LSTMAEを用いた長期死亡率予測

明治大学大学院　先端数理科学研究科　現象数理学専攻　松山研究室
博士前期課程1年　冨山　蓮

【研究の背景と概要】

- 近年、日本をはじめとする先進国では長寿化が進んでいる。この長寿化がもたらす社会・経済的な負担の加速度的な増加は、長寿リスク（Longevity Risk）とよばれ、そのリスク評価には長期の死亡率予測が不可欠である。死亡率を予測する際に用いられる代表的なモデルがLee-Carter(LC)モデルであり、年齢x、暦年tの死亡率$m_{x,t}$を次のように表すモデルである。

$$\log(m_{x,t}) = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t} \, , \varepsilon_{x,t} \sim N(0, \sigma^2)$$

（α_x：対数死亡率平均, β_x： κ_tに対応する年齢別感応度, κ_t：死亡率改善度曲線, $\varepsilon_{x,t}$：平均0のノイズ項）

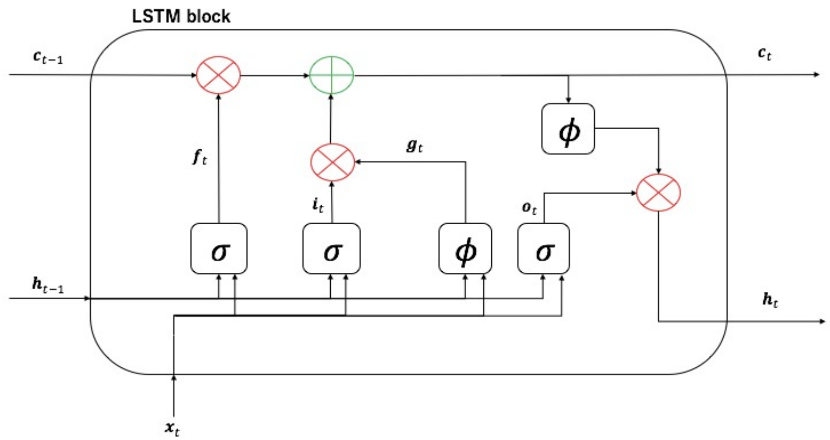
- 本研究ではニューラルネットワーク(NN)手法の中でも時系列との親和性が最も高いRecurrent Neural Network(RNN)の構造を持つLong Short Term Memory(LSTM)とLSTM Autoencoder(LSTMAE)を用いたLee-Carter(LC)モデルの拡張を提案する。

- LCモデルは特異値分解(SVD）による次元削減の第一段階と、得られた時系列成分への時系列モデルの当てはめの二段階で推定される。本研究では、第一段階のSVDによる次元削減をLSTMを用いた自己符号化器であるLSTMAEに置き換え、第二段階の時系列モデルをLSTMに置き換えることでLCモデルの非線形拡張を行う。その結果、年齢成分と時系列成分を分離しない整合性のとれた年齢別時系列を、大災害等の一時的外部要因の影響をうけにくい形で生成でき、旧来のLCモデルや単純にLSTMを用いた先行研究に比して高い死亡率推定精度を実現できることを示す。

LSTMとLSTMAE

シーケンシャルデータを利用した分析ではRNNで起こりうる勾配消失問題の可能性を大きく下げるLSTMを利用する。
LSTM内のそれぞれのゲートの出力は次のように計算される。

