

AI による数法則発見の時系列データへの 拡張と金融データへの応用

Modeling and Visualization of Social Reality
Using Latent Profile Analysis and Number Law Discovery Methods
for Evidence-Based Policy Making

蒲田 涼馬 (Ryoma Gamada)
u455007@st.pu-toyama.ac.jp

富山県立大学大学院 工学研究科 電子・情報工学専攻
情報基盤工学講座

N212, 09:30-10:00 Tuesday, February 13, 2024.

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

情報技術の発達により、社会における様々なデータを観測・収集することが可能に

→ 経済分析においても将来予測などの研究が急速に発展.
しかし要因分析に関する研究はそれほど進んでいるとは言えない.

経済に影響を与える要因を分析する研究

因果探索による要因分析.
シンボリック回帰を用いた要因分析.

本研究

シンボリック回帰を用いて分析を行う.

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

なぜシンボリック回帰？

経済分野では，原因と結果の間に成り立つ関係性が重要

→ 複数の要因が複雑に影響しあうため，因果探索では具体的にどのように絡み合っているかを詳細に分析できない。

研究の意義

専門的な知識がなくても為替リスクヘッジ戦略を最適化することができる。

アプローチ

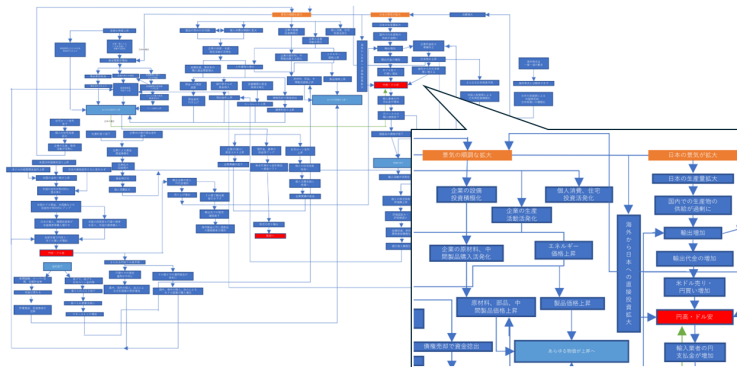
経済波及メカニズムに関する時系列データを分析

→ どの変数がどのように絡み合っているかを数理モデルで表す。

数理モデルの例

$$\text{USD/JPY} = 2.0 \cdot (\text{Nikkei225}) + 2.6 \cdot (\text{OIL}) - 1.0 \cdot (\text{VIX})$$

経済における様々な要因間の関係を表す経済の波及メカニズムがある。



はじめに

経済における波及メカニズム

統計データの特徴と研究の概要

End to end symbolic regression with transformers

今回やったこと

データの拡張

実験概要 2

おわりに

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

公開されているデータ

経済波及メカニズムを参考に用いるデータを調査したところ、約 50 の変数が収集可能であった。

Table 1: 公開されている様々な経済データ

データ項目	
為替レート	金利
コモディティ価格	エネルギー価格
マネーストック	ボラティリティ指数
出来高	スプレッド
株価指数	ニュース

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

目的

時系列経済データを用いて、時系列を考慮した数法則の発見を行いデータ間の関係性をモデル化する手法を提案する.

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

使用する手法の概要

機械学習を用いたシンボリック回帰手法である「End to end symbolic regression with Transformers」を拡張させ、時系列を考慮した分析を行う。可読性と得られる情報量を重視し、人間が式を見ることである程度その式が何を表しているのかをわかるレベルのものを生成させる。

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

End to end Symbolic Regression with Transformers の概要

データから数式を自動発見する深層学習アプローチ

従来のシンボリック回帰の課題: 計算コストが非常に高い.

Transformer アプローチの着想:

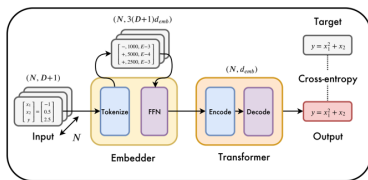
数式は、演算子、定数、変数といった要素が並んだシーケンスとして表現する.

例) $y = x + 2 \cdot \sin(z) \rightarrow + \times * 2 \sin z$ Transformer はシーケンスデータの複雑なパターン学習と高速なシーケンス生成能力を持つため、数式発見に応用できる.

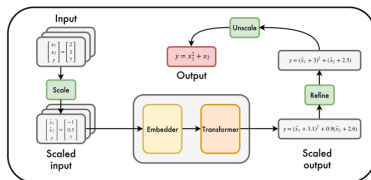
Transformer によるシンボリック回帰 (前準備)

