スパース正則化に基づく関数データ判別による遺伝子データ解析

滋賀大学データサイエンス学部　松井秀俊

1 はじめに

近年の分析・測定機器の発展に伴い，取得されるデータの量・頻度は共に飛躍的に上昇している．その中でも，経時的に測定されたデータを取得し分析の対象とすることで，一点の数値ではなく時間変化に伴う連続的なデータの特徴を考慮に入れることができる．このようなデータに対しては，観測時点の不均一性などにより古典的な多変量解析手法を直接適用することが困難になる場合がしばしばある．Ramsay and Silverman (2005) によって確立された関数データ解析（functional data analysis）は，経時測定データを関数化処理し，関数化データ集合に基づく解析を行う手法で，これにより上記の問題点を解消でき，生命科学，システム工学，気象学といったさまざまな分野でその有用性が報告されている．関数データ解析では，従来の多変量解析手法を関数データの枠組みへ拡張した手法が数多く提唱されているが，その一つである関数ロジスティック回帰モデルを適用することで，関数データの判別が可能となる．

本研究では，関数データの判別問題におけるモデル選択問題について検討する．変数選択に対するアプローチとして注目を集めているスパース正則化は，推定の過程で L_1 ノルムを含む制約を課した最適化問題を解く方法で，理論，応用の両側面から幅広い研究が行われている (例えば， Hastie et al., 2015)．特に，ロジスティック回帰モデルに対してスパース正則化を適用することで，モデルの推定と同時に，判別に影響を与えている変数を選択できる．

本報告では，関数データに対する多群ロジスティック回帰モデルにスパース正則化を適用することで，パラメータの推定と，判別に寄与している変数の選択を同時に行う方法について述べる．特にここでは，制約として elastic net (Zou and Hastie, 2005) 型および sparse group lasso (Friedman et al., 2010) 型に基づく 2 種類の制約を紹介し，それぞれがもたらす効果について紹介する．モデルの推定については，正則化最尤法の枠組みで推定法を紹介し，それに伴う制約の形状を紹介する．そして，これらの手法を，遺伝子発現データの解析へ適用した結果について報告する．

2 関数ロジスティック回帰モデル

いま，n 個体の関数データとそのラベルの組 $\{(\boldsymbol{x}_i(t), g_i);\ i = 1, \ldots, n\}$ が与えられているとする．ここで，$\boldsymbol{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))^T$ は関数データとして与えられた説明変数，$g_i \in \{1, \ldots, L\}$ は $\boldsymbol{x}_i$ が属する群のラベルとする．ここで扱う判別モデルである関数ロジスティック回帰モデルは，次で与えられる．

$$\log\left\{\frac{\Pr(g_i = l|\boldsymbol{x}_i)}{\Pr(g_i = L|\boldsymbol{x}_i)}\right\} = \beta_{0l} + \sum_{j=1}^{p}\int x_{ij}(t)\beta_{jl}(t)dt. \tag{1}$$

ここで，β_{0l} は定数項，$\beta_{jl}(t)$ は係数関数である．関数データ $x_{ij}(t)$ および係数関数 $\beta_{jl}(t)$ は，次のように基底関数展開によって表されていると仮定する．

$$x_{ij}(t) = \sum_{m=1}^{M_j} w_{ijm}\phi_{jm}(t) = \boldsymbol{w}_{ij}^T\boldsymbol{\phi}_j(t), \quad \beta_{jl}(t) = \sum_{m=1}^{M_j} b_{jlm}\phi_{jm}(t) = \boldsymbol{b}_{jl}^T\boldsymbol{\phi}_j(t). \tag{2}$$

ここで，$\boldsymbol{\phi}_j(t) = (\phi_{j1}(t), \ldots, \phi_{jM_j}(t))^T$ は基底関数ベクトル，$\boldsymbol{w}_{ij} = (w_{ij1}, \ldots, w_{ijM_j})^T$ は関数化の段階で得られた係数ベクトル，$\boldsymbol{b}_{jl} = (b_{jl1}, \ldots, b_{jlM_j})^T$ は未知の係数パラメータベクトルとする．

