« Consensus LASSO » : Inférence conjointe de réseaux de gènes dans des conditions expérimentales multiples

Nathalie Villa-Vialaneix [1] & Magali SanCristobal [2,3]

[1] *Unité BIA, INRA Toulouse, Castanet Tolosan, France -* nathalie.villa@toulouse.inra.fr

[2] *UMR444 Laboratoire de Génétique Cellulaire, INRA Toulouse, 31326 Castanet Tolosan, France -* magali.san-cristobal@toulouse.inra.fr

[3] *Département de Génie Mathématiques et Modélisation, INSA Toulouse, 31077 Toulouse, France*

Résumé. Nous présentons ici une méthode pour l'inférence de réseaux de co-expression génique à partir de données d'expression obtenues dans des conditions expérimentales différentes. Cette approche est basée sur une double pénalité, permettant d'une part l'obtention d'une solution parcimonieuse et, d'autre part, la proximité entre les divers réseaux inférés dans les diverses conditions et un réseau consensus commun à toutes les conditions. [1]

Mots-clés. graphe, biostatistique, inférence de réseau, modèle graphique Gaussien

Abstract. In this paper, a novel method is introduced for inferring co-expression networks from samples obtained in different experimental conditions. This approach is based on a double penalization : a first penalty aims at inferring a sparse solution ; then, the second part is used to make the networks obtained in different conditions consistent with a consensual network used to represent the dependency structure between genes, whatever the condition.

Keywords. graph, biostatistics, network inference, Gaussian graphical model ...

1 Introduction

Le développement rapide des techniques d'acquisition haut débit a conduit à la production de jeux de données transcriptomiques dans lesquels l'expression de plusieurs milliers de gènes est mesurée simultanément sur un nombre relativement restreint d'individus (en général, moins de cent). Comprendre les relations complexes qui existent entre ces gènes est devenu une question d'un intérêt majeur en biologie des systèmes. Une approche commune pour aborder ce problème est d'utiliser un modèle graphique et d'inférer un *réseau de co-expression génique* à partir des données transcriptomiques. Un grand réseau est

1. Ce travail a été réalisé en partie grâce à l'appui du European Eadgene_S network.

alors défini, dans lequel les sommets représentent les gènes et les arêtes une relation ou une dépendance donnée entre deux gènes. Un grand nombre de méthodes assez différentes ont été développées pour la construction du réseau, utilisation des corrélations [1], d'un modèle graphique Gaussien [4], réseaux bayésiens [9, 10]...

Il est fréquent, en biologie des systèmes, que les données transcriptomiques soient collectées dans des conditions expérimentales variées afin d'essayer de comprendre l'impact de la condition d'intérêt sur le fonctionnement de l'organisme : (par exemple, [11] étudie l'expression de gènes chez des patients obèses avant et après un régime). Couramment, des *gènes différentiellement exprimés* sont recherchés, c'est-à-dire, des gènes dont l'expression est significativement différente selon la condition expérimentale. Une question plus difficile est alors de comprendre, non seulement comment la condition expérimentale d'étude influence l'expression individuelle de chacun des gènes, mais aussi comment elle impacte la manière dont les gènes interagissent entre eux. Il s'agit de comprendre quelles paires de gènes sont liés (co-exprimés ou régulés l'un par l'autre), *indépendamment* de la condition et lesquels sont liés pour une condition particulière, tout cela, sous l'hypothèse réaliste qu'un fonctionnement commun existe dans l'organisme, quelle que soit la condition.

Une approche naïve pour aborder cette question consiste à inférer un réseau différent dans chacune des conditions et ensuite de comparer les deux réseaux mais cette méthodologie n'exploite que partiellement l'information disponible, faisant fi, notamment, de la similarité des processus sous-jacents dans les différentes conditions. Particulièrement dans le cas où le nombre d'échantillons disponibles est faible, ce qui est fréquent en biologie, une telle stratégie peut s'avérer assez inefficace. Plusieurs stratégies alternatives ont donc été développées pour tenir compte de la similarité entre les diverses conditions expérimentales : [2], [7] et [3], dans le cadre du modèle graphique Gaussien avec des pénalités parcimonieuses ou bien [6] en procédant en deux temps avec une classification non supervisée suivie d'un modèle d'inférence. Notre proposition se rapproche de celles de [2, 7, 3], en se plaçant dans le cadre du modèle graphique Gaussien : une pénalisation à deux composantes est introduite permettant à la fois l'inférence d'un réseau par condition et assurant la proximité entre les réseaux correspondant aux différentes conditions. Nous présentons la méthode plus en détails dans la section suivante (section 2) et l'illustrons sur des exemples simples dans la section 3.