9/23

Transformer によるシンボリック回帰



Training



Inference

Embedder

入力データをトークン化する。

埋め込みルックアップテーブルを使ってトークンをベクトルに変換する (512 次元)。

ベクトルを FFN に入力し、短いベクトルに圧縮する。

この処理をすべてのデータ点に対して行う。

これを Transformer 本体に渡す。

Transformer のメカニズム

Transformer はエンコーダーとデコーダーから成り、Attention メカニズムが核を担っている。

■ Attention メカニズム

シーケンス内の各要素が、ほかのどの要素に注意を向けるべきかを動的に判断し、その重要度について重みづけを行う。

$$Scores = QK^T \quad (1)$$

$$ScaledScore = \frac{QK^T}{\sqrt{d_k}} \quad (2)$$

$$AttentionWeight = softmax(\frac{QK^T}{\sqrt{d_k}}) \quad (3)$$

$$Attention = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

ここで Q はクエリ行列、 K はキー行列、 V は値、 d_k はキーの次元を意味する。

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

■ Feed-Forward Network

ベクトルを圧縮する際やベクトルを拡張する際に使われる。
Transformer モデルでは両方の使われ方がされている。

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (5)$$

ここで x は前のステップからの入力, W_1, W_2, b_1, b_2 はモデルが学習する重みとバイアスを意味する。

エンコーダー

入力されたデータ点を読み込み, それらがどんな関数の特徴を持っているかを分析する。

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

エンコーダー

Embedder によってベクトル化された N 個のデータ点を受け取る。
関連性の計算 (self-attention) を行い、各データ点がほかのすべてのデータ点とどの程度関連しているかを計算する。
FFN によって関連性を加味した情報をさらに深く処理する。
関連性の計算と FNN による処理を 4 層繰り返す。
入力データ全体の特徴を要約した情報をデコーダーに渡す。

エンコーダーの self-attention

入力された全データ点の、相互の関連性の強さを計算する。
自分自身を含むすべての入力データ点を参照する。
これによって各データ点のベクトルにデータセット全体における他の点との関係性を埋め込む。

デコーダー

エンコーダーからの情報と数式の一部を受け取る (初回は開始トークンのみを受け取る).

自己参照 (初回は開始トークンのみ, 2 回目以降はトークン列を受け取る) エンコーダーからの要約情報に注目し, 作る関数を理解する.

上記 2 角情報をもとに, 数式の次に来るべきトークンを予測する.

自己参照から予測までの流れを終了トークンが出力されるまで 1 トークンずつ繰り返す.

完成した数式を出力する.

デコーダーの自己参照

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}} + M)V \quad (6)$$

注目しているトークンとほかの全トークンとの関連度を計算する.

マスクを利用して未来のトークンに関する部分を $-\infty$, それ以外の部分を 0 にする.

これに softmax 関数を適用することでモデルが過去と現在のトークンのみを参照して次のトークンを予測するようにする.

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

クロスエントロピーによる数式の評価

$$H(p, q) = - \sum_i p(x_i) \log q(x_i) \quad (7)$$

ここで $p(x_i)$ は正解の確率分布, $q(x_i)$ はモデルが予測した確率分布を意味する.

これを計算することで生成した数式の評価を行い, これを高めていくように学習していったモデルを作る.

システム概要



はじめに

経済における波及メカニズム

統計データの特徴と研究の概要

End to end symbolic regression with transformers

今回やったこと

データの拡張

実験概要 2

おわりに

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

今回やったこと

実験用システムの実装
システムの数値実験.

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

用いたデータ

以下のデータを用いて実験を行った。

上の表の変数は月足、下の表のデータは月足で実験を行う。

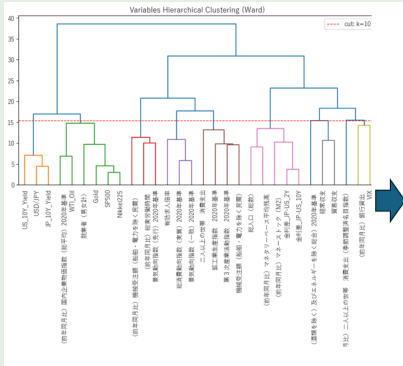
月足データは日足データに拡張して実験を行う。

総人口	鉱工業生産指数	総消費動向指数	マネタリーベース平均残高	輸出物価指数
就業者数	第3次産業活動指数	景気動向指数	経常収支	国内企業物価指数
有効求人倍率	消費者物価指数	マネーストック	貿易収支	
国内企業物価指数	二人以上の世帯消費支出	銀行貸出	輸入物価指数	

SP500	日米金利差(2年)	アメリカ債券利回り(10年)	VIX指数
日経平均225	日米金利差(10年)	オイル価格	
USD/JPY	日本債券利回り(10年)	金価格	

実験内容

レジームごとに作成した実験計画法を使って、シンボリック回帰を行う。
シンボリック回帰によって得られた数理モデルの妥当性について評価する。



Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6	Cluster_7
0 Nikkei225_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	(前年同月比) マネタリーベース 平均預高_lag1	(前年同月比) 機械受注額 (船 舶・電力を除く 反照)_lag1
1 USD/JPY_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	金利差 JP- US_2Y_lag1	機械受注額 (船 舶・電力を除く 反照)_lag1
2 SP500_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	総人口 (総数) _lag1	景気動向指数 (先行) 2020年 基準_lag1
3 Nikkei225_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	(季節調整済み) 鉱工業生産指数 2020年基準_lag1	機械受注額 (船 舶・電力を除く 反照)_lag1
4 Gold_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	総人口 (総数) _lag1	景気動向指数 (先行) 2020年 基準_lag1
5 経常収支 (男女 計)_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	総人口 (総数) _lag1	景気動向指数 (先行) 2020年 基準_lag1
6 Nikkei225_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	金利差 JP- US_10Y_lag1	景気動向指数 (先行) 2020年 基準_lag1
7 Nikkei225_lag1	貿易収支 マナーズトック (M2) 目録数_lag1	(前年同月比) (前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	(前月比) 二人以 上の世帯 消費支 出 (季節調整済み 目録数)_lag1	VIX_lag1	総人口 (総数) _lag1	景気動向指数 (先行) 2020年 基準_lag1