(2) 式の仮定と，$\pi_l(\boldsymbol{x}_i; \boldsymbol{b}) = \Pr(g_i = l|\boldsymbol{x}_i)$，$\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_{(L-1)}^T)^T$，$\boldsymbol{b}_l = (\beta_{0l}, \boldsymbol{b}_{1l}^T, \ldots, \boldsymbol{b}_{pl}^T)^T$ という表記を用いると，モデル (1) は次のように書きかえることができる．

$$\log\left\{\frac{\pi_l(\boldsymbol{x}_i; \boldsymbol{b})}{\pi_L(\boldsymbol{x}_i; \boldsymbol{b})}\right\} = \beta_{0l} + \sum_{j=1}^{p} \boldsymbol{w}_{ij}^T\Phi_j\boldsymbol{b}_{jl} = \boldsymbol{z}_i^T\boldsymbol{b}_l. \tag{3}$$

ここで $\boldsymbol{z}_i = (1, \boldsymbol{w}_{i1}^T\Phi_1, \ldots, \boldsymbol{w}_{ip}^T\Phi_p)^T$，$\Phi_j = \int \boldsymbol{\phi}_j(t)\boldsymbol{\phi}_j^T(t)$ とした．これにより，関数ロジスティック回帰モデル (1) の推定問題は，モデル (3) に含まれるパラメータベクトル $\boldsymbol{b}$ を推定する問題に帰着される．ここで，モデル (3) より，j 番目の説明変数に関わる係数パラメータの数は $M_j(L-1)$ 個となることに注意されたい．また，群ラベルを表す目的変数 $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i(L-1)})^T$ を次のように定義する．

$$\boldsymbol{y}_i = \begin{cases} (0, \ldots, 0, \overset{(l)}{1}, 0, \ldots, 0)^T & \text{if } g_i = l, \quad l = 1, \ldots, L-1, \\ (0, \ldots, 0)^T & \text{if } g_i = L. \end{cases}$$

これより，上記モデルは次の確率関数をもつことがわかる．

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{b}) = \prod_{l=1}^{L-1} \pi_l(\boldsymbol{x}_i; \boldsymbol{b})^{y_{il}} \pi_L(\boldsymbol{x}_i; \boldsymbol{b})^{1-\sum_{h=1}^{L-1} y_{ih}}.$$

3 モデル推定

関数ロジスティック回帰モデル (3) に含まれる係数パラメータ $\boldsymbol{b}$ を正則化最尤法，すなわち次で定義される正則化対数尤度関数の最大化により推定する．

$$\ell_\lambda(\boldsymbol{b}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{b}) - nP_{\lambda,\alpha}(\boldsymbol{b}). \tag{4}$$

ただし $P_{\lambda,\alpha}(\boldsymbol{b})$ はパラメータに対する制約関数で，制約そのものの強さを規定する正則化パラメータ $\lambda > 0$ と，追加の調整パラメータ $\alpha \in [0, 1]$ である．これに L_1 ノルムを含む関数を仮定することで，いくつかのパラメータを 0 と推定できる．本報告では，次の二種類の制約と，これらがもたらす効果について紹介する．

3.1 Elastic net 型制約

モデル (3) を通して変数選択を行うにあたり，次で定義される elastic net 型の制約を導入する (Kayano et al., 2016).

$$P_{\lambda,\alpha}(\boldsymbol{b}) = \frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\lambda_j\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2^2 + n\alpha\sum_{j=1}^{p}\lambda_j\left\{\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2^2\right\}^{\frac{1}{2}}. \tag{5}$$

ただし $\lambda_j = \sqrt{M_j}\lambda$ とする．第 1 項は ridge 制約に，第 2 項は group lasso 制約 (Yuan and Lin, 2006) に対応している．2 節で述べたように，関数ロジスティック回帰モデルにおいては，1 つの説明変数に関わる係数パラメータの数が $M_j(L-1)$ 個となる．従って，スパース正則化を用いて変数選択を行う場合，これらのパラメータをまとめて 0 と縮小する必要があるため，lasso の代わりに group lasso が用いられる．また，Elastic net 型の制約を適用することで，多重共線性などによる推定量の不安定性を除去できる (Zou and Hastie, 2005).

