ランダム分布埋め込み理論による短期予測精度向上手法の日本株における可能性検証

杉友盛佑 前多啓一 *

本研究では、ある時刻のたくさんの観測変数の値からランダムに変数を選んでその時点でのアトラクターの状態を推定するランダム分布埋め込み手法を、日本株の将来リターン予測に用いた時、単純な線形回帰等の従来手法に対し予測精度が向上することを示す。また、金融市場に対しランダム分布埋め込み手法を適用する際の留意点、実務においての具体的な今後の適用展望も示していく。

1. はじめに

株式運用において、トレード対象となる株式のリターン予測を精度良く行うことは、運用において重要な問題である。だが、金融データはシグナルノイズレシオが非常に低く、データ間の関係性が複雑に絡み合い、時系列に十分なサンプル数を得ることも難しく、その予測は容易ではない。しかし、金融データのなかでも、株式の特徴として、時系列方向のデータ量は多くはないが、銘柄数は非常に多く、同時計測が可能であるという特徴がある。そこで、2018 年 10 月に提案された、多変数の同時計測からなる短い時系列データから、重要なターゲット変数の将来の変化を高精度に予測するための数学理論である、ランダム分布埋め込み手法と株式市場におけるリターン予測は親和性が高いと考えられる。本研究では、この手法を用いたリターンの回帰予測と従来の代表的予測手法である、最小二乗法による線形回帰、LASSO 回帰を用いた予測とを比較することで、ランダム分布埋め込み手法の有効性を評価していく。

2. 先行研究と研究背景

1. アトラクタの再構成

自然界において観測，測定される不規則時系列信号の解析は「カオス時系列解析」として研究されてきた．不規則時系列データを決定論的力学系の観点から解析しようとする場合，まず初めに行わなければならないのはアトラクタの再構成である [S91]．アトラクタ再構成の中でもっともよく使われるのは時間遅れアトラクタを用いた再構築である．時間遅れアトラクタとは、ある変数 $x_k(t)$ に関して、時間遅れ座標系 $(x_k(t), x_k(t+\tau), x_k(t+2\tau), \cdots)$ を用いて力学系のアトラクターを再構築することである。このとき、Takens の埋め込み定理 [T81] および，一般化埋め込み定理 [S91] により，時間遅れ座標系の次元が一定以上大きければ、力学系のオリジナルなアトラクターから，再構築したアトラクターへの埋め込みが存在する．一方，非時間遅れアトラクタとは、$x_i(t)$ からランダムに M 個（M は時間遅れ座標系の次元と同じ）を選び、それからなる座標系 $(x_{m1}(t), x_{m2}(t), \cdots, x_{mM}(t))$ を用いて力学系のアトラクターを再構築することであり，こちらにもオリジナルアトラクターからの埋め込み Γ が存在する．

2. ランダム分布埋め込み手法

2018 年 10 月に、合原ら [A18] によって提案された、高次元で長さが短い時系列データを高精度に予測するための手法である。まず、観測データ $x_i(t), i = 1, \cdots, n$(t は時刻) に対して、時間遅れアトラクタと非時間遅れアトラクタを再構築する。このとき，埋め込み理論により，埋め込み Φ，Γ と協調的な微分同相写像 Ψ が存在することになり [S91]，これらをサンプルから学習することで，非時間遅れアトラクタを用いて，時間遅れアトラクタの精度が高い予測を可能にするというものである．

* 連絡先: 杉友盛佑、エピックパートナーズインベストメンツ株式会社、sugitomo@epicgroup.jp, 前多啓一，東京大学大学院数理科学研究科，maeta@ms.u-tokyo.ac.jp

3. 日本株への適用

次に、上記ランダム分布埋め込み手法をどのように日本株のリターン予測に適用するかを考える。今回の手法を適用するにおいて留意すべき点は、観測データにおける各変数が、同一の力学系に起因するサンプルであるのかという点である。リターン予測の際によく用いられるリスクファクター等は同一の力学系に起因するとは考えにくい一方，同業種に含まれる個別銘柄の各リターンは、同一の力学系に起因する可能性が高い.（ここ，理由がもう少し必要かと思います）そこで、本研究においては、同業種に含まれる個別銘柄のリターンを用いて、特定銘柄のリターンを予測していくことを目的とする。

