Sélection de variables dans les modèles mixtes fonctionnels

Madison Giacofci
Equipe SAM

Sophie Lambert-Lacroix (TIMC - Grenoble)
Franck Picard (LBBE - Lyon)

Journée de rentrée des doctorants en statistique
Mardi 20 novembre 2012

Contexte

- On s'intéresse au problème de sélection de variables
- Considérons le modèle de régression pour $i = 1, \ldots, N$:

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_i \quad \text{où } \mathbf{E}_i \sim \mathcal{N}(0, \sigma_E^2 \mathbf{I})$$

  avec $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$ vecteur de variables explicatives

- But de la sélection de variables :
  - Avoir un modèle interprétable
  - Réduire la variance des estimateurs
  - Stabiliser les estimateurs

Régression pénalisée

- Dans les années 90, développement des techniques de régression pénalisée pour la sélection de variables
  - La plus populaire : Régression LASSO basée sur la norme ℓ_1 des paramètres [Tib96]

$$\widehat{\boldsymbol{\beta}} = \arg\min \|\bar{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

    avec λ paramètre de régularisation à déterminer.

  $\Longrightarrow$ Sélection de variables grâce à la présence d'une singularité en zéro.
- Dans un cadre de régression non paramétrique fonctionnelle

$$Y_i(t_m) = \mu(t_m) + E_i(t_m)$$

- En projetant sur une base d'ondelettes, le modèle sur les coefficients empiriques devient :

$$d_{ijk} = \beta_{jk} + \varepsilon_{ijk} \quad \text{où } \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

Seuillage

- Cadre fonctionnel : design orthogonal
  $\Rightarrow$ LASSO est équivalent à un seuillage doux des coefficients [DJ94] défini par :
  $$\widehat{\beta}_{jk} = \mathrm{sign}(d_{\bullet jk})\,(d_{\bullet jk} - \lambda)_+$$
- Associé avec le seuillage dit dur :
  $$\widehat{\beta}_{jk} = d_{\bullet jk}\mathbf{1}_{\{|d_{\bullet jk}|>\lambda\}}$$
- Dépendent d'un paramètre de régularisation λ
  $\Rightarrow$ Seuil universel : $\lambda = \widehat{\sigma}_\varepsilon\sqrt{2\log M}$ [DJ94]
- Seuillages near-minimax : vitesse de convergence optimale dans la classe des espaces de Besov à un facteur $\log M$ près.