3.2 Sparse group lasso 型制約

(4) 式の制約関数 $P_{\lambda,\alpha}(\boldsymbol{b})$ として，次の sparse group lasso 型制約を適用することを考える (Matsui, 2017).

$$P_{\lambda,\alpha}(\boldsymbol{b}) = n(1-\alpha)\sum_{j=1}^{p}\lambda_j\left\{\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2^2\right\}^{\frac{1}{2}} + n\alpha\sum_{j=1}^{p}\lambda_j\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2. \tag{6}$$

第 1 項は，制約 (5) にもあった group lasso 項で，変数選択の役割を持つ．一方で第 2 項は，各変数において l 群 ($l = 1, \ldots, L-1$) と L 群の判別に関わる係数をそれぞれグループ化したものである．この項を導入することで，各変数において，特定の群の判別への寄与の有無を選択できるようになる．すなわち，第 2 項は判別における決定境界を選択する役割を持つ．まとめると，(6) 式の制約は，変数選択および決定境界選択の 2 つの効果を持つ.

4 適用例

提案した手法を，経時測定された 2 種類の遺伝子発現データの解析へ適用する．適用結果の詳細については当日報告する.

一つは，多発性硬化症患者に対する遺伝子組み換え型インターフェロン β (rIFN-β) 治療開始後の，時間経過に伴う遺伝子発現量の変化を測定したデータである (Baranzini et al., 2004)．このデータは，治療により予後が良好であった群と良好でなかった群の 2 群からなる．そこで，予後良好，不良の判別に寄与している遺伝子の選択を，elastic net 型正則化に基づく関数ロジスティック回帰モデルを適用して行った結果を報告する．なお，この適用結果の詳細については Kayano et al. (2016) で述べられている.

もう一つは，イースト菌に含まれている約 800 個の細胞周期関連遺伝子に対して，cDNA マイクロアレイを用いて発現量の経時変化を計測し得られたデータである．Spellman et al. (1998) は，6 種類の実験により得られた経時測定データに基づき遺伝子を 5 つの細胞周期にクラスタリングした．本研究では，Spellman et al. (1998) によって行われた実験が，実際に細胞周期の分類に効果があったかを検証するために，提案手法を適用した.

5 まとめと今後の課題

本報告では，経時測定されたデータを関数データとして扱い，判別に寄与している変数を選択するための方法について紹介した．関数データに対するロジスティック回帰モデルを推定するためのスパース正則化法として，elastic net 型および sparse group lasso 型の制約を導出し，それぞれの制約の性質について紹介した．前者は判別に関わる変数を選択できる一方で，後者については，多群判別の場合，各変数がどの判別に寄与しているかを選択することができる．そして，提案手法を遺伝子データの解析に適用し，その有効性について議論した．

本報告で述べた手法では，用いたモデルやアルゴリズムについて検討の余地がある．特に近年，スパース正則化に対する推定アルゴリズムとして，より汎用的な方法が提案されているため，この方法を今回提案した方法の推定に取り入れ，モデル構築から評価までの一連のプロセスを確立させたい．

参考文献

Baranzini, S. E., Mousavi, P., Rio, J., Caillier, S. J., Stillman, A., Villoslada, P., Wyatt, M. M., Comabella, M., Greller, L. D., Somogyi, R., et al. (2004), "Transcription-based prediction of response to IFNβ using supervised computational methods," *PLoS Biol.*

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0376*.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalization*, Boca Raton: Chapman & Hall/CRC.

Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S., and Miyano, S. (2016), "Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to timedependent biomarker detection," *Biostatistics*, 17, 235–248.

Matsui, H. (2017), "Selection of variables and decision boundaries for functional data via bi-level selection," *arXiv preprint arXiv:1702.02010*.

Ramsay, J. and Silverman, B. (2005), *Functional data analysis 2nd ed.*, New York: Springer.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998), "Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell*, 9, 3273–3297.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, 68, 49–67.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B*, 67, 301–320.