2 Consensus LASSO

2.1 Notations et rappels sur le modèle graphique Gaussien

Supposons que l'expression de p gènes est étudiée dans k conditions différentes $(X_j^c)_{j=1,\dots,p,\, c=1,\dots,k}$ et que les variables $(X_j^c)_j$ sont observées indépendamment sur n_c individus : $(X_{ij}^1)_{i=1,\dots,n_1,\, j=1,\dots,p} \dots (X_{ij}^k)_{i=1,\dots,n_k,\, j=1,\dots,p}$.

Dans le cadre du modèle graphique Gaussien, on suppose que $X^c = (X_1^c, \dots, X_p^c)$ suit une loi normale $\mathcal{N}(0, \Sigma^c)$ et on tente d'estimer les matrices de concentration

$\mathbf{K}^c = (\Sigma^c)^{-1}$ dont les éléments sont en relation avec les corrélations partielles $s^c_{jj'} = \mathbb{C}\text{or}\left(X^c_j, X^c_{j'} | (X^c_l)_{l\neq j,j'}\right)$ par

$$s^c_{jj'} = -\frac{\mathbf{K}^c_{jj'}}{\sqrt{\mathbf{K}^c_{jj}\mathbf{K}^c_{j'j'}}}.$$

Un graphe sous-jacent peut alors être défini pour modéliser la structure de corrélation partielle entre les gènes, $\mathcal{G}^c = (V^c, E^c)$, faisant correspondre l'ensemble des arêtes du graphe avec l'ensemble des paires de sommets dont la corrélation partielle est non nulle :

$$V^c = \{1, \ldots, p\} \qquad \text{et} \qquad \left\{(j,j') \in E^c \Leftrightarrow j \neq j' \text{ and } s^c_{jj'} \neq 0)\right\}.$$

L'estimation des coefficients $(s^c_{jj'})_{j,j',c}$ à partir des observations est effectuée en considérant $k \times p$ modèles linéaires :

$$X^c_j = \mathbf{X}^c_{\backslash j}\beta^c_j + \epsilon^c_j$$

où on peut montrer que $\beta^c_{jj'} = -\frac{\mathbf{K}^c_{jj'}}{\mathbf{K}^c_{jj}}$. Ces régressions linéaires sont résolues simultanément en considérant l'optimisation de la pseudo-vraisemblance

$$\mathcal{L}(\mathbf{K}|\mathbf{X}) = \sum_{c=1}^{k}\left[\frac{n_c}{2}\log\det\mathbf{D}^c - \frac{n_c}{2}\text{Tr}\left((\mathbf{D}^c)^{-1/2}\mathbf{K}^c\widehat{\Sigma}^c\mathbf{K}^c(\mathbf{D}^c)^{-1/2}\right)\right] - \frac{np}{2}\log(2\pi)$$

où $n = \sum_{c=1}^{k} n_c$, $\widehat{\Sigma}^c$ est la matrice de variance empirique $n^{-1}(\mathbf{X}^c)^T(\mathbf{X}^c)$ et $\mathbf{D}^c = \mathbb{D}\text{iag}(\mathbf{K}^c_{11}, \ldots, \mathbf{K}^c_{pp})$. Afin d'assurer la parcimonie de la solution, une pénalisation par la norme L^1 de la matrice $\mathbf{K}^c$, est introduite par de nombreux auteurs : dans le cas $k = 1$, cette méthode est connue sous le nom de « Graphical LASSO » [5]. Dans le cas où k est arbitraire, [2] montre alors que la maximisation de la vraisemblance pénalisée est équivalente à résoudre les p problèmes de minimisation quadratiques, $\forall\, j = 1, \ldots, p$,

