一般化線形混合モデルにおけるモデル選択

吉田拓真[1]　　勘場 大[1]　　内藤 貫太[2]

[1] 島根大学大学院総合理工学研究科

[2] 島根大学総合理工学部　　E-mail: naito@riko.shimane-u.ac.jp

1　はじめに

一般化線形スプライン回帰において，含まれるパラメータの推定に正則化法を用いるとき，あらかじめ，ノットの個数と平滑化パラメータを決定しておく必要があり，一般化情報量規準 GIC(小西, 北川 (2004)) などが利用される．しかし，情報量規準による平滑化パラメータの決定はグリッドサーチによるため，多くの平滑化パラメータを含むモデルの場合，その計算量は膨大なものとなり，効率的であるとはいえない．

その一方で，一般化スプライン回帰は，スプラインの部分の係数をランダム効果として捉えることにより一般化線形混合モデルとなり，罰則付き擬似尤度 (PQL) から推定量，予測量を求めることができ，一般化スプライン回帰における平滑化パラメータが直接推定できる．本講演では，これらの議論を GIC に応用させることで，時間効率の良い新たな情報量規準を提案する．

2　問題設定

n 個のデータ $\{(y_i, x_{i1}, \cdots, x_{ip}) | i = 1, \cdots, n\}$ が観測されたとき，未知の回帰関数 $f(\boldsymbol{x}) = E[y|\boldsymbol{x}], (\boldsymbol{x} = [x_1 \ \cdots \ x_p]^T)$ の推定を考える．そこで，$\boldsymbol{x}_i = [x_{i1} \cdots x_{ip}]^T$ を与えたときの y_i が指数分布族の密度

$$f(y_i; \eta_i) = \exp\left(\frac{y_i \eta_i - b(\eta_i)}{\phi} + h(y_i, \phi) \right), \quad i = 1, \cdots, n$$

を持つと仮定する．ここで，$b'(\eta_i) = \mu_i = E[y_i|\boldsymbol{x}_i]$ であり，$\phi b''(\eta_i) = V[y_i|\boldsymbol{x}_i]$ である．$\eta(\cdot)$ を未知の回帰関数として，$\eta_i = \eta(\boldsymbol{x}_i)$ を扱い，

$$\eta(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{K_1} u_{1k}(x_{i1} - \kappa_{1k})_+ + \cdots + \sum_{k=1}^{K_p} u_{pk}(x_{ip} - \kappa_{pk})_+, \quad i = 1, \cdots, n$$

を想定する．ここで，$\kappa_{11}, \cdots, \kappa_{1K_1}, \cdots, \kappa_{p1}, \cdots, \kappa_{pK_p}$ はノット，$(x)_+ = \max\{x, 0\}$ である．また，$\boldsymbol{\eta} = [\eta(x_i) \ \cdots \ \eta(x_n)]^T$，$\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \cdots \ \beta_p]^T$，$\mathbf{u}_j = [u_{j1} \ \cdots \ u_{jK_j}]^T$，$\mathbf{u} = [\mathbf{u}_1^T \ \cdots \ \mathbf{u}_p^T]^T$，$\boldsymbol{z}_i = [(x_{i1} - \kappa_{11})_+ \ \cdots \ (x_{i1} - \kappa_{1K_1})_+ \ \cdots \ (x_{ip} - \kappa_{1p})_+ \ \cdots \ (x_{ip} - \kappa_{pK_p})_+]^T$，$\boldsymbol{Z} = [\boldsymbol{z}_1 \ \cdots \ \boldsymbol{z}_n]^T$，$\boldsymbol{X} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]^T$ を導入すると，$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{u}$ と表せ，$\boldsymbol{c}_i = [\boldsymbol{x}_i^T \ \boldsymbol{z}_i^T]^T$，$\boldsymbol{C} = [\boldsymbol{X} \ \boldsymbol{Z}] = [\boldsymbol{c}_1 \ \cdots \ \boldsymbol{c}_n]^T$ とおき，$\boldsymbol{w} = [\boldsymbol{\beta}^T \ \mathbf{u}^T]^T$ とすると $\boldsymbol{\eta} = \boldsymbol{C}\boldsymbol{w}$ と表せる．

このとき，正則化法による $\boldsymbol{\beta}$，$\mathbf{u}$ の推定量は，更新式

$$\begin{bmatrix} \boldsymbol{\beta}^{new} \\ \mathbf{u}^{new} \end{bmatrix} \leftarrow (\boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C} + n\phi\boldsymbol{\lambda})^{-1} \boldsymbol{C}^T \boldsymbol{W} \left\{ \boldsymbol{C} \begin{bmatrix} \boldsymbol{\beta}^{old} \\ \mathbf{u}^{old} \end{bmatrix} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \right\} \tag{1}$$

より得られ，収束値 $[\hat{\boldsymbol{\beta}} \ \hat{\mathbf{u}}]$ を用いて $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\mathbf{u}}$ となる．ただし，$\boldsymbol{\mu} = [\mu_1 \ \cdots \ \mu_n]^T$，$b'(\eta_i) = \mu_i = E[y_i|\boldsymbol{x}_i]$，$\boldsymbol{W} = \mathrm{diag}[b''(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{u})]$ であり，$\boldsymbol{\lambda} = \mathrm{diag}[0 \ \cdots \ 0 \ \lambda_1 \ \cdots \ \lambda_p]$ は平滑化パラメータである．

