Inférence de réseaux pour des données gaussiennes inflatées en zéros par double troncature

Clémence Karmann, Anne Gégout-Petit, Aurélie Muller-Gueudin

Inférence de réseaux pour des données gaussiennes inflatées en zéros par double troncature

Clémence Karmann [1,2], Anne Gégout Petit [1,2] & Aurélie Gueudin [1,2,3]

[1] *INRIA Nancy, équipe BIGS, 615 Rue du Jardin-Botanique, 54600 Villers-lès-Nancy ; clemence.karmann@inria.fr*

[2] *Université de Lorraine, Institut Elie Cartan de Lorraine, UMR 7502 ; anne.gegout-petit@univ-lorraine.fr*

[3] *aurelie.gueudin@univ-lorraine.fr*

Résumé. On s'intéresse à inférer la structure de dépendances conditionnelles dans le cas de données gaussiennes inflatées en zéros par double troncature (à droite et à gauche). On dispose d'un p-vecteur gaussien X observé au travers du vecteur tronqué $Y := X\mathbb{1}_{a\leq X\leq b}$. L'objectif est de retrouver la matrice de précision de X à partir d'observations de Y. Pour ce faire, on propose une procédure d'estimation qui consiste à estimer d'abord chacun des termes de la matrice de covariance par maximum de vraisemblance, puis la matrice de précision à l'aide du graphical Lasso. On donne un résultat théorique concernant la convergence de la matrice de précision estimée par cette méthode.

Mots-clés. Gaussiennes tronquées, inférence de graphe, zéro-inflaté, matrice de précision.

Abstract. We are interested in inferring the structure of conditional dependencies in the case of Gaussian data zero-inflated by double truncation. We have a Gaussian p-vector X observed through the truncated vector $Y := X\mathbb{1}_{a\leq X\leq b}$. The objective is to find the precision matrix of X from observations of Y. To do this, we propose an estimation procedure, that consists of first estimating each of the terms of the covariance matrix by maximum likelihood estimation, and then the precision matrix using the graphical Lasso. We give a theoretical result about the convergence of the estimated precision matrix.

Keywords. Truncated gaussian, graph inference, zero-inflated, precision matrix.

1 Modèle théorique

Soient $\mu = (\mu_j)_{j=1,\dots,p}$, $\Sigma^* = (\Sigma^*_{jk})_{1\leq j,k\leq p}$ et $X \sim \mathcal{N}_p(\mu, \Sigma^*)$. On considère le vecteur Y défini par $Y_j = \mathbb{1}_{a_j\leq X_j\leq b_j} X_j$ pour tout $j \in \{1,\dots,p\}$ où $a_j, b_j \in \mathbb{R}$ connus, tels que $a_j < b_j$. Quitte à estimer μ_j et Σ^*_{jj} grâce aux méthodes présentées dans la littérature des gaussiennes univariées "doublement" tronquées (voir par exemple, Cohen (1957)), on se restreint ici au cas où X est centré et réduit, c'est-à-dire $\mu_j = 0$ et $\Sigma^*_{jj} = 1$ pour tout $j \in \{1,\dots,p\}$.

Rappelons que le modèle gaussien est particulièrement approprié à l'inférence de graphe de dépendance conditionnelle. Ces dépendances conditionnelles sont en effet spécifiées par la matrice de précision $\Theta^* := (\Sigma^*)^{-1}$ du vecteur gaussien X, qui fournit facilement cette structure latente de graphe. Plus précisément, ce graphe contient une arête entre les variables X_j et X_k si :

$$\begin{aligned} X_j \longleftrightarrow X_k &\iff X_j \not\perp\!\!\!\perp X_k \mid (X_l)_{l\neq j,k} \\ &\iff \mathrm{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) \neq 0 \\ &\iff \Theta^*_{jk} \neq 0. \end{aligned}$$

L'objectif est de retrouver la structure latente de graphe, donc la structure de dépendances conditionnelles entre les variables du vecteur X donnée par Θ^*, à partir d'observations du vecteur Y.

Ce modèle peut par exemple être utilisé dans le cadre de modélisation d'interactions de populations microbactériennes. La troncature à gauche modélise des phénomènes liés aux méthodes de réplications et la troncature à droite est en fait une hypothèse (non restrictive) introduite pour obtenir des résultats théoriques.