$$\frac{1}{2}\beta_j^T\widehat{\Sigma}_{\backslash j\backslash j}\beta_j + \beta_j^T\widehat{\Sigma}_{j\backslash j} + \lambda\sum_{c=1}^{k}\frac{1}{n_c}\|\beta^c_j\|_1 \tag{1}$$

avec $\beta^c_j = (\beta^c_{jk})_{k\neq j} \in \mathbb{R}^{p-1}$ et $\beta^c_{jk} = (\mathbf{K}^c)^{-1}_{jj}\mathbf{K}^c_{jk}$, $\beta_j = \left(\beta^1_j, \ldots, \beta^k_j\right)^T$, $\widehat{\Sigma}^c_{\backslash j\backslash j}$ est la matrice $\widehat{\Sigma}^c$ privée de la ligne et de la colonne j, $\widehat{\Sigma}_{\backslash j\backslash j}$ est la matrice diagonale par blocs $\widehat{\Sigma}_{\backslash j\backslash j} = \mathbb{D}\text{iag}\left(\widehat{\Sigma}^1_{\backslash j\backslash j}, \ldots, \widehat{\Sigma}^k_{\backslash j\backslash j}\right)$, $\widehat{\Sigma}^c_{j\backslash j}$ est la j-ème ligne de $\widehat{\Sigma}^c$ sans son j-ème élément et $\widehat{\Sigma}_{j\backslash j}$ est le vecteur $\left(\widehat{\Sigma}^1_{j\backslash j}, \ldots, \widehat{\Sigma}^k_{j\backslash j}\right)$.

2.2 Estimations jointes

L'optimisation de la pseudo-vraisemblance $\mathcal{L}(\mathbf{K}|\mathbf{X})$ est équivalente à l'optimisation séparée de k fonctions de vraisemblance correspondant chacune à une condition. Dans

le cadre du graphical LASSO, afin d'assurer une plus grande cohérence entre les divers réseaux inférés, divers auteurs ont proposé des solutions distinctes : remplacer $\widehat{\Sigma}^c$ par une matrice contenant un mélange des matrices de variances des diverses conditions [2] ou bien pénaliser $\mathcal{L}(\mathbf{K}|\mathbf{X}) - \lambda\|\mathbf{K}^c\|_1$ par une deuxième pénalité $P((\mathbf{K}^c)_c)$ qui assure la ressemblance entre les réseaux obtenus dans différentes conditions. Différentes pénalités ont été proposées :

- [2] propose deux types de pénalités, $P((\mathbf{K}^c)_c) = \sum_{ij}\sqrt{\sum_c (\mathbf{K}^c_{ij})^2}$ (Group-LASSO) et $P((\mathbf{K}^c)_c) = \sum_{ij}\left[\sqrt{\sum_c (\mathbf{K}^c_{ij})^2_+} + \sqrt{\sum_c (\mathbf{K}^c_{ij})^2_-}\right]$ (sign-coherent Group-LASSO). Ces pénalités vont assurer, respectivement, que les réseaux inférés pour chacune des conditions sont identiques ou bien ne diffèrent que par des arêtes de signes distincts mais en faible nombre ;
- [3] propose l'utilisation de la pénalité $P((\mathbf{K}^c)_c) = \sum_{c\neq c'}\|K^c - K^{c'}\|_1$ qui force les corrélations partielles des différentes conditions à être identiques pour la plupart d'entre elles ;
- [7] introduit la pénalité $P((\mathbf{K}^c)_c) = \sum_{c\neq c'}\sum_j \|K^c_j - K^{c'}_j\|^2$ qui est une autre forme de pénalité de type « Group-LASSO » où les arêtes différentes entre les différentes conditions sont forcées sur un nombre limité de sommets.

