正則化に基づく深層状態空間モデル
Deep state space models with regularization

数学専攻　村松 洋介
MURAMATSU, Yosuke

1　はじめに

近年, デジタル化の進展により膨大なデータが収集されるようになり, 多変量時系列データの活用の機会が増えている. 様々なモデルを包含した汎用性の高い時系列データのモデルとして状態空間モデルがある. 状態空間モデルは様々な要因を分解してモデル化するため, データ間の関連性や影響を把握しやすい. 一方で高次元の時系列データ, とくに相関をもつパネルデータのモデルとして十分な表現能力を有しているとはいい難い. この課題は, 機械学習の方法論である深層学習を用いることで解決することができる. しかし, 深層学習では結果がブラックボックス化して出力されるため, 解釈することが難しい.

そこで近年, 状態空間モデルなどの古くから使われている時系列モデルの考え方をベースに深層ニューラルネットワークを取り込んだモデルが提案されている. これにより, 状態空間モデルと深層学習の利点をそれぞれ取り込むことが期待できる. なかでも状態空間モデルをベースとしているモデルを本研究では深層状態空間モデルとよぶ. 深層状態空間モデルとしては, 線形ガウス状態空間モデルをベースとしている DSSM (Deep State Space Model) [3] や, 非線形ガウス状態空間モデルをベースとしている DeepKF (Deep Kalman Filter) [2] など多くのモデルが提案されている. しかし, 依然として深層学習では結果はブラックボックス化されて得られるため結果の解釈は困難である. そこで本研究では DSSM において正則化項の導入を提案する. 正則化項を導入することで, 変数選択による解釈性の向上や予測精度の向上を目指す.

2　状態空間モデル

状態空間モデルは, 直接観測できない潜在的な状態を導入し, この潜在的な状態の時間変化によって観測値の時間変化を表現するモデルである. 状態空間モデルは, モデルが線形性を有するか否か, ノイズがガウス分布に従うか否かで類別される. ここでは最も単純な, 線形性を有しノイズがガウス分布に従う線形ガウス状態空間モデルを紹介する.

時刻 t における N 次元の観測値時系列を $\boldsymbol{x}_t$, 状態を L 次元とし $\boldsymbol{z}_t$ とする. このとき線形ガウス状態空間モデルは

$$\boldsymbol{x}_t = \boldsymbol{A}_t \boldsymbol{z}_t + \varepsilon_t, \qquad \varepsilon_t \sim N(\boldsymbol{0}, \boldsymbol{E}_t), \tag{2.1}$$

$$\boldsymbol{z}_t = \boldsymbol{F}_t \boldsymbol{z}_{t-1} + \boldsymbol{R}_t \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N(\boldsymbol{0}, \boldsymbol{Q}_t), \tag{2.2}$$

$$\boldsymbol{z}_0 \sim N(\boldsymbol{y}_0, \boldsymbol{P}_0) \tag{2.3}$$

と定義される. これは, 観測値 $\boldsymbol{x}_t$ は状態 $\boldsymbol{z}_t$ とホワイトノイズ ε_t によって変動することを表しており, 状態 $\boldsymbol{z}_t$ は 1 時点前の状態 $\boldsymbol{z}_{t-1}$ とノイズ $\boldsymbol{\eta}_t$ により時間変化していることを表している. ここで, (2.1) 式は観測方程式, (2.2) 式は状態方程式と呼ばれる. また, $\boldsymbol{z}_0$ を初期状態, ε_t を観測値撹乱項, $\boldsymbol{\eta}_t$ を状態撹乱項と呼ぶ. ただし, 各方程式の各係数 $\boldsymbol{A}_t$, $\boldsymbol{F}_t$, $\boldsymbol{R}_t$ はサイズがそれぞれ $N \times L$, $L \times L$, $L \times R$ の行列である.