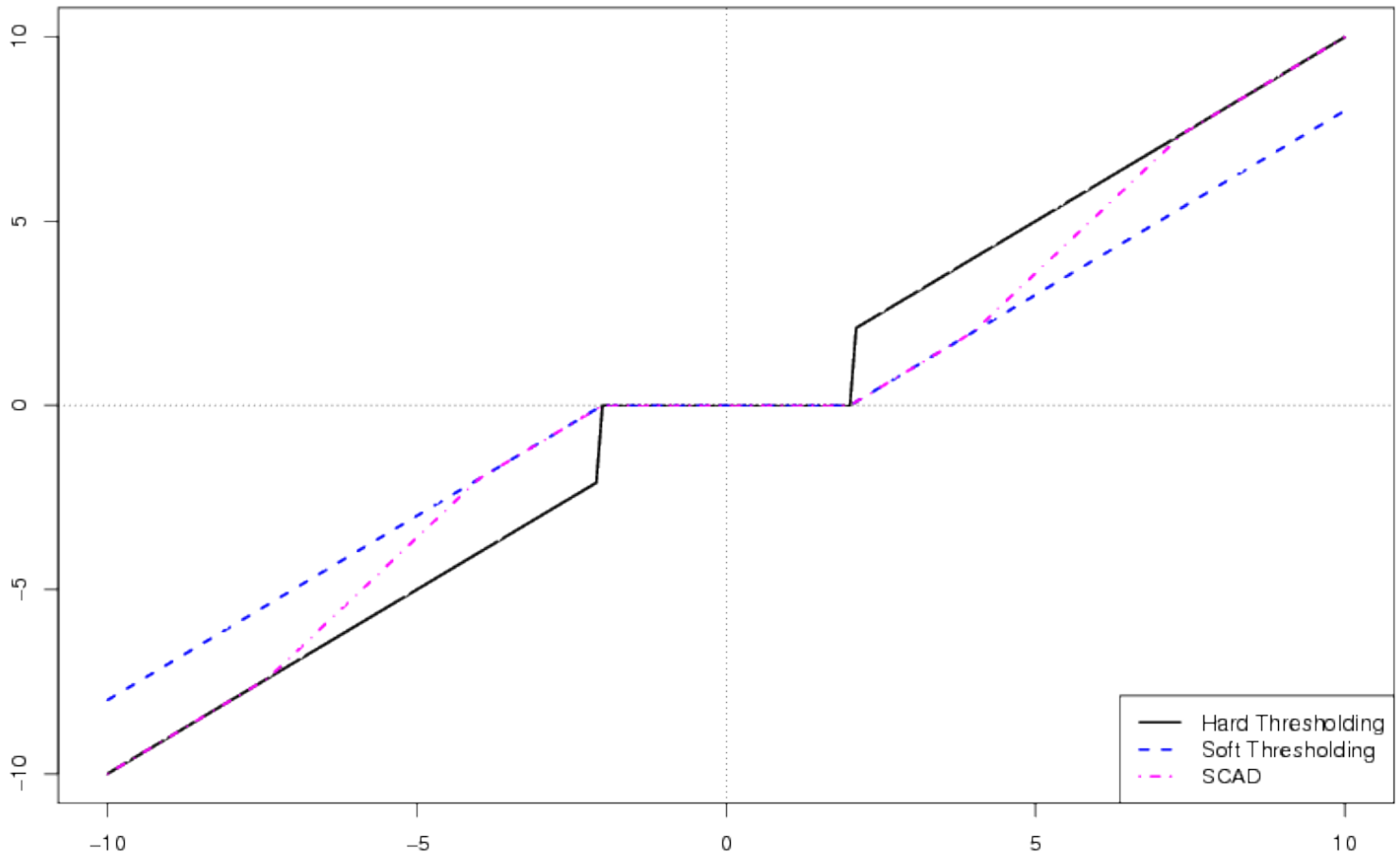
Seuillage SCAD

- Inconvénients : biais sur les grands coefficients (seuillage doux), instabilité des estimations (seuillage dur)
- Seuillage SCAD [FL01]
  $\Rightarrow$ Réalise un compromis entre les seuillages doux et dur

$$\widehat{\beta}_{jk} = \begin{cases} \text{sign}(d_{\bullet jk})\,(d_{\bullet jk} - \lambda)_+ & \text{si } |d_{\bullet jk}| \leq 2\lambda \\ \frac{1}{a-2}\left[(a-1)d_{\bullet jk} - a\lambda \text{sign}(d_{\bullet jk})\right] & \text{si } 2\lambda > |d_{\bullet jk}| \leq a\lambda \\ d_{\bullet jk} & \text{si } |d_{\bullet jk}| > a\lambda \end{cases}$$

avec a et λ paramètres de régularisation (habituellement, $a = 3.7$ par des arguments bayésiens).

Représentation des différents seuillages



Seuillage SCAD [2]

- Equivalent à un problème de régression pénalisée :

$$\text{pen}_{\text{SCAD}}(\beta_{jk};\lambda) = \lambda|\beta_{jk}|\mathbf{1}_{\{\beta_{jk}\leq\lambda\}} + \frac{(a+1)\lambda^2}{2}\mathbf{1}_{\{\beta_{jk}>a\lambda\}} - \frac{|\beta_{jk}^2| - 2a\lambda|\beta_{jk}^2| + \lambda^2}{2(a-1)}\mathbf{1}_{\{\lambda<\beta_{jk}\leq a\lambda\}}$$

- Utilisé avec le seuil universel
- Bonnes propriétés de convergence (vitesse near-minimax avec le seuil universel)
- Possède des propriétés oraculaires (sélection de variables et normalité asymptotique).

Modèle mixte fonctionnel

- Ajout d'effets individuels fonctionnels pour modéliser la variabilité inter-individuelle

$$Y_i(t_m) = \mu(t_m) + U_i(t_m) + E_i(t_m)$$

conduisant sur les coefficients d'ondelettes :

$$d_{ijk} = \beta_{jk} + \theta_{ijk} + \varepsilon_{ijk}$$

avec : $\begin{cases} \varepsilon_i & \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}) \\ \boldsymbol{\theta}_i & \sim \mathcal{N}(0, \mathbf{G}_{\boldsymbol{\theta}}) \end{cases}$ et $\mathbf{G}_{\boldsymbol{\theta}} = \text{Diag}(2^{-j\eta}\gamma_{jk}^2)$.