Nous proposons ici une approche similaire, en pénalisant les problèmes d'optimisation de l'équation (1) par une pénalité de type L^2 entre chacun des coefficients $(\beta^c_j)_c$ et une solution « consensus », β^{cons}_j, représentative d'une certaine solution « globale », inter-conditions qui dépend de l'ensemble des $(\beta^c_j)_c$. De manière plus précise, les p problèmes d'optimisation suivants sont résolus :

$$\frac{1}{2}\beta_j^T\widehat{\Sigma}_{\backslash j\backslash j}\beta_j + \beta_j^T\widehat{\Sigma}_{j\backslash j} + \lambda\sum_{c=1}^{k}\frac{1}{n_c}\|\beta^c_j\|_1 + \mu\sum_{c=1}^{k} w_c\|\beta^c_j - \beta^{\text{cons}}_j\|_2^2 \tag{2}$$

où w_c est un poids accordé à chacune des conditions (des choix naturels étant $w_c = 1/k$ ou bien $w_c = \frac{n_c}{n}$). Selon les objectifs visés, plusieurs choix peuvent être effectués pour la valeur de β^{cons}_j, dont par exemple :

- $\beta^{\text{cons}}_j = \beta^{c*}_j$ avec $c* = \arg\min_c |\beta^c_j|$: dans ce cas, le réseau consensus induit est l'intersection des réseaux obtenus dans les diverses conditions ;
- $\beta^{\text{cons}}_j = \sum_{c=1}^{k}\frac{n_c}{n}\beta^c_j$: dans ce cas, le réseau consensus induit est en général l'union des réseaux obtenus dans les diverses conditions. Ce choix présente l'avantage d'être dérivable selon les $(\beta^c_j)_c$ et de permettre une résolution simple de l'équation (2) comme le montre la proposition suivante :

Proposition 1 *Pour $\beta^{cons}_j = \sum_{c=1}^{k}\frac{n_c}{n}\beta^c_j$, les problèmes d'optimisation de l'équation (2) peuvent s'écrire sous la forme d'un problème d'optimisation quadratique standard avec pénalisation parcimonieuse :*

$$\frac{1}{2}\beta_j^T S_j(\mu)\beta_j + \beta_j^T\widehat{\Sigma}_{j\backslash j} + \lambda\sum_{c=1}^{k}\frac{1}{n_c}\|\beta^c_j\|_1$$

avec $S_j(\mu) = \widehat{\Sigma}_{\backslash j \backslash j} + 2\mu A^T A$ *où* A *est une matrice* $k(p-1) \times k(p-1)$ *qui ne dépend pas de* j*.*

Ainsi, la solution de l'équation (2) peut être trouvée de manière très rapide avec une méthode de sous-gradient comme décrit dans [8].

3 Illustration

Nous présentons ici un petit exemple d'application sur un jeu de données simulées très simple : des données d'expression (50 simulations) ont été générées selon une distribution Gaussienne (bruitée) à partir d'une matrice de corrélations partielles connue. Deux graphes sous-jacents relativement similaires ont été utilisés, modélisant deux conditions expérimentales, avec des signes parfois distincts pour les corrélations partielles des deux graphes. Les résultats obtenus par la méthode décrite dans la section 2.2 (« Consensus LASSO ») sont comparés au graphe réel et aux graphes inférés par la méthode « Intertwinned LASSO » et à une approche où les deux graphes sont inférés indépendamment l'un de l'autre. Les résultats sont présentés dans la figure 1. Les résultats obtenus sont

