Une variante des tests de sphéricité pour l'adéquation de données transcriptomiques à un graphe de régulations génétiques.

Vincent Guillemot[1,2], Arthur Tenenhaus[1] et Vincent Frouin[2]

[1] Département Signaux et Systèmes Électroniques
Supélec, 3, rue Joliot Curie, 91190 Gif-sur-Yvette, France
[2] Laboratoire d'Exploration Foctionnelle des Génomes
CEA, 2, rue gaston Crémieux, 91000 Evry, France
vincent.guillemot@cea.fr

Abstract: *When analyzing genomic data, the researcher often encounters the situation where different genetic regulation graphs can be determined on the same dataset with several genetic regulation graph inference methods. One graph is very often compared to another with the help of databases gathering already known interactions between genes: the more known interactions an inferred graph contains, the better it is considered as. We propose a different approach, adapted from the theory of tests, to determine whether a graph fits the dataset.*

Keywords: Genomics, graph, hypothesis test, sphericity test.

1 Introduction

La détermination de graphes décrivant les interactions entre gènes et produits de gènes (ou graphes d'interactions génétiques) est une problématique centrale en bioinformatique et il existe plusieurs stratégies pour les déterminer. D'une part, ces graphes de régulations génétiques peuvent provenir de bases de données faisant l'inventaire de toutes les régulations observées dans des expériences de biologie. D'autre part, de plus en plus de méthodes sont développées pour inférer des graphes de régulations génétiques directement à partir de jeux de données transcriptomiques. Citons par exemple la méthode *Graphical Lasso* (*glasso*) [7] qui estime la matrice de variance covariance inverse d'une classe de profils d'expression. Cette méthode fait l'hypothèse qu'un profil d'expression est une réalisation d'une variable aléatoire X suivant une loi normale multivariée de moyenne μ et de matrice de variance covariance Σ : $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. En se plaçant dans le cadre des modèles graphiques gaussiens [3], la matrice de précision $\mathbf{\Sigma}^{-1}$ de la variable X permet de remonter à un graphe d'indépendances conditionnelles entre les différentes composantes de X (les gènes). Ainsi, la méthode *glasso* permet, à partir de profils d'expression, de déterminer un graphe d'indépendances conditionnelles entre gènes, que l'on considère comme un graphe de régulations génétiques.

Plusieurs questions se posent alors :
- comment vérifier qu'un graphe, quel que soit sa provenance, soit en adéquation avec un jeu de données transcriptomiques particulier ?
- parmi plusieurs graphes, quel est le celui qui est le mieux en adéquation avec les données dont on dispose ?

La réponse que l'on apporte le plus souvent est basée sur la comparaison du graphe obtenu avec un ou plusieurs graphes provenant de bases de données. Malheureusement, les bases de données contenant ces graphes de régulations génétiques de référence sont certainement incomplètes et il est même parfois possible qu'un œil expert y détecte des erreurs. La littérature propose des alternatives sous la forme de tests « locaux » d'adéquation d'un jeu de données gaussien à un graphe. C'est-à-dire que l'on cherche à déterminer s'il manque des arêtes dans ce graphe [4,5] ou si au contraire certaines arêtes sont fausses [3]. Mais de telles méthodes locales ne permettent pas de comparer plusieurs graphes entre eux, sans compter qu'elles sont coûteuses en temps de calcul. Nous proposons une approche plus globale, et certainement préliminaire aux approches précédemment citées, pour déterminer si un graphe $\mathcal{G}$ donné correspond de façon significative à la matrice de précision de données supposées gaussiennes multivariées. Répondre à ce problème permettra, entre autres, de déterminer parmi un ensemble de graphes inférés par des méthodes différentes, lequel correspond le mieux aux données, même si le graphe exact sous-jacent n'est pas connu. Dans le cadre des modèles graphiques gaussiens, une matrice de précision $\mathbf{\Sigma}_0^{-1}$ est déduite du graphe inféré $\mathcal{G}$ et nous calculons ensuite une statistique de sphéricité qui permet de tester l'hypothèse nulle suivante

$$\mathcal{H}_0 = \ll \mathbf{\Sigma} = \mathbf{\Sigma}_0 \gg \tag{1}$$

contre l'hypothèse alternative $\mathcal{H}_1 = \ll \mathbf{\Sigma} \neq \mathbf{\Sigma}_0 \gg$. Ainsi, nous répondons à la question posée uniquement sur la base des observations obtenues après l'expérience transcriptomique effectuée.