2 Procédure d'estimation

2.1 Première étape : estimation de la matrice de covariance

On souhaite dans un premier temps estimer la matrice de covariance Σ^* du vecteur X à partir d'observations du vecteur Y.

Estimer cette matrice de covariance de façon empirique à partir des observations $\mathbf{Y}$ du vecteur Y conduirait à des résultats médiocres. Une autre idée pourrait être de l'estimer par maximisation de la vraisemblance du vecteur Y. Or, cette vraisemblance est très difficile à écrire. On voit que la vraisemblance d'un couple de variables du vecteur Y, donnée en (1), est une somme de quatre termes permettant de traiter les quatre cas possibles (selon la nullité de chacune des variables). Ainsi, la vraisemblance du vecteur Y se découperait en 2^p termes correspondant à des intégrales multiples de la densité du vecteur gaussien X en dehors des points de troncature.

Au regard de ces difficultés, on propose d'estimer cette matrice de covariance terme à terme en se ramenant à l'étude des couples (Y_j, Y_k), $j < k$ de variables du vecteur Y. Soit $(j,k) \in \{1,\dots,p\}^2$, $j < k$. On note $f_{jk}(x,y)$ la densité du couple gaussien

$(X_j, X_k) \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \Sigma^*_{jk} \\ \Sigma^*_{jk} & 1 \end{pmatrix} \right)$. On a $f_{jk}(x,y) = f(x,y,\Sigma^*_{jk})$ où :

$$f(x,y,\sigma) = \frac{1}{2\pi\sqrt{1-\sigma^2}} \exp\left[-\frac{x^2 - 2\sigma xy + y^2}{2(1-\sigma^2)} \right].$$

La vraisemblance du couple (Y_j, Y_k) est $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ où y correspond à la réalisation du vecteur Y et :

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k), \tag{1}$$

avec :

— $n_{ab}(y_j, y_k) = \mathbb{1}_{\zeta_j = a, \zeta_k = b}$ où $\zeta_l = \begin{cases} 1 \text{ si } y_l \in [a_l, b_l] \setminus \{0\}, \\ 0 \text{ si } y_l = 0. \end{cases}$

— $\displaystyle\sum_{a,b=0}^{1} n_{ab}(y_j, y_k) = 1$

— $\phi_{11,jk}(\sigma, y_j, y_k) = f(y_j, y_k, \sigma)$

— $\phi_{01,jk}(\sigma, y_j, y_k) = \phi_{01,jk}(\sigma, y_k) = \displaystyle\int_{[a_j, b_j]^c} f(x, y_k, \sigma) \mathrm{d}x$

— $\phi_{10,jk}(\sigma, y_j, y_k) = \phi_{10,jk}(\sigma, y_j) = \displaystyle\int_{[a_k, b_k]^c} f(y_j, y, \sigma) \mathrm{d}y$

— $\phi_{00,jk}(\sigma, y_j, y_k) = \phi_{00,jk}(\sigma) = \displaystyle\iint_{[a_j, b_j]^c \times [a_k, b_k]^c} f(x, y, \sigma) \mathrm{d}x \mathrm{d}y.$

On dispose d'un n-échantillon $\mathbf{Y} := (Y^{(1)}, \ldots, Y^{(n)})$ du vecteur Y. La log-vraisemblance vaut alors $L^{(n)}_{jk}(\Sigma^*_{jk}, \mathbf{y})$ où :

$$\begin{aligned} L^{(n)}_{jk}(\sigma, \mathbf{y}) &= \sum_{i=1}^{n} \sum_{a,b=0}^{1} n_{ab}(y_j^{(i)}, y_k^{(i)}) \log\left(\phi_{ab,jk}(\sigma, y_j^{(i)}, y_k^{(i)}) \right) \\ &= \sum_{\substack{i=1 \\ i : y_j^{(i)} = y_k^{(i)} = 0}}^{n} \log\left(\phi_{00,jk}(\sigma) \right) + \sum_{\substack{i=1 \\ i : y_j^{(i)} = 0, y_k^{(i)} \neq 0}}^{n} \log\left(\phi_{01,jk}(\sigma, y_k^{(i)}) \right) \\ &+ \sum_{\substack{i=1 \\ i : y_j^{(i)} \neq 0, y_k^{(i)} = 0}}^{n} \log\left(\phi_{10,jk}(\sigma, y_j^{(i)}) \right) + \sum_{\substack{i=1 \\ i : y_j^{(i)} \neq 0, y_k^{(i)} \neq 0}}^{n} \log\left(\phi_{11,jk}(\sigma, y_j^{(i)}, y_k^{(i)}) \right), \end{aligned}$$

où $\mathbf{y} := (y^{(1)}, \ldots, y^{(n)})$ correspond à la réalisation du n-échantillon $\mathbf{Y}$.

