La méthode `shock` : réduction de dimension en inférence de réseaux

Mélina Gallopin[a] et **Emilie Devijver**[b]

[a]Laboratoire MAP5
Université Paris Descartes
45 rue des Saints Pères, 75006 Paris
melina.gallopin@parisdescartes.fr

[b]Department of Mathematics and Leuven Statistics Research Center
KU Leuven
Leuven, Belgium
emilie.devijver@wis.kuleuven.be

Mots clefs : Inférence de réseaux, modèle graphique gaussien, sélection de variables, sélection de modèle non-asymptotique, heuristique de pente

On considère une matrice $\mathbf{y}$ de taille $n \times p$ correspondant à n observations $\mathbf{y}_i$ pour $i = 1, \ldots, n$ et p variables $\mathbf{y}^j$ pour $j = 1, \ldots, p$. On suppose que les vecteurs $\mathbf{y}^1, \ldots, \mathbf{y}^p$ sont n réalisations issues de variables aléatoires $\mathbf{Y}^1, \ldots, \mathbf{Y}^p$. On souhaite inférer les dépendances entre les p variables aléatoires à partir des n observations, $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Pour représenter et visualiser ces dépendances, on considère le graphe $G = (V, E)$ où $V = \{1, \ldots, p\}$ est l'ensemble des nœuds représentant les variables aléatoires $\mathbf{Y}^1, \ldots, \mathbf{Y}^p$ et $E \subset V \times V$ est l'ensemble des arêtes du graphe représentant les dépendances entre les variables aléatoires. Dans le cadre du modèle graphique gaussien, chaque observation $\mathbf{y}_i$ est supposée issue d'une loi normale mutivariée de dimension p, de moyenne nulle $\mathbf{0}$ et de variance $\mathbf{\Sigma}$, matrice définie positive de taille $p \times p$:

$$\mathbf{y}_i \overset{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$$

Dans ce modèle, les dépendances conditionnelles entre deux variables $\mathbf{Y}^j$ et $\mathbf{Y}^{j'}$ conditionnellement aux autres variables sont directement liées aux coefficients de la matrice la matrice de covariance inverse, notée $\Theta = \mathbf{\Sigma}^{-1}$. Un coefficient nul $\theta_{jj'} = 0$ indique l'indépendance des variables j et j' conditionnellement à toutes les autres variables du jeu de données [1]. Inférer le graphe de dépendance entre les p variables revient donc à détecter les coefficients non nuls de la matrice Θ. On considère la log-vraisemblance pénalisée suivante, où λ est le paramètre de régularisation de la pénalité ℓ_1 imposée sur les éléments de la matrice Θ à estimer.

$$L_\lambda(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \text{trace}(\boldsymbol{S}\mathbf{\Theta}) - \lambda \,||\, \Theta \,||_1 \,.$$

Le problème d'estimation de la matrice à partir de cette vraisemblance pénalisée est connu sous le nom du lasso graphique ou *graphical lasso* [2].
En pratique, si le nombre d'échantillons disponibles est faible, la performance de la méthode d'inférence de réseaux est mauvaise [3]. Dans ce contexte, il convient donc de réduire le plus possible le nombre de variables à inclure dans le réseau afin de réduire le nombre de paramètres à estimer. Pour cela, nous faisons l'hypothèse que le réseau à inférer peut être approché par un réseau à structure modulaire, ce qui revient à chercher à approcher la distribution de $\mathbf{y}_i$ par $\mathcal{N}_p(\mathbf{0}, \Sigma_B)$ où Σ_B est une matrice de covariance à structure diagonale

par bloc :

$$\Sigma_B = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_K \end{pmatrix}$$

Pour sélectionner la meilleure structure diagonale par bloc, *i.e.* la meilleure partition des variables B, on utilise une procédure en deux étapes. L'exploration de l'ensemble des partitions possibles des p variables, notée $\mathcal{B}$, n'est pas possible. Dans un premier temps, nous obtenons une collection de partitions pertinentes, notée $\mathcal{B}^\Lambda$, correspondant aux différentes structures diagonales par bloc obtenues par seuillage de la matrice de covariance empirique en valeur absolue [4]. Puis nous sélectionnons la meilleure partition à l'aide du critère de sélection de modèle non-asymptotique suivant :

$$\hat{B} = \underset{B \in \mathcal{B}^\Lambda}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{f}_B(\mathbf{y}_i)) + \operatorname{pen}(B) \right\},$$
$$\operatorname{pen}(B) = \kappa D_B,$$

où κ est un coefficient à calibrer à partir des données par l'heuristique des pentes [5]. Deux méthodes existent pour calibrer ce coefficient, la régression robuste et le saut de dimension. Ces deux méthodes consistent à détecter le coefficient κ à l'aide d'une régression robuste réalisée entre la log-vraisemblance et la dimension du modèle pour les modèles complexes où à l'aide du saut de dimension détecté sur la fonction représentant la dimension du modèle en fonction du paramètre κ à calibrer [6].
Le package `shock` implémente cette méthode de réduction de dimension du problème d'inférence initial, et permet de décomposer un problème d'inférence ambitieux en plusieurs sous problèmes de moindre dimension.

Références

[1]Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. *Wiley Publishing*.
[2]Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432-441.
[3] Verzelen, N. (2012). Minimax risks for sparse regressions : Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6 :38-90.
[4]Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale Graphical Lasso. *Journal of Machine Learning Research*, 13 :781-794.
[5]Birge, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory Related Fields*, 138(1-2).
[6] Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics : overview and implementation. *Statistics and Computing*, 22(2) :455470.