Pour toute la suite, nous adoptons des notations identiques à celles du livre d'Anderson [2]. Soit une variable aléatoire de loi $\mathcal{N}(\mu, \mathbf{\Sigma})$. On considère un échantillon de cette variable aléatoire, un individu étant noté $\mathbf{x}_\alpha, \alpha = 1, ..., (n+1)$. On veut tester l'hypothèse (1). Les matrices $\mathbf{A}$ et $\mathbf{S}$ sont définies de la façon suivante :

$$\mathbf{A} = \sum_{\alpha=1}^{n+1} (\mathbf{x}_\alpha - \overline{\mathbf{x}}) (\mathbf{x}_\alpha - \overline{\mathbf{x}})' \text{ et } \mathbf{S} = \frac{1}{n}\mathbf{A}.$$

2 État de l'art

Les tests de sphéricité font l'objet d'une littérature abondante dans le cas où le nombre d'individus est raisonnablement plus important que le nombre de variables, certains sont basés sur le critère du rapport de vraisemblance [2]. Cependant, d'après Ledoit et Wolf [6], le test du rapport de vraisemblance est « dégénéré » dans le cas où $n + 1 < p$.

Une autre statistique est alors proposée par Ledoit et Wolf, mais uniquement pour un test de sphéricité simple (*i.e.* $\mathbf{\Sigma}_0 = I_p$), la transformation pour passer au test de $\mathcal{H}_0$ dans le cas général n'est pas précisée :

$$W = \frac{1}{p}\operatorname{tr}([\mathbf{S} - I_p]^2) - \frac{p}{n}\left(\frac{1}{p}\operatorname{tr}(\mathbf{S})\right)^2 + \frac{p}{n}.$$

Deux hypothèses supplémentaires sont nécessaires pour pouvoir approximer asymptotiquement la loi de W :

(a) le nombre de variables et d'individus sont des fonctions croissantes d'un indice entier k, $p = p_k$ et $n = n_k$ de sorte que $\lim_{k\to+\infty} p_k = +\infty$, $\lim_{k\to+\infty} n_k = +\infty$ et qu'il existe un réel $c > 0$ tel que $\lim_{k\to+\infty} p_k/n_k = c$;

(b) pour chaque indice k, X_k est une matrice $(n_k+1)\times p_k$ de n_k+1 observations i.i.d. d'une variable aléatoire gaussienne multivariée de moyenne μ_k et de matrice de variance covariance $\mathbf{\Sigma}_k$. Soit

$\lambda_i^{(k)}, i = 1, ..., p_k$ les valeurs propres de $\mathbf{\Sigma}_k$. On suppose que la valeur moyenne de ces valeurs propres $\overline{\lambda} = 1/p_k \sum_{i=1} p_k \lambda_i^{(k)} > 0$ et que la dispersion $\delta^2 = 1/p_k \sum_{i=1} p_k (\lambda_i^{(k)} - \overline{\lambda})^2$ sont toutes deux indépendantes de l'indice k.

Sous ces deux hypothèses supplémentaires, et à condition que $(\overline{\lambda} - 1)^2 + \delta^2 = 0$, Ledoit et Wolf ont montré que

$$nW - p \xrightarrow{D} \mathcal{N}(1, 4), \text{ quand } n \to +\infty \text{ et } p \to +\infty, \tag{2}$$

où $\xrightarrow{D}$ indique la convergence en distribution.

3 Statistique proposée

La première étape consiste à transformer le graphe inféré $\mathcal{G}$ en une matrice de précision. On note $\mathbf{Adj}$ la matrice d'adjacence de $\mathcal{G}$, $\mathbf{D}$ la matrice diagonale dont chaque coefficient diagonal est le degré de connectivité du gène concerné. Le lien entre $\mathcal{G}$ et la matrice de précision peut se faire de la manière suivante : si les nœuds i et j ne sont pas connectés dans le graphe $\mathcal{G}$, alors le coefficient en ligne i et colonne j de Σ_0^{-1} est nul, on voit donc que la matrice $\mathbf{Adj}$ remplit cette première contrainte. Une contrainte classique supplémentaire imposée à la matrice de précision est qu'elle doive être inversible (bien que cela ne soit pas nécessaire absolument) ! Une matrice de précision acceptable pour la variable X est alors

$$\mathbf{\Sigma}_0^{-1} = \mathbf{Adj} + \mathbf{D} + I_p.$$

Il suffit ensuite, pour tester si le graphe est en adéquation avec les données, de tester l'hypothèse $\mathcal{H}_0 = \ll \mathbf{\Sigma} = \mathbf{\Sigma}_0 \gg$.

Nous ne proposons pas pour l'instant d'utiliser la propriété 2 pour développer un test avec comme résultat une p-value. Nous allons juste nous servir de la statistique pour classer les graphes inférés les uns par rapport aux autres. La statistique est inspirée de W, mais est généralisée au cas où $\mathbf{\Sigma}_0^{-1}$ n'est pas forcément égal à I_p grâce à la transformation proposée par [2]. Cette nouvelle statistique est notée W' :

$$W' = \frac{1}{p} \operatorname{tr}([\mathbf{S\Sigma^{-1}}_0 - I_p]^2) - \frac{p}{n} \left(\frac{1}{p} \operatorname{tr}(\mathbf{S\Sigma^{-1}}_0) \right)^2 + \frac{p}{n}.$$