Finalement, on estime Σ^* par $\widetilde{\Sigma}^{(n)}$ en estimant chacun de ses coefficients Σ^*_{jk} par maximisation de la log-vraisemblance de l'échantillon du couple $(Y_j^{(i)}, Y_k^{(i)})_{i=1,\dots,n}$:

$$\widetilde{\Sigma}^{(n)}_{jk} = \underset{|\sigma|\leq 1}{\operatorname{argmax}}\ L^{(n)}_{jk}(\sigma, \mathbf{y}) \tag{2}$$

2.2 Deuxième étape : estimation de la matrice de précision

Notre objectif étant de retrouver le graphe sous-jacent, on va ensuite utiliser l'estimateur de la matrice de précision Θ^* donné par le graphical Lasso introduit par Friedman et al. (2008). Le graphical Lasso est une procédure utilisée dans le modèle graphique gaussien qui consiste à estimer la matrice de précision en maximisant la log-vraisemblance pénalisée du modèle gaussien sur l'ensemble des matrices définies positives de dimension $p \times p$:

$$\underset{\Theta \succ 0}{\operatorname{argmax}}\ \log\det(\Theta) - \operatorname{trace}(\Theta S) - \lambda_n ||\Theta||_{1,\text{off}},$$

où $||\Theta||_{1,\text{off}} = \sum_{\substack{j,k=1 \\ j\neq k}}^{p} |\Theta_{jk}|$, S est la matrice de covariance empirique des observations du vecteur gaussien X et $\lambda_n > 0$ est le paramètre de la pénalisation Lasso.

Dans notre cas, on ne peut pas obtenir la matrice de covariance empirique car les observations du vecteur X sont inaccessibles. Au lieu de calculer la matrice de covariance empirique des X comme la matrice de covariance empirique des observations du vecteur Y, on remplace cette matrice de covariance S par l'estimateur $\widetilde{\Sigma}^{(n)}$ de Σ^* obtenu à l'étape précédente. $\widehat{\Theta}^{(n)}$ est donc défini comme l'unique solution du problème d'optimisation convexe suivant :

$$\widehat{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}}\ \log\det(\Theta) - \operatorname{trace}(\Theta\widetilde{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\text{off}}. \tag{3}$$

3 Résultats théoriques

L'objectif de cette partie est d'étudier les propriétés de $\widehat{\Theta}^{(n)}$. On énonce au préalable trois hypothèses :

(H1) Pour tout $j < k$, $|\Sigma^*_{jk}| \neq 1$. Ainsi, il existe $\delta > 0$ tel que pour tout $j < k$, $|\Sigma^*_{jk}| < 1 - \delta$.

(H2) Soit $j < k$. On considère l'application $g : \sigma \in [-1+\delta, 1-\delta] \mapsto \mathbb{E}\Big(L^{(n)}_{jk}(\sigma, \mathbf{y})\Big)$. Alors :

- $-1+\delta$ et $1-\delta$ ne sont pas des points critiques de g,

- g admet un nombre fini de points critiques,
- tous les points critiques de g, différents de Σ^*_{jk}, sont non-dégénérés, c'est-à-dire :

$$\text{pour tout } \sigma \neq \Sigma^*_{jk},\ g'(\sigma) = 0 \Rightarrow g''(\sigma) \neq 0.$$

Notons que Σ^*_{jk} est un point critique non-dégénéré de g.

(H3) Cette hypothèse est technique et ne sera pas énoncée par souci de simplicité. L'intuition sous-jacente est de limiter l'influence des termes relatifs à des "non-arêtes" sur les termes relatifs à des arêtes.

Voici un résultat intermédiaire concernant la matrice de covariance estimée $\widetilde{\Sigma}^{(n)}$, qui s'appuie sur des résultats récents de Mei et al. (2017).

