機械学習を用いた時系列ヘルスケアデータの解析と応用

東京大・医科学研究所　長谷川 嵩矩　東京大・医科学研究所　新井田 厚司
東京大・医科学研究所　山口 類　東京大・医科学研究所　井元 清哉

1. はじめに

本報告では, 線形ガウス型状態空間モデルとその拡張モデルに関して、状態変数とパラメータ値を推定する方法として正則化を用いたスパース学習法を提案すると同時に, 提案手法を時系列の健康診断データに適応したときの状態変数同士もしくは状態変数と観測変数の制御関係と，遺伝的形質や生活習慣などの外部環境因子が将来の結果に与える影響を推定する.

2. 状態空間モデルによる表現

いま, n $(=1,\ldots,N)$ 番目の受診者の t 年度における q 個の健診結果を要素として持つベクトルを $\boldsymbol{y}_t^{(n)} = (y_{t,1}^{(n)},\ldots,y_{t,q}^{(n)})'$, m 個の遺伝的形質もしくは生活習慣変数 ($e.g.$, 喫煙習慣) を要素として持つベクトルを $\boldsymbol{z}_t^{(n)} = (z_{t,1}^{(n)},\ldots,z_{t,m}^{(n)})'$, 健診結果 $\boldsymbol{y}_t^{(n)}$ を制御する p 個の状態変数を $\boldsymbol{x}_t^{(n)} = (x_{t,1}^{(n)},\ldots,x_{t,p}^{(n)})'$ とする. このとき, n 番目の受診者の次年度の検診結果 $\boldsymbol{y}_{t+1}^{(n)}$ が, 前年度の検診結果と生活習慣の影響によって，以下の線形ガウス型状態空間モデルによって表されると仮定する.

$$\boldsymbol{x}_t^{(n)} = (A+I)\boldsymbol{x}_{t-1} + G\boldsymbol{z}_{t-1}^{(n)} + \boldsymbol{v}_t, \tag{1}$$

$$\boldsymbol{y}_t^{(n)} = H\boldsymbol{x}_t^{(n)} + \boldsymbol{w}_t. \tag{2}$$

ここで，$A \in R^{p\times p}$ は状態変数同士の制御関係を与えるシステム行列，$I \in R^{p\times p}$ は単位行列，$G \in R^{p\times m}$ は遺伝的形質と生活習慣の状態変数に対する影響を与える行列，$H \in R^{p\times q}$ は状態変数と健診結果の関連性を与える観測行列，$\boldsymbol{v}_t \in R^p$ と $\boldsymbol{w}_t \in R^q$ はそれぞれシステム誤差と観測誤差とし，平均 $\mathbf{0}$ 分散共分散行列 Q と R のガウス分布に従うとする. ここで，N 人の受診者に関する，欠損を含む健診結果の時系列データが与えられたときに，それぞれの状態変数とパラメータ値を推定する問題を考える．また本報告では，観測ノイズが健診データ特有の分布を持つ場合に関してまで取り扱う.

3. 正則化付き線形状態空間モデルのパラメータ推定

線形ガウス型状態空間モデルにおける状態変数の条件付き確率分布は，Kalman filter を用いて効率的に計算可能であることが一般的に知られており，計算された状態変数の条件付き確率分布を EM アルゴリズムの枠組みで用いることで，観測データの対数尤度を局所的に最大化するようなパラメータ値を推定する方法もまた提案されている．本報告においては，式 (1) と (2) によって表される状態空間モデルにおいて，A と G もしくは H が疎行列であることを仮定し，状態変数とパラメータ値を推定する方法論を提案する．EM アルゴリズムを用いて正則化対数尤度を局所的に最大化し，疎なシステム行列 A を持つ線形ガウス型状態空間モデルのパラメータ値を推定すると同時に適切な正則化項の値を決定する方法論が遺伝子ネットワークの制御関係の推定問題に対して提案されており (Hasegawa *et al.*, 2014)，本報告ではその拡張と健診データへの応用例を紹介する予定である.

参考文献

T. Hasegawa, R. Yamaguchi, M. Nagasaki, S. Miyano and S. Imoto. Inference of gene regulatory networks incorporating multi-Source biological knowledge via a state space model with L1 regularization, *PLoS One*, vol.9(8), e105942, 2014.