この線形ガウス状態空間モデルにおける状態の推定を効率的に行う計算アルゴリズムとしてカルマンフィルタがある. カルマンフィルタは初期状態 $\boldsymbol{z}_0$ から出発して各時点 t に対して状態 $\boldsymbol{z}_t$ の 1 期先予測 $\boldsymbol{y}_{t|t-1} = \mathrm{E}[\boldsymbol{z}_t|\boldsymbol{x}_{1:t-1}]$ とその推定誤差分散 $\boldsymbol{P}_{t|t-1} = \mathrm{Var}[\boldsymbol{z}_t|\boldsymbol{x}_{1:t-1}]$, および, 状態 $\boldsymbol{z}_t$ のフィルタ化推定量

$\boldsymbol{y}_{t|t} = \mathrm{E}[\boldsymbol{z}_t|\boldsymbol{x}_{1:t}]$ とその推定誤差分散 $\boldsymbol{P}_{t|t} = \mathrm{Var}[\boldsymbol{z}_t|\boldsymbol{x}_{1:t}]$ を逐次的に求める計算アルゴリズムであり, その更新式は

$$\boldsymbol{v}_t = \boldsymbol{x}_t - \boldsymbol{A}_t\boldsymbol{y}_{t|t-1}, \quad \boldsymbol{V}_t = \boldsymbol{A}\boldsymbol{P}_{t|t-1}\boldsymbol{A}_t^T + \boldsymbol{E}_t, \tag{2.4}$$

$$\boldsymbol{y}_{t|t} = \boldsymbol{y}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{v}_t, \quad \boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{V}_t\boldsymbol{K}_t^T, \tag{2.5}$$

$$\boldsymbol{y}_{t+1|t} = \boldsymbol{F}_t\boldsymbol{y}_{t|t}, \quad \boldsymbol{P}_{t+1|t} = \boldsymbol{F}_t\boldsymbol{P}_{t|t}\boldsymbol{F}_t^T + \boldsymbol{R}_t\boldsymbol{Q}_t\boldsymbol{R}_t^T \tag{2.6}$$

である. ただし $\boldsymbol{x}_{1:t} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t\}$ である. ここで, $\boldsymbol{K}_t$ はカルマンゲインと呼ばれ $\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{A}_t^T\boldsymbol{V}_t^{-1}$ である.

3 深層状態空間モデル

2 節で述べた状態空間モデルは, 高次元の時系列データ, とくに相関をもつパネルデータのモデルとして十分な表現能力を有しているとはいい難い. この問題は深層学習を用いることで解決できる. 時系列データに対応した深層学習の手法として再帰型ニューラルネットワーク (Recurrent Neural Network, RNN) がある. RNN は非線形で柔軟なモデルなので複雑な現象を扱うことが可能である. 一方で, 結果がブラックボックス化されて出力されるため結果を解釈することが難しい.

そこで, 状態空間モデルと深層学習それぞれの優れた点を取り入れたモデルとして深層状態空間モデルが提案されている. 本節では線形ガウス状態空間モデルに深層学習を取り入れたモデルである Deep State Space Model (DSSM) [3] を紹介し, DSSM において正則化項の導入を提案する.

3.1 DSSM : Deep State Space Model

DSSM では N 個の時系列データを同時に考える. 時刻 t における i 番目の観測値時系列を $x_t^{(i)} \in \mathbb{R}$, $i = 1, ..., N$ とし, 時刻 T_i までの観測値 $x_{1:T_i}^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, ..., x_{T_i}^{(i)}\right\}$ が与えられているとする. また, 観測値時系列 $x_t^{(i)}$ に関連する外部変数を $\boldsymbol{u}_t^{(i)}$ とし, 外部変数は観測値が与えられていない時刻 $T_i + 1$ から $T_i + \tau$ でも観測されている, すなわち $\boldsymbol{u}_{1:T_i+\tau}^{(i)} = \left\{\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}, ..., \boldsymbol{u}_{T_i+\tau}^{(i)}\right\}$ が与えられているとする.