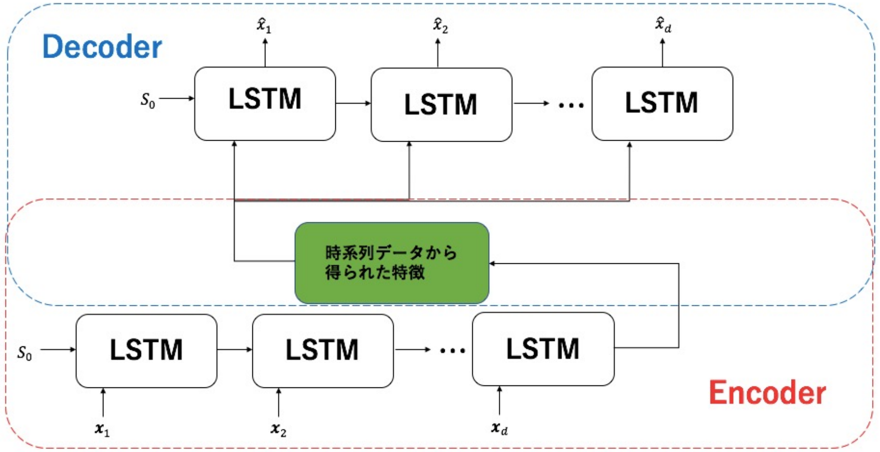
$$\begin{cases} \boldsymbol{f}_t = \sigma\left(W_f^T \boldsymbol{x}_t + U_f^T \boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right) : \text{忘却ゲート} \\ \boldsymbol{i}_t = \sigma\left(W_i^T \boldsymbol{x}_t + U_i^T \boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right) : \text{入力ゲート} \\ \boldsymbol{o}_t = \sigma\left(W_o^T \boldsymbol{x}_t + U_o^T \boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right) : \text{出力ゲート} \\ \boldsymbol{g}_t = \phi\left(W_g^T \boldsymbol{x}_t + U_g^T \boldsymbol{h}_{t-1} + \boldsymbol{b}_g\right) : \text{入力補助} \end{cases}$$



時刻 t のLSTMの出力と長期メモリセルの出力は次のようになる。

$$\begin{cases} \boldsymbol{c}_t = \boldsymbol{c}_{t-1} \odot \boldsymbol{f}_t + \boldsymbol{g}_t \odot \boldsymbol{i}_t : \text{長期メモリセル} \\ \boldsymbol{h}_t = \phi(\boldsymbol{c}_t) \odot \boldsymbol{o}_t : \text{LSTMの出力} \end{cases}$$

右下図がLSTMAEの構造であり、通常のAEと同じでEncoder部分とDecoder部分に大きく分かれており、特徴はEncoder部分では時系列の情報から固定長のベクトルを抽出し、次元削減を行うことである。



提案手法

HMD(Human Mortality Database)から日本のデータを引用している。
年齢範囲：0歳~99歳、歴年範囲：1960年~2018年、
トレーニング：1960年~1989年、テスト：1990年~2018年

本研究ではLSTMAEとLSTMを使用した二段階推定の予測手法を提案する。
推定の手順：($m_{x,t}$は年齢x、歴年tの死亡率とし、α_xは年齢xの対数死亡率平均とする。)
(1)各時刻tに対して次のようにベクトル$\boldsymbol{u}_t$を定義する。

$$\boldsymbol{u}_t := \left[\log\left(m_{age_{min},t}\right) - \alpha_{age_{min}} \quad \cdots \quad \log\left(m_{age_{max},t}\right) - \alpha_{age_{max}}\right]^T$$

(2-1)提案手法1
LSTMAEで推定結果$\hat{\boldsymbol{u}}_t$は次のように計算される。

$$\hat{\boldsymbol{u}}_t = f_{dec}\left(f_{enc}\left(\boldsymbol{u}_{t_{min}}, \ldots, \boldsymbol{u}_{t_{max}}\right)\right)$$

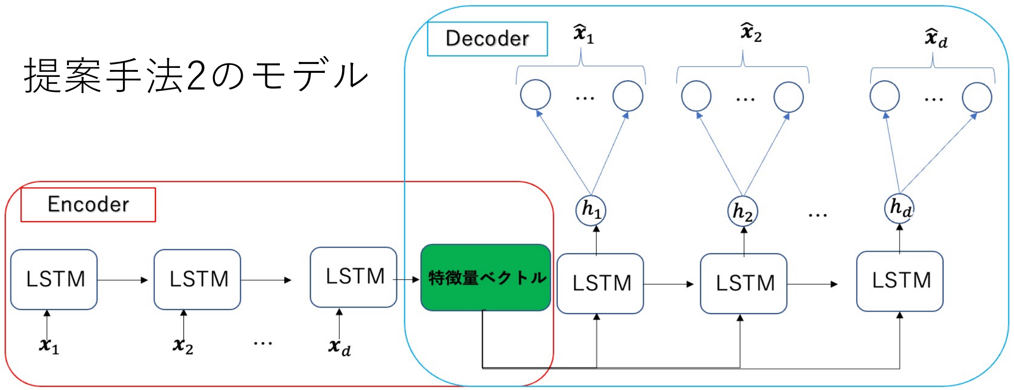
(2-2)提案手法2
(Decoder部分のLSTMの出力を一つとする)
LSTMAEで推定結果$\hat{\boldsymbol{u}}_t$は次のように計算される。