3　統計的モデル選択

正則化法によって推定されたスプラインモデル $f(y|\hat{\boldsymbol{w}},\hat{\phi})$ のノットの個数 $\boldsymbol{K}=\{K_1,\cdots,K_p\}$，平滑化パラメータ $\boldsymbol{\lambda}$ の選択は GIC を用いて，GIC = GIC$(\boldsymbol{K},\boldsymbol{\lambda})$ を最小とする $\boldsymbol{K}$，$\boldsymbol{\lambda}$ として得られる．しかし，GIC を最小にする $\lambda_1,\cdots,\lambda_p$ は，グリッドサーチによって決定するものであるため，時間コストの観点からは効率的であるとは言えない．そこで我々は GIC によるモデル選択において効率の良い平滑化パラメータの決定を以下に提案する．

4　一般化線形混合モデル

$\boldsymbol{y}=[y_1\ \cdots\ y_n]^T$ の密度における $\mathbf{u}$ を $\mathbf{u}\sim N(\mathbf{0},\boldsymbol{G})$ とすると，モデルは一般化線形混合モデルとなる．ただし，$\boldsymbol{G}=\mathrm{diag}[\sigma_1^2\mathbf{1}_{K_1}\ \cdots\ \sigma_p^2\mathbf{1}_{K_p}]$ である．そこで，PQL を用いて，$\boldsymbol{\beta}$ の推定量と $\mathbf{u}$ の予測量を求めると，PQL による $[\boldsymbol{\beta}^T\ \mathbf{u}^T]^T$ の更新式は

$$\begin{bmatrix}\boldsymbol{\beta}^{new}\\ \mathbf{u}^{new}\end{bmatrix}\ \leftarrow\ (\boldsymbol{C}^T\boldsymbol{W}\boldsymbol{C}+\phi\boldsymbol{B})^{-1}\boldsymbol{C}^T\boldsymbol{W}\left\{\boldsymbol{C}\begin{bmatrix}\boldsymbol{\beta}^{old}\\ \mathbf{u}^{old}\end{bmatrix}+\boldsymbol{W}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\} \tag{2}$$

で与えられる．ここで，$\boldsymbol{B}=\mathrm{diag}[0\ \cdots\ 0\ (1/\sigma_1^2)\mathbf{1}_{K_1}\ \cdots\ (1/\sigma_p^2)\mathbf{1}_{K_p}]$，$\boldsymbol{\mu}=b'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\mathbf{u})$，$\boldsymbol{W}=\mathrm{diag}[b''(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\mathbf{u})]$ である．(2) は，(1) において，$\boldsymbol{B}=n\boldsymbol{\lambda}$ とした式と一致している．したがって，$\boldsymbol{B}$ の中に含まれるパラメータを推定することができれば，平滑化パラメータ $\lambda_1,\cdots,\lambda_p$ の推定値は $\hat{\lambda}_i=1/(n\hat{\sigma}_i^2),(i=1,\cdots,p)$ として得られることになる．$\boldsymbol{B}$ の推定量は，擬似データ $\boldsymbol{y}_p=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\mathbf{u}+\boldsymbol{W}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\mathbf{u}+\boldsymbol{\varepsilon}_p$ を考えたとき，$\boldsymbol{y}_p\sim N(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{V})$，$\boldsymbol{V}=\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T+\phi\boldsymbol{W}$ を仮定し，最尤法により推定できる (Ruppert et al(2003))．そこで，GIC において，$\hat{\boldsymbol{\lambda}}=\hat{\boldsymbol{B}}/n$ としたとき，ノットの個数 $\boldsymbol{K}$ を GIC$_P$ = GIC$(\boldsymbol{K},\hat{\boldsymbol{B}}/n)$ を最小とするものとして選択する．

GIC$_P$ では平滑化パラメータを推定するので，第 3 節で扱った手法のようにグリッドサーチにより探索する必要が無い分，計算コストの観点から効率的であると言える．

5　比較

線形モデルや線形加法モデル，変化係数モデルに対して，実データへの当てはめ，予測精度，シミュレーションにおける GIC と GIC$_P$ の比較を行った結果，予測精度や曲線の形状はほとんど同じであったが，計算時間は GIC$_P$ の方がはるかに早いことがわかった．

6　まとめと展望

5 節より，GIC と GIC$_P$ は，共に良いモデル選択規準となることがわかったが，計算時間が早い分，GIC$_P$ は GIC の簡便版として有用である．また，セミパラメトリック回帰の枠組みの中で，空間モデル (Ruppert et al(2003)) にも GIC$_P$ を適用できるだろう．

参考文献

[1] Ruppert,D., Wand,M.P. and Carroll,R.J.(2003). *Semiparametric Regression*, Cambridge University Press.

[2] 小西貞則，北川源四郎 (2004)．情報量規準．朝倉書店．