数を 1 と変更している．学習時のパラメーター最適化には Adam [5] を用いた．

2.3 AR 過程の学習

定常性を満たすモデルで最も代表的な AR(1) 過程 $S_t = \alpha S_{t-1} + \epsilon_t$ ここで ϵ_t は正規分布 $N(0, \sigma)$ に従う誤差項, $S_t = Y_t - \beta X_t$ を LSTM に学習させる事を考える．

ここで，[7] でも指摘されている通り，株価ペアが共和分性を満たしても，スプレッドの定常性を表現するパラメータ α が時期によって変化することには注意を要する．従来研究では α を時期によらず一定として扱い，予測精度の低下をもたらしていたと考えられる．

そこで，実際のスプレッドのデータから α と σ を推定する．TOPIX 指数および日経平均株価のペアに対して，1985 年 1 月から 2006 年 12 月末までの 20 日間における AR(1) の α の推移は図 1 に示す通りである．

当該期間における α の中央値は 0.7552 であることから，α が変化し，$\{0.55, 0.65, 0.75, 0.85, 0.95\}$ の 5 つの値を取っていると推定した．またこの時の σ についても同一期間の実データからフィッティングし推定した．次に，1 バッチあたり各 α を持つ AR(1) に従う時系列

[2] https://github.com/chainer/chainer/blob/master/examples/

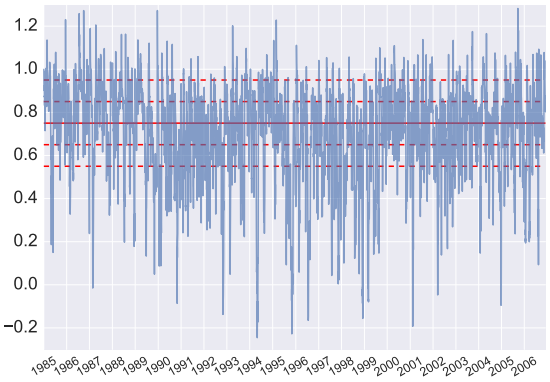

図 1: 1985 年 1 月から 2006 年 12 月末までの 20 日間における AR(1) の α の推移

長 20 のサンプルを 50 生成し，入力値として，0 から 1 の範囲に基準化し，バッチサイズ 250 として，損失関数の値が十分に低下する 600 回まで学習させた．

α については次章で報告する株価のペア以外にも適用できる一般的な値と考えており，提案手法は実質 σ をデータから推定するだけで他の対象にも応用できる，即ち分散を計測するのに十分な短期間のデータがあれば適用できる，と考えているが，この点については他の株価データにより今後確認していく必要がある．

3 実証分析

提案手法の有効性を評価するために，TOPIX 指数と日経平均株価のペアを対象に分析を行った．これらは統計的に共和分性が担保されているだけでなく，経験的にも当該ペアは共和分性を満たしていると考えられている．データは Bloomberg から取得した．検証期間は 2007 年 1 月から 2017 年 7 月末までとし，20 営業日後のスプレッドを予測対象とした．スプレッドを作成する際に使用するパラメータ Y は日経平均株価，X は TOPIX 指数とし，定数 β については，年始最初の営業日のタイミングで推定を行い，毎年更新する．また，AR 過程学習時に使用する σ は 1985 年 1 月から 2006 年 12 月末までの値を用いた．なお，簡単のため収益率の計算はスプレッドの予測が上昇なら式 (1) 下落なら式 (2) を 20 で割った値とした．

$$(S_{t+20} - S_t)/(Y + abs(\beta X)) \tag{1}$$

$$-(S_{t+20} - S_t)/(Y + abs(\beta X)) \tag{2}$$

また，比較対象として，実際のスプレッドを学習した LSTM(raw) を加える．学習に用いたスプレッドの期間は 1985 年 1 月から 2006 年 12 月末までであり，当該期間における時系列長 20 のサンプルを 250 生成し，入力値として，0 から 1 の範囲に基準化し，AR 過程を学習させた LSTM と学習パラメータを一致させ，バッチサイズ 250 として，損失関数の値が十分に低下する 600 回まで学習させた．