評価指標

評価指標として決定係数と DA(Directional accuracy) を考える

DA → 時系列データの変動方向を正しく予測できた割合を示す指標

時刻 t における実測値を y_t , 予測値を $\hat{y}_t$ とする.

ここで $(t-1)$ からの差分を以下のように定義する.

$$\Delta y_t = y_t - y_{t-1} \quad (8)$$

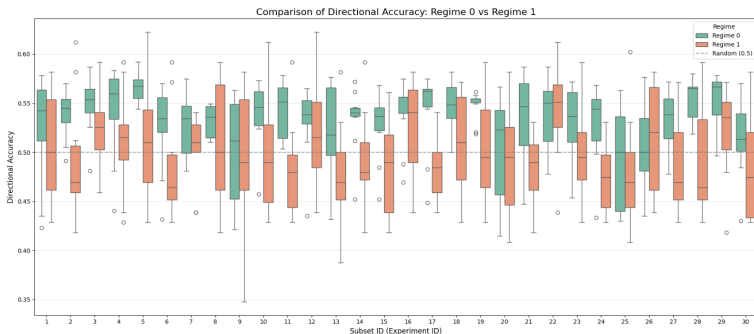
$$\Delta \hat{y}_t = \hat{y}_t - y_{t-1} \quad (9)$$

$$I(x, y) = \begin{cases} 1 & \text{if } \text{sign}(x) = \text{sign}(y) \\ 0 & \text{if } \text{sign}(x) \neq \text{sign}(y) \end{cases} \quad (10)$$

$$DA = \frac{1}{N} \sum_{t=1}^N I(\Delta y_t, \Delta \hat{y}_t) \quad (11)$$

実験結果

レジーム 0 に比べて、レジーム 1 では DA が低い傾向にあった。
しかし、レジーム 1 でも、実験ナンバーによっては高い DA を示すものもあった。
→ 変数の組み合わせによって精度が大きく異なることを示唆する？



はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

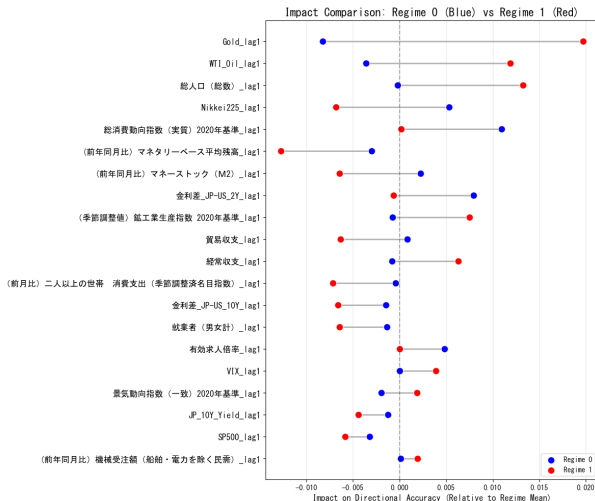
データの拡張

実験概要 2

おわりに

実験結果

はじめに
経済における波及
メカニズム
統計データの特徴
と研究の概要
End to end
symbolic
regression with
transformers
今回やったこと
データの拡張
実験概要 2
おわりに



はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

考察

レジームごとの精度に明らかな違いがあることがわかった.

→ 市場の駆動要因が異なることを示唆している

例えば, レジーム 0 の場合は金利差や株価指数などがプラスの影響を与えており, 市場が理論通りに動いていることを示している.

一方, レジーム 1 では VIX や OIL などの変数が影響を与えており, 市場が不安定な状態であることを示している.

また, 純粹に学習に使うデータ数がレジーム 0 はレジーム 1 の約 4 倍あることも影響している可能性がある.

はじめに

経済における波及
メカニズム

統計データの特徴
と研究の概要

End to end
symbolic
regression with
transformers

今回やったこと

データの拡張

実験概要 2

おわりに

まとめ

今回は両方のレジームについて実験を行い、比較、考察を行った.

- 妥当性は割とありそう.
- 変数によって精度が大きく異なる.
- レジームごとに影響を与える変数が異なる (現状).

今後の展望

- もう少し、分析に力を入れてみる (具体的な有意差などを調べる).
- クラスタ数と変数の絞り込み方法をいじって実験してみる.
- 可読性を向上させるプログラムも作成し、今回と同様の実験を行う.
- 論文を書きつつ、手法について勉強、プログラムの確認.