DSSM では, 観測方程式, 状態方程式, 初期状態をそれぞれ

$$x_t^{(i)} = \boldsymbol{A}_t^{(i)}\boldsymbol{z}_t^{(i)} + G_t^{(i)}\varepsilon_t^{(i)}, \quad \varepsilon_t^{(i)} \sim N(0, 1), \tag{3.1}$$

$$\boldsymbol{z}_t^{(i)} = \boldsymbol{F}_t^{(i)}\boldsymbol{z}_{t-1}^{(i)} + \boldsymbol{R}_t^{(i)}\boldsymbol{\eta}_t^{(i)}, \quad \boldsymbol{\eta}_t^{(i)} \sim N\left(\boldsymbol{0}, \boldsymbol{I}\right), \tag{3.2}$$

$$\boldsymbol{z}_0^{(i)} \sim N\left(y_0^{(i)}, P_0^{(i)}\right) \tag{3.3}$$

と定義する. ここで, 各方程式と初期状態のパラメータ $\boldsymbol{\Theta}_t^{(i)} = \left\{\boldsymbol{A}_t^{(i)}, G_t^{(i)}, \boldsymbol{F}_t^{(i)}, \boldsymbol{R}_t^{(i)}, y_0^{(i)}, P_0^{(i)}\right\}$ は, 外部変数 $\boldsymbol{u}_{1:t}^{(i)}$ を入力としたニューラルネットワークの出力として得られる, すなわち

$$\boldsymbol{\Theta}_t^{(i)} = \mathrm{NN}_\theta\left(\boldsymbol{u}_{1:t}^{(i)}\right) \tag{3.4}$$

であるとする. ここで NN_θ は i によらず共通の RNN であることに注意する. この構造により, 複数の時系列間で関連性を持たせてモデル化できる.

RNN のパラメータ θ は時系列 $\left\{x_{1:T_i}^{(i)}\right\}_{i=1}^N$ を観測する確率を最大にする, つまり, 条件付き対数尤度関数

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p\left(x_{1:T_i}^{(i)} \middle| \boldsymbol{u}_{1:T_i}^{(i)}\right) = \sum_{i=1}^{N}\sum_{t=1}^{T_i} \log p\left(x_t^{(i)} \middle| x_{1:t-1}^{(i)}, \boldsymbol{u}_{1:T_i}^{(i)}\right) \tag{3.5}$$

を最大にするように学習する. (3.5) 式の各因子 $\log p\left(x_t^{(i)}\middle|x_{1:t-1}^{(i)}, \boldsymbol{u}_{1:T_i}^{(i)}\right)$ は正規分布の再生性より $p\left(x_t^{(i)}\middle|x_{1:t-1}^{(i)}, \boldsymbol{u}_{1:T_i}^{(i)}\right) = N\left(\mu_t^{(i)}, \Sigma_t^{(i)}\right)$ であり, 平均 $\mu_t^{(i)}$ と分散 $\Sigma_t^{(i)}$ はカルマンフィルタの結果よりそれぞれ

$$\mu_t^{(i)} = \boldsymbol{A}_t^{(i)} \boldsymbol{F}_t^{(i)} y_{t-1|t-1}^{(i)}, \tag{3.6}$$

$$\Sigma_t^{(i)} = \boldsymbol{A}_t^{(i)} \left(\boldsymbol{F}_t^{(i)} \boldsymbol{P}_{t-1|t-1}^{(i)} {\boldsymbol{F}_t^{(i)}}^T + \boldsymbol{R}_t^{(i)} {\boldsymbol{R}_t^{(i)}}^T \right) {\boldsymbol{A}_t^{(i)}}^T + {G_t^{(i)}}^2 \tag{3.7}$$

となる.

RNN により θ が推定できれば, 観測値を確率的に予測することができる. まず, パラメータ θ を学習したことで RNN の構造は決定しているため, RNN に $\boldsymbol{u}_{T_i+1:T_i+\tau}^{(i)}$ を入力することで $\boldsymbol{\Theta}_{T_i+1:T_i+\tau}^{(i)}$ が求まる. また, 時刻 T_i での状態の事後確率 $p\left(\boldsymbol{z}_{T_i}^{(i)}\middle|x_{1:T_i}^{(i)}\right)$ も求まっているので, この分布からサンプル $\boldsymbol{z}_{T_i}^{(i)}$ を得る. 以降は $t = 1, 2, ..., \tau$ について,