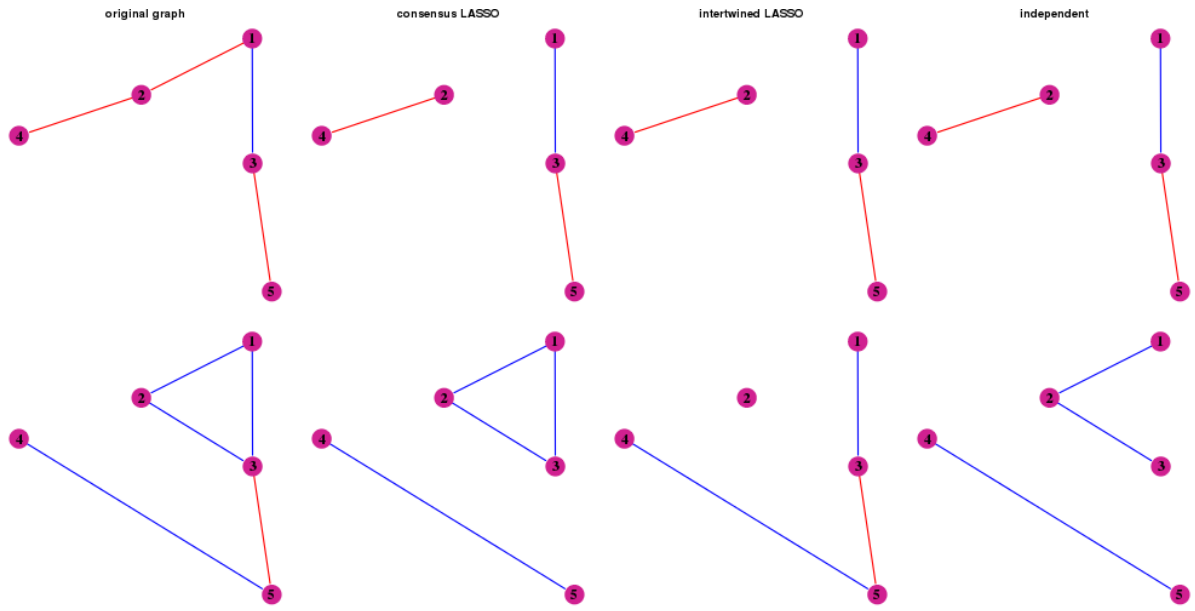

FIGURE 1 – Graphe de corrélations partielles initial (gauche) et graphes inférés par les différentes méthodes (dans l'ordre, Consensus LASSO, intertwinned LASSO et inférence indépendante). Les couleurs, rouge et bleu, représentent le signe de la corrélation partielle (négative ou positive).

assez similaires et relativement cohérents avec le graphe initial et présentent quelques

différences : l'approche « Intertwinned LASSO » échoue à prédire correctement l'arête $(1, 2)$, qui a un signe opposé dans les deux conditions ; par contre, « Consensus LASSO », qui prédit correctement cette arête, échoue à prédire l'arête $(3, 5)$ qui a un signe cohérent dans la condition 2, de même que l'approche indépendante qui se trompe de la même manière sur la prédiction de l'arête $(1, 3)$.

Références

[1] A. Butte and I. Kohane. Mutual information relevance networks : functional genomic clustering using pairwise entropy measurements. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 418–429, 2000.

[2] J. Chiquet, Y. Grandvalet, and C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4) :537–553, 2011.

[3] P. Danaher, P. Wang, and D. Witten. The joint graphical lasso for inverse covariance estimation accross multiple classes. Preprint arXiv 1111.0324v3. Submitted for publication.

[4] D. Edwards. *Introduction to Graphical Modelling*. Springer, New York, 1995.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441, 2008.

[6] N. Jung, M. Maumy-Bertrand, L. Vallat, and F. Bertrand. Inférence conjointe de réseaux de gènes dans de multiples états. In *Actes des 44èmes Journées de Statistique de la SFdS*, Bruxelles, Belgique, 2012.

[7] K. Mohan, J.Y. Chung, S. Han, D. Witten, S.I. Lee, and M. Fazel. Structured learning of Gaussian graphical models. In *Proceedings of NIPS (Neural Information Processing Systems) 2012*, Lake Tahoe, Nevada, USA, 2012.

[8] M.R. Osborne, B. Presnell, and B.A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2) :319–337, 2000.

[9] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, San Francisco, California, USA, 1998.

[10] J. Pearl and S. Russel. *Bayesian Networks*. Bradford Books (MIT Press), Cambridge, Massachussets, USA, 2002.

[11] N. Viguerie, E. Montastier, J.J. Maoret, B. Roussel, M. Combes, C. Valle, N. Villa-Vialaneix, J.S. Iacovoni, J.A. Martinez, C. Holst, A. Astrup, H. Vidal, K. Clément, J. Hager, W.H.M. Saris, and D. Langin. Determinants of human adipose tissue gene expression : impact of diet, sex, metabolic status and *cis* genetic regulation. *PLoS Genetics*, 8(9) :e1002959, 2012.