$$\hat{\boldsymbol{u}}_t = W^T h_t + \boldsymbol{b} = W^T f_{dec(LSTM)}\left(f_{enc}\left(\boldsymbol{u}_{t_{min}}, \ldots, \boldsymbol{u}_{t_{max}}\right)\right) + \boldsymbol{b}$$

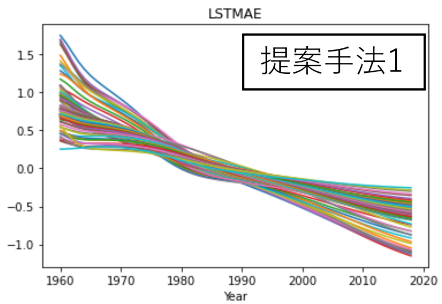
(3)(2-1)と(2-2)で推定された結果を利用してLSTMで外挿を行う。ここで、$u_t^{(x)}$をベクトル$\boldsymbol{u}_t$の$(x+1)$番目の要素とし、次のように外挿を行う。

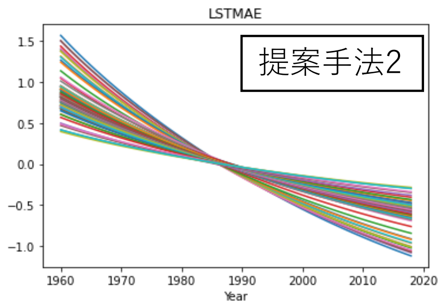
$$\begin{bmatrix} u_{t+1}^{(age_{min})} \\ \vdots \\ u_{t+1}^{(age_{max})} \end{bmatrix} = f_{LSTM}\left(\begin{bmatrix} u_t^{(age_{min})} \\ \vdots \\ u_t^{(age_{max})} \end{bmatrix}\right)$$

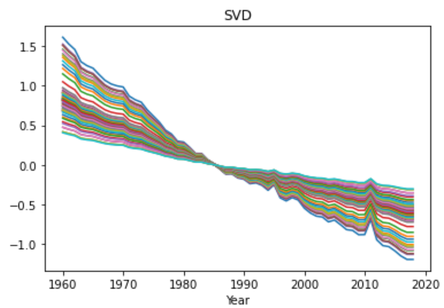
提案手法2のモデル



対数死亡率の再現



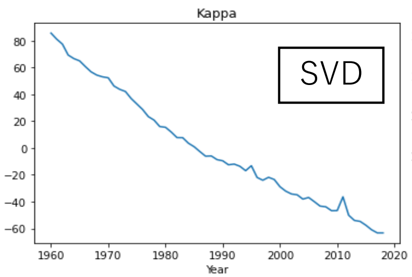


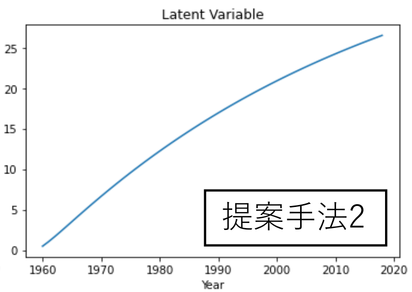


再現精度 (全データ)	MAE
提案手法1	3.364×10^{-2}
提案手法2	5.763×10^{-2}
SVD	5.759×10^{-2}

SVDでは2011年付近で大きく振れており、東日本大震災の影響があると考えられる。しかし、提案手法1,2では減少し続ける結果を得ることができた。
長期の死亡率予測の目的では、地震のような一過性の情報の反映は望ましいことではなく、LSTMAEの次元削減の方が適していると考えられる。
また、再現精度を比較すると提案手法1が他の二つに比べ再現精度が高く、提案手法2とSVDではほとんど変わらない再現精度を出すことができている。