$$\hat{x}_{T_i+t}^{(i)} = \boldsymbol{A}_{T_i+t}^{(i)} \boldsymbol{z}_{T_i+t}^{(i)} + G_{T_i+t}^{(i)} \varepsilon_{T_i+t}^{(i)}, \tag{3.8}$$

$$\boldsymbol{z}_{T_i+t+1}^{(i)} = \boldsymbol{F}_{T_i+t+1}^{(i)} \boldsymbol{z}_{T_i+t}^{(i)} + \boldsymbol{R}_{T_i+t+1}^{(i)} \boldsymbol{\eta}_{T_i+t+1}^{(i)} \tag{3.9}$$

と再帰的に計算することで予測値を導出することができる.

3.2 正則化に基づく深層状態空間モデル

本節では DSSM において L_1 正則化項, L_2 正則化項の導入を提案する. まず主に過学習を抑制することが知られている L_2 正則化項を導入する, ただし, 特定の条件下ではスパースな構造を学習するという研究 [4] があり, L_2 正則化を導入した場合でも変数選択による解釈性の向上が期待できる.

DSSM に組み込まれている RNN のパラメータを θ として, 最適化する目的関数を $E_{L_2}(\theta)$, θ の条件付き対数尤度関数 $\mathcal{L}(\theta)$ とおき, L_2 正則化項を導入した目的関数

$$E_{L_2}(\theta) = \mathcal{L}(\theta) + \lambda \sum_{i=1}^{N} \sum_{t=1}^{T_i} \left|\left| \boldsymbol{A}_t^{(i)} \right|\right|_2^2$$

を提案する. このとき λ は正則化パラメータである. この目的関数 $E(\theta)$ は微分可能であるため通常の DSSM と同様に確率的勾配降下法を用いて RNN のパラメータ θ の勾配を導出しパラメータの学習を行うことができる.

次に, 係数を真に 0 に縮小する効果が知られている L_1 正則化項の導入を提案する. 目的関数は

$$E_{L_1}(\theta) = \mathcal{L}(\theta) + \sum_{i=1}^{N} \sum_{t=1}^{T_i} \lambda \left|\left| \boldsymbol{A}_t^{(i)} \right|\right|_1 \tag{3.10}$$

である. ここで, λ は正則化パラメータである. ただし, L_1 正則化項は微分可能ではないため, L_2 正則化項と同様に重みの勾配を計算することはできない. そこで L_1 正則化項と似た形をしている SCAD 正則化項の 2 次近似 [1] と同様の方法で近似する. まず,

$$\sum_{j=1}^{L} p_\lambda\left(\left|A_{t,j}^{(i)}\right|\right) \coloneqq \sum_{j=1}^{L} \lambda \left|A_{t,j}^{(i)}\right| = \lambda \left|\left| \boldsymbol{A}_t^{(i)} \right|\right|_1$$

とする. ただし, $\boldsymbol{A}_t^{(i)} = \left(A_{t,1}^{(i)}, A_{t,2}^{(i)}, ..., A_{t,L}^{(i)}\right)^T$ である. このとき $p_\lambda\left(\left|A_{t,j}^{(i)}\right|\right)$ の $A_{t,j}^{(i)} = \tilde{A}_{t,j}^{(i)}$ のまわりでのテイラー展開による 2 次近似は

$$p_\lambda\left(\left|A_{t,j}^{(i)}\right|\right) \approx \lambda \left|\tilde{A}_{t,j}^{(i)}\right| + \frac{\lambda\left({A_{t,j}^{(i)}}^2 - {\tilde{A}_{t,j}^{(i)}}^2\right)}{2\left|\tilde{A}_{t,j}^{(i)}\right|}$$