表 1: 各モデルにおける正答率および収益率

手法	正答率	収益率
AR(1)	56.01 %	13.45 %
LSTM(raw)	56.59 %	14.06 %
LSTM(AR)	58.33 %	17.26 %

前述の AR 過程を学習した LSTM を用いる提案手法の結果 LSTM(AR) と，標準的な AR の予測結果 AR(1)，および，実際のスプレッドで学習した LSTM で予測した結果 LSTM(raw) の比較を表 1 に示す．

また各収益率の推移を図 2 に，F 値を表 2 に示す．表 2 においては，価格が上昇した場合 1, 下落した場合 0, 平均 ave 毎に結果をわけて示す．

正答率・収益率ともに，LSTM(raw) は標準的な方法である AR(1) の結果を再現できており，LSTM がスプレッドの時系列をの学習する能力を持つ事を示している．また提案手法による結果 LSTM(AR) は正答率・収益率ともに AR(1) および LSTM(raw) を上回っている．特に収益率の変化 (図 2) は LSTM(raw) を安定して上回っている．

実際のスプレッドを直接学習した LSTM(raw) より一旦パラメータ α, σ を取り出した結果を使って生成した人工データを学習した LSTM(AR) が精度良く予想できる結果は興味深い．AR(1) を用いた共和分を満たす時系列データ分析の妥当性，言い替えると共和分に関連する既存研究が妥当であった事を示唆していると考える．

4 まとめ

本研究では，株価そのものではなく，共和分性を満たす株価ペアのスプレッドを機械学習で予測する手法を提案した．具体的には，LSTM を用いて，定常性を満足するモデルで最も代表的な AR(1) 過程に従う人工的な時系列データを事前学習する．従来研究で一定とされていた共和分の性質を決めるパラメータが時間により変化する事を LSTM を使ってモデル化する事が提案手法の特徴である．

学習した LSTM を用いてペア・トレード戦略に適用した結果，単純な AR(1) 過程や実データを直接学習させた LSTM よりも良好な結果が得られた．

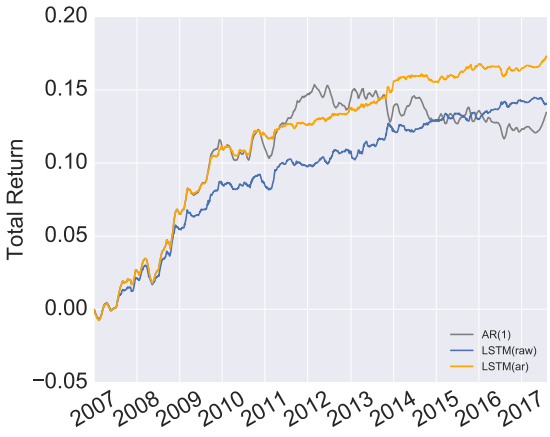

図 2: 各モデルの収益率

表 2: 各モデルの F 値

method		precision	recall	f1-score
AR(1)	0	50 %	63%	56 %
	1	63 %	51%	56 %
	ave	57 %	56%	56 %
LSTM(raw)	0	52 %	25%	34 %
	1	58 %	81%	68 %
	ave	55 %	57%	53 %
LSTM(AR)	0	53 %	46%	50 %
	1	62 %	68%	65 %
	avg	58 %	58%	58 %

参考文献

[1] Walter Enders and Pierre L Siklos. Cointegration and threshold adjustment. *Journal of Business & Economic Statistics*, Vol. 19, No. 2, pp. 166–176, 2001.

[2] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pp. 251–276, 1987.

[3] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, Vol. 19, No. 3, pp. 797–827, 2006.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[5] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *In The International Conference on Learning Representations (ICLR)*, 2015.

[6] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, Vol. 5, 2015.

[7] 中川慧. 非線形共和分関係に基づくペアトレード戦略. テクニカルアナリストジャーナル, Vol. 3, pp. 1–8, 2016.