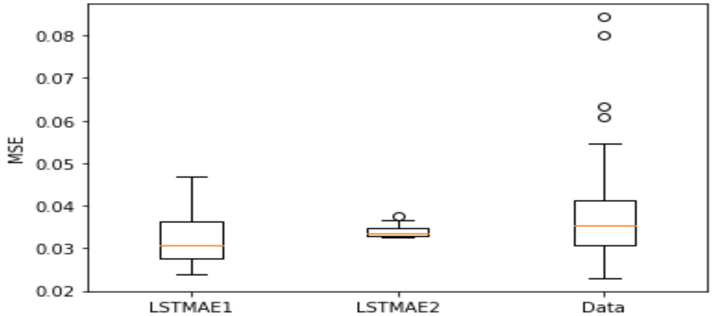
提案手法2のDecoder部分におけるLSTMの各時刻tで出力される潜在変数は右図のようになり、SVDの時系列改善度曲線κ_tと比較する。SVDのκ_tは直線上に減少しているのが分かるが、提案手法2の潜在変数では死亡率改善が上昇していることが示唆されており、直線上には増加せず、徐々に増加が減衰する結果を得られた。





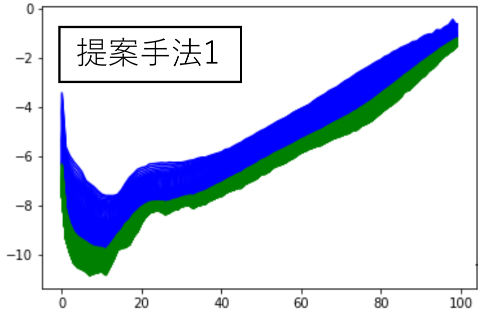
外挿の結果と課題

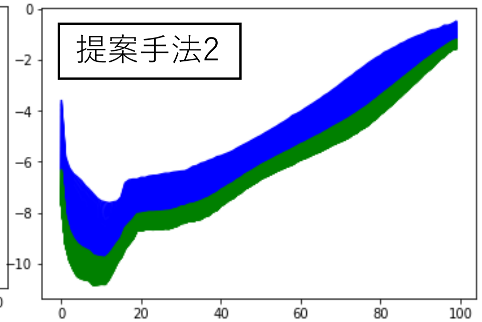
予測精度(seed=10)	提案手法1	提案手法2	Data+LSTM	SVD+LSTM	SVD+ARIMA
MSE	2.587×10^{-2}	3.508×10^{-2}	3.372×10^{-2}	4.278×10^{-2}	1.108×10^{-1}



提案手法1で最も良い予測精度を出すことができ、精度面ではデータに直接LSTMで予測した場合と提案手法2が続いている。NNの不安定性を考慮し、いくつかのシードで予測精度を確認した結果が左図の箱髭図になる。データに直接LSTMを利用すると大きくバラつくことがわかる。提案手法2では他の二つの手法に比べて予測精度は低く出るがシード値による不安定性が解消出来ていることが分かる。

超長期の死亡率予測を考える際には、対数死亡率の形状安定性（不自然なノイズ除去）が鍵となってくると考えられる。右図は2018年までのデータを利用して100年先の超長期死亡率予測した結果である。提案手法1に比べて提案手法2の方が形状が安定していることが分かる。





まとめと課題：
LSTMAEを用いることで地震等の一時的な外部要因の影響を受けにくい次元削減が可能となり、精度の改善だけではなく、今後のポストコロナ(COVID-19)における長期の死亡率予測において有用であると考えられる。
長期の死亡率予測における形状安定性の確保には課題が残り、提案手法2では長期の形状安定性と精度のトレードオフが生じている。