- Représentation parcimonieuse de l'effet fixe μ dans le domaine des ondelettes
  $\Rightarrow$ On peut penser qu'il en est de même pour $U_i(t)$

Vraisemblance pénalisée

- Proposition : Retrouver cette parcimonie par une sélection des effets fixes $\boldsymbol{\beta}$ et des variances des effets aléatoires $\boldsymbol{\gamma}$.
- Idée : basée sur les techniques de vraisemblance pénalisée en utilisant une pénalité de type SCAD.
- Problème d'optimisation

$$q(\lambda_1, \lambda_2) = \log \mathcal{L}(\mathbf{d}; \boldsymbol{\beta}, \mathbf{G}_{\boldsymbol{\theta}}, \sigma_\varepsilon^2) + \text{pen}_{\text{SCAD}}(\boldsymbol{\beta}; \lambda_1) + \text{pen}_{\text{SCAD}}(\boldsymbol{\gamma}; \lambda_2)$$

$\Rightarrow$ On cherche à minimiser ce critère

Propriétés asymptotiques

- Pour $\Upsilon = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ et sous certaines hypothèses concernant la vraisemblance et le terme de pénalité

Théorème

On suppose les hypothèses (H1)-(H7) vérifiées. De plus, on suppose que $\sqrt{\frac{N}{M}}\lambda \to \infty$ *si* $\lambda \to 0$ *et* $\frac{M^5}{N} \to 0$ *quand* $N \to \infty$. *Alors, avec une probabilité tendant vers 1, l'estimateur* $\sqrt{N/M}$*-consistant* $\widehat{\Upsilon} = \begin{bmatrix} \widehat{\Upsilon}^1 \\ \widehat{\Upsilon}^2 \end{bmatrix}$ *vérifie :*

i. *(Parcimonie)* $\quad \widehat{\Upsilon}_2 = 0$
ii. *(Normalité Asymptotique)*

$$\sqrt{N}\mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\left[\mathcal{I}(\Upsilon_0^1) + \mathbf{P}_N\right]\left[\widehat{\Upsilon}^1 - \Upsilon_0^1 + \left(\mathcal{I}(\Upsilon_0^1) + \mathbf{P}_N\right)^{-1}\mathbf{p}_N\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{H})$$

où $\mathbf{A}_N$ *est une matrice de taille* $q \times s_N$ *telle que* $\mathbf{A}_N\mathbf{A}_N^T \to \mathbf{H}$, *avec* $\mathbf{H}$ *matrice symétrique positive.*

Procédure d'estimation

- Présence de données non-observées (effets individuels)
  $\Longrightarrow$ Utilisation de l'algorithme EM
- Se déroule en 2 étapes
  - Etape E : Prédiction des effets aléatoires (calcul de l'espérance conditionnelle du critère $q(\lambda_1, \lambda_2)$)
  - Etape M : Maximisation selon les paramètres $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ et σ_ε^2 séparément (variante ECM)
- Conduit à des seuillages de type SCAD
  - $\boldsymbol{\beta} \rightsquigarrow$ Seuillage des données corrigées des prédictions des effets aléatoires
  - $\boldsymbol{\gamma} \rightsquigarrow$ Seuillage des données centrées et normalisées par les prédictions des effets aléatoires

Prochaines étapes

- Etude de simulation pour evaluer les performances :
  - en terme de sélection de variables sur les effets fixes et aléatoires
  - en terme de précision des estimateurs
  - question subsidiaire : quel est le gain par rapport au seuillage habituel avec estimation robuste de la variance ?
- Applications à des données réelles issues de la biologie moléculaire.

D.L. Donoho and I.M. Johnstone.
Ideal spatial adaptation by wavelet shrinkage.
Biometrika, 81(3) :425–455, 1994.

J. Fan and R. Li.
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American Statistical Association, 96(456) :1348–1360, 2001.

R. Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society Series B, 58(1) :267–288, 1996.