Proposition 3.1. *On suppose les hypothèses **(H1)** et **(H2)**. Soit $0 < \rho < 1$. Il existe des constantes B, C et D connues telles que si n satisfait $\dfrac{n}{\log n} \geq C \log\left(\dfrac{B}{\rho}\right)$, alors la matrice de covariance estimée $\widetilde{\Sigma}^{(n)}$ définie par (2) vérifie :*

$$\mathbb{P}\left(\left|\left|\widetilde{\Sigma}^{(n)} - \Sigma^*\right|\right|_\infty \geq D\sqrt{\frac{\log n}{n} \log\left(\frac{B}{\rho}\right)} \right) \leq \rho \frac{p(p-1)}{2},$$

où $||A||_\infty = \max\limits_{j,k \in \{1,\dots,p\}} |A_{jk}|$ est la norme infinie de la matrice A vue comme un élément de $\mathbb{R}^{p^2}$.

Avant d'énoncer le résultat sur la convergence de la matrice de précision estimée $\widehat{\Theta}^{(n)}$, introduisons encore quelques notations :

— si $M \in \mathcal{M}_{r,m}(\mathbb{R})$, $A \subset [\![1, r]\!]$ et $B \subset [\![1, m]\!]$, M_{AB} désigne la matrice extraite $(m_{ij})_{i \in A, j \in B}$,

— $S = S(\Theta^*) := E(\Theta^*) \cup \{(1,1), \dots, (p,p)\}$ où $\Theta^* = (\Sigma^*)^{-1}$ et $E(\Theta^*) = \{(j,k) \in \{1,\dots,p\}^2,\ j \neq k,\ \Theta^*_{jk} \neq 0\}$,

— $\Gamma^* = \Sigma^* \otimes \Sigma^*$ où $\otimes$ désigne le produit de Kronecker. On a : $\Gamma^*_{(j,k),(l,m)} = \mathrm{cov}(X_j X_k, X_l X_m)$,

— $\kappa_{\Sigma^*} := |||\Sigma^*|||_\infty = \max\limits_{j=1,\dots,p} \sum\limits_{k=1}^{p} |\Sigma^*_{jk}|$,

— $\kappa_{\Gamma^*} := \left|\left|\left| \left(\Gamma^*_{SS}\right)^{-1} \right|\right|\right|_\infty$.

Théorème 3.1. *On suppose **(H1)**, **(H2)** et **(H3)**. Soit $c > 2$, $\widehat{\Theta}^{(n)}$ l'unique solution de (3) et d le degré maximal du graphe défini par :*

$$d = \max_{j=1,\dots,p} \Big| \{k \in [\![1, p]\!] : \Theta^*_{jk} \neq 0\} \Big|.$$

Il existe des constantes B, C et D connues telles que pour n vérifiant

$$\frac{n}{\log n} > D^2 \log\left(Bp^c\right) \max\left\{\frac{\sqrt{C}}{D}, 6(1+8\alpha^{-1})d \max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa_{\Sigma^*}^3\kappa_{\Gamma^*}^2\}\right\}^2$$

et $\lambda_n = \dfrac{8D}{\alpha}\sqrt{\dfrac{\log n}{n}\log\left(Bp^c\right)}$ *le paramètre de pénalisation de l'équation Lasso* (3)*, on a, avec probabilité* $1 - \dfrac{1}{p^{c-2}}$ *:*

(a) L'estimateur $\widehat{\Theta}^{(n)}$ *de* Θ^* *satisfait :*

$$||\widehat{\Theta}^{(n)} - \Theta^*||_\infty \leq 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}.$$

(b) $E(\widehat{\Theta}^{(n)}) \subset E(\Theta^*)$ *et l'arête* (j,k) *est correctement retrouvée dès que :*

$$\left|\Theta^*_{jk}\right| > 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}.$$

Le paramètre c du Théorème 3.1 est en fait un paramètre à définir par l'utilisateur. Plus c est grand, plus la probabilité pour laquelle les deux résultats du théorème tiennent sera grande. En revanche, de grandes valeurs de ce paramètre c conduisent également à de plus fortes exigences sur la taille n de l'échantillon.

La preuve de ce résultat utilise les résultats de Ravikumar et al. (2011).

Bibliographie

Cohen, A. Clifford (1957). On the solution of estimating equations for truncated and censored samples from normal populations, *Biometrika*, 44, pp. 225–2236.

Mei, S., Bai, Y. and Montanari, A. (2017). The landscape of empirical risk for non-convex losses, *https ://arxiv.org/abs/1607.06534*.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence, *Electronic Journal of Statistics*, 5, pp. 935–980.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, 9, pp. 432–441.