4 Simulations et résultats

Les résultats que nous avons obtenus sont uniquement basés sur des données simulées. Nous générons $(n + 1)$ réalisations *i.i.d* d'une variable gaussienne multivariée de moyenne nulle et de matrice de variance covariance $\mathbf{\Sigma}_0$ (de taille $(p \times p)$ avec $n + 1 = 30$ et $p = 150$. L'algorithme est le suivant :

(1) générer un graphe à p nœuds aléatoirement en utilisant le modèle d'Erdős-Rényi [1] de telle sorte que le graphe obtenu soit très épars (seulement 10 % des arêtes sont créées),

(2) considérer ce graphe, noté $\mathcal{G}_{ref}$ pour générer un jeu de données gaussien multivarié de matrice de variance covariance $\mathbf{\Sigma}_{ref} = (\mathbf{Adj} + \mathbf{D} + I_p)^{-1}$ (avec les notations adoptées dans le paragraphe précédent),

(3) obtenir des graphes dont on veut savoir s'il sont en adéquation avec les données, pour cela, utiliser la méthode *glasso* et contrôler la qualité des graphes par le paramètre de régularisation noté ρ de la méthode.

Le problème sur ces données simulées est donc le suivant : pour chaque ensemble de $n+1$ profils d'expression simulés, la méthode *glasso* fournit un graphe $\mathcal{G}$ correspondant à une matrice de précision $\mathbf{\Sigma}_0^{-1}$. La première mesure de la qualité de ce graphe à laquelle nous nous référons est τ, le taux d'arêtes communes entre $\mathcal{G}_{ref}$ et $\mathcal{G}$.

La figure 1(a) présente une comparaison entre la statistique W' et τ. Chaque point correspond aux mesures de W' et τ pour une simulation de $n+1$ profils d'expression. Pour chaque valeur du paramètre ρ, nous avons généré 20 jeux de données de dimensions 30×150 de façon indépendante.

La figure 1(b) présente également W' en fonction de τ, mais pour le même jeu de données : on dispose d'un graphe de $\mathcal{G}_{ref}$, et un jeu de données est simulé pour ce graphe, on applique ensuite la méthode *glasso* pour inférer des graphes de qualité variable.

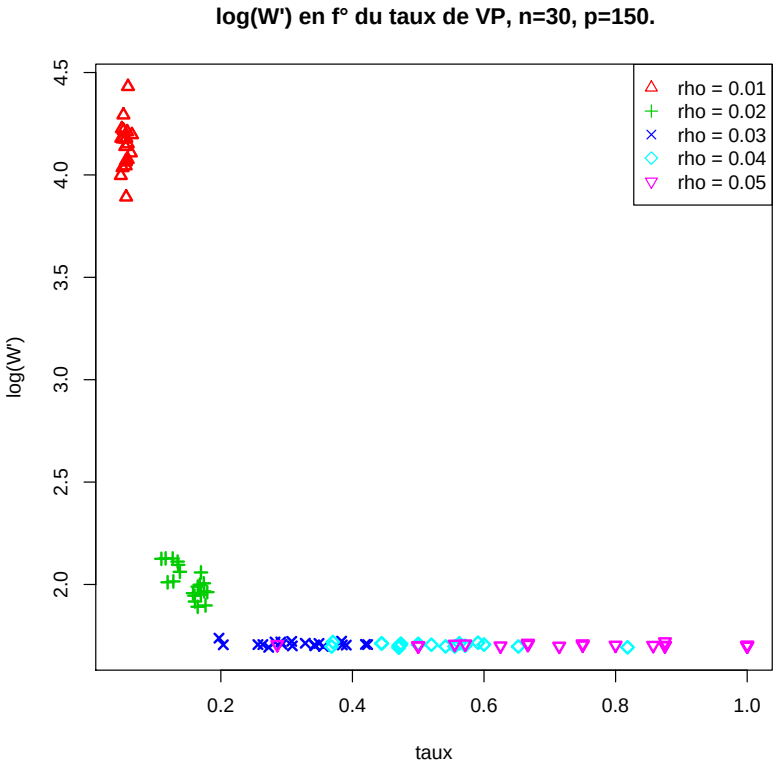
Les deux figures 1(a) et 1(b) permettent de constater que l'évolution de W' est décroissante en fonction de τ. Ainsi, lorsque W' est grand, et même sans la connaissance du graphe sous-jacent aux données, on peut conclure au rejet de $\mathcal{H}_0$. De plus, lorsque deux graphes sont proposés pour un seul jeu de données, les statistiques obtenues sont tout à fait comparables et permettent de déterminer quel graphe est le plus en adéquation avec les données.

Une difficulté subsiste cependant, qui nous a empêché d'utiliser la propriété 2 de convergence asymptotique de W' : sous $\mathcal{H}_0$, $nW' - p$ n'est pas en moyenne égal à 1. Cela remet donc en cause la validité, pour nos données simulées, des hypothèses (a) et (b) évoquées plus haut.