と表すことができる. よって (3.5) 式を利用し, SCAD 正則化項の 2 次近似 [1] と同様に (3.10) 式を

$$
\begin{aligned}
E_{L_1}(\theta) &= \mathcal{L}(\theta) + \sum_{i=1}^{N}\sum_{t=1}^{T_i} \lambda \left|\left| \boldsymbol{A}_t^{(i)} \right|\right|_1 \\
&\approx \mathcal{L}(\tilde{\theta}) + \sum_{i=1}^{N}\sum_{t=1}^{T_i} \left\{ \frac{\partial}{\partial \boldsymbol{A}_t^{(i)}} \log p\left(x_t^{(i)} \middle| x_{1:t-1}^{(i)}, \boldsymbol{u}_{1:T_i}^{(i)} \right) \right\}^T \left(\boldsymbol{A}_t^{(i)} - \tilde{\boldsymbol{A}}_t^{(i)} \right) \\
&\quad + \frac{1}{2} \left(\boldsymbol{A}_t^{(i)} - \tilde{\boldsymbol{A}}_t^{(i)} \right)^T \sum_{i=1}^{N}\sum_{t=1}^{T_i} \left\{ \frac{\partial^2}{\partial \boldsymbol{A}_t^{(i)} \partial \boldsymbol{A}_t^{(i)^T}} \log p\left(x_t^{(i)} \middle| x_{1:t-1}^{(i)}, \boldsymbol{u}_{1:T_i}^{(i)} \right) \right\} \left(\boldsymbol{A}_t^{(i)} - \tilde{\boldsymbol{A}}_t^{(i)} \right) \\
&\quad + \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{j=1}^{L} \left\{ \lambda \left| \tilde{A}_{t,j}^{(i)} \right| + \frac{\lambda \left({A_{t,j}^{(i)}}^2 - {\tilde{A}_{t,j}^{(i)}}^2 \right)}{2 \left| \tilde{A}_{t,j}^{(i)} \right|} \right\}
\end{aligned}
$$

と近似する. ただし, $\tilde{\theta}$ は DSSM の観測方程式の係数 $A_{t,j}^{(i)}$ が $\tilde{A}_{t,j}^{(i)}$ と推定されたときに学習されているネットワークのパラメータである. この近似した目的関数はネットワークのパラメータ θ に関しての微分操作が可能であり, DSSM と同様にパラメータの推定が行える. ただし, $\tilde{A}_{t,j}^{(i)} = 0$ のときは $1/\left|\tilde{A}_{t,j}^{(i)}\right|$ が計算できないため, 非 0 要素のみに対して計算する. そのため, この近似手法では一度 0 と推定されたパラメータは再び非 0 に推定されることはない.

4 今後の展望

本研究では, 線形ガウス状態空間モデルをベースに深層学習を取り込んだ DSSM において正則化項の導入を提案した. これにより予測精度や解釈性の向上が期待できる.

今後の課題として次の 2 つを挙げる. 1 つ目は本研究で提案したモデルについても解釈可能であるかは用いるデータに大きく依存することである. 潜在状態から観測値が生成されるメカニズムの解釈性は向上することが期待できるが, 潜在状態自体の解釈については改善する必要があると考える.

2 つ目は, 数多くある深層状態空間モデルのうち, どのモデルが解釈性に優れているのかの判断が難しいことである. いままで提案されてきた深層状態空間モデルは各モデルの扱うデータの特性やモデルの開発の背景が異なるため, 統一的に比較することができない. 様々な特性のデータに対して各モデルを用いて分析を行い, 各モデルの性能を比較することが必要であると考える.

参考文献

[1] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Lilelihood and its Oracle Properties, *journal of the American Statistical Association* 96. 1348-1360.

[2] Krishnan, R. G., Shalit, U. and Sontag, D. (2015). Deep Kalman Filters, *arXiv preprint, arXiv*:1511.05121.

[3] Rangapuram, S. S., Seeger, M., Gasthaus, J., Stella, L., Wang, Y. and Januschowski, T. (2018). Deep state space models for time series forecasting, In: *A*dvances in Neural Information Processing Systems, pp.7785-7794.

[4] 谷口敦司, 浅野渉, 谷沢昭行. (2019). 重み係数のスパース化による深層ニューラルネットワークのコンパクト化技術, 東芝レビュー 74 巻 4 号.