4. ガウス過程回帰

ガウス過程回帰はノンパラメトリックな回帰モデルである[1]．2 変数 $\boldsymbol{x} \in \mathbb{R}^n$, $t \in \mathbb{R}$ に対し，基底関数 $\phi : \mathbb{R}^n \to \mathbb{R}^n$ を用いて，$t = \boldsymbol{w}^T\phi(x) + \varepsilon$ という関係が成り立っていると仮定する．ただし，重み付け $\boldsymbol{w}$ と誤差 ε について，$\boldsymbol{w} \sim N(\boldsymbol{0}, \alpha^{-1}I_n)$ および，$\varepsilon \sim N(0, \beta^{-1})$ が成り立っているとする．このとき，テストデータ $(\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2)\cdots, (\boldsymbol{x}_n, t_n) \in \mathbb{R}^n \times \mathbb{R}$ と新しい入力 $\boldsymbol{x}_{n+1}$ から，えられる出力 t_{n+1} の分布 $p(t_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n, t_1, t_2, \cdots, t_n) = N(t_{n+1}|m, \sigma^2)$ を推定する．カーネル関数 $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ を，$k(x, x') = \alpha^{-1}\phi(x)^T\phi(x')$ と定め，

$$
\begin{aligned}
K &= \{k(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j}, \\
\boldsymbol{t} &= (t_1, \cdots, t_n)^T, \\
\boldsymbol{k} &= (k(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}), \cdots, k(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}))
\end{aligned}
$$

とするとき，最適な推定は以下で与えられる．

$$
\begin{aligned}
m &= \boldsymbol{k}^T K^{-1}\boldsymbol{t} \\
\sigma^2 &= k(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1}) + \beta^{-1} - \boldsymbol{k}^T K^{-1}\boldsymbol{k}
\end{aligned}
$$

[1]詳しくは [B13] を参照されたい.

3. 実験

1. 実験手順

ユニバースを TOPIX500 として、東証 33 業種を用いていくつかの業種内においてランダム分布埋め込み手法を適用していく。本研究においては、ある程度銘柄数が業種内に存在し、内需･外需での結果の変化等も見ていくため、建設、化学、食品、機械の 4 業種を予測対象とする。ランダム分布埋め込み手法に関しては、以下の手順で予測を行う。与えられたデータは、n 個の観測ポイントにおける関数 $x : \mathbb{R} \to \mathbb{R}^n t \mapsto (x_1, \cdots, x_n)$ の時間 $t_1, \cdots, t_m$ でのデータである。推定するのは k 番目の特定の変数 x_k の翌日日中リターンである。まず、$(1, 2, \cdots, n)$ のなかから，L 個の数が入っているタプルを s 個選ぶ。そして、l 番目のタプルから，次の値を最小化する $\psi_l : \mathbb{R}^L \to \mathbb{R}$ をガウス過程回帰を用いて推定する．

$$
\sum_{i=1}^{m-1} |x_k(t_{i+1}) - \psi_l(x_{l_1}(t_i), x_{l_2}(t_i), \cdots, x_{l_L}(t_i))|
$$

その後、各 ψ_l より 1 ステップの推定 $\tilde{x}_k^l(t+1) = \psi_k^l(x_{l_1}(t), \cdots, x_{l_L}(t))$ を計算し、集めてできた推定の集合から，カーネル密度推定を行うことで，確率密度関数 $p(x)$ を推定する。そして、確率密度関数の歪度 γ を計算し，γ が 0.5 以下であれば採用し，$\tilde{x}_k(t+\tau) = \int xp(x)dx$ を推定として確定する。（実は，プログラム内で以下の考察をやっていないので，プログラム修正してやり直さなければいけません）そうでなければ，以下のように推定値を修正する。交差検証によりインサンプルエラー δ_l を計算し，それに従って r 個のベストなサンプルを選び出す。

$$
\tilde{x}_k(t+\tau) = \sum_{i=1}^{r} \omega_i \tilde{x}_k^{l_i}(t+\tau)
$$

ここで，$\omega_i = \dfrac{\exp(-\delta_i/\delta_1)}{\sum_j \exp(-\delta_j/\delta_1)}$ である。本研究では、予測期間を 2016 年とし、$L = 10, s = 3$ として推定を行った。データは各業種に含まれる各銘柄の日中リターンで

ある。そして、各業種に含まれる各銘柄を、それぞれ 1 期間ずつ予測し、予測期間全体での実際の日中リターンとの平均二乗誤差 (MSE) の、業種全体での平均値を精度予測用の指標とする。比較対象として、ランダム分布埋め込み手法を用いずに、同様の期間で各銘柄をその業種の他の銘柄を使って、同様に $L = 10$ とした時に線形回帰、Lasso 回帰で予測した時に、その実リターンとの MSE の平均値を計算する。

2. 実験結果

	Linear	Lasso	埋め込み手法
建設	0.75427	0.75644	0.6943
化学	3.2203	2.6910	2.0573
食品	1.7092	0.5125	0.3745
機械	0.6737	0.5925	0.51308

TABLE 1—

TABLE 1 に実験の結果を示す。結果として、ランダム分布埋め込み手法が全ての業種で最も精度のよい手法となった。ここで、建設・化学・機械の業種に比べ、食品において本手法の改善幅が大きい。ランダム分布埋め込み手法が機能するためには、前提として、分析する変数群が、同じアトラクターにのっている必要がある。その意味において、建設や化学や機械に比較して、食品業種に含まれる銘柄群は、同じアトラクターにのっている、つまり銘柄間の関係性が比較的近い業種であるということが結果から推察できる。

4. 結論と考察

本研究では、ある時刻のたくさんの観測変数の値からランダムに変数を選んでその時点でのアトラクターの状態を推定するランダム分布埋め込み手法を、日本株の将来リターン予測に用いた時、単純な線形回帰、Lasso 回帰に対し予測精度が向上することを示した。また、その予測精度の改善幅は、業種によって違いが存在し、業種に含まれる銘柄群の性質が、同じアトラクターにどの程度のっているかどうかに依存すると推察することができる。

本研究の今後の展望として、例えば複数のボラティリティインデックスなど、より同じアトラクターにのっている可能性の高い金融商品に対して適用することで、より高精度の予測精度を目指すということがまず考えられる。また、ランダム分布埋め込み手法に用いる回帰手法に、LSTM 等のアルゴリズムなどを用いてることで予測精度の改善を目指すということも考えられる。更には、本手法による従来手法からの予測精度改善幅を利用することで、例えばペアトレードなど、銘柄間の性質の近さが重要になる投資手法における銘柄選択のフィルタリングに用いる、などが考えられる。

REFERENCES

[A18] Kazuyuki Aihara et al. "Randomly distributed embedding making short-term high-dimensional data predictable" 2018

[B13] Christopher M. Bishop "Pattern Recognition and Machine Learning" 2013

[E11] Ethan R. Deyle et al. "Generalized Theorems for Nonlinear State Space Reconstruction" 2011

[IA97] 池口徹, 合原一幸. (1997). 力学系の埋め込み定理と時系列データからのアトラクタ再構成 (力学系理論-応用数理における新しい展開). 応用数理, 7(4), 260270.

[T81] Dold, E. a, Takens, F., Teissier, B. (1981). Danamical Systems and Turbulence. J. Jacod Proceedings J. Staffans. VIII Algebraic Topology Proceedings. Edited by P. Hoffman and V. Sna1th. XI Complex Analysis Proceedings (Vol. 701).

[S91] Sauer, T., Yorke, J. A., Casdagli, M. (1991). Embedology.

[PCFS80] Packard, N. H., Crutchfield, J. P., Farmer, J. D., Shaw, R. S. (1992). Geometry from a time series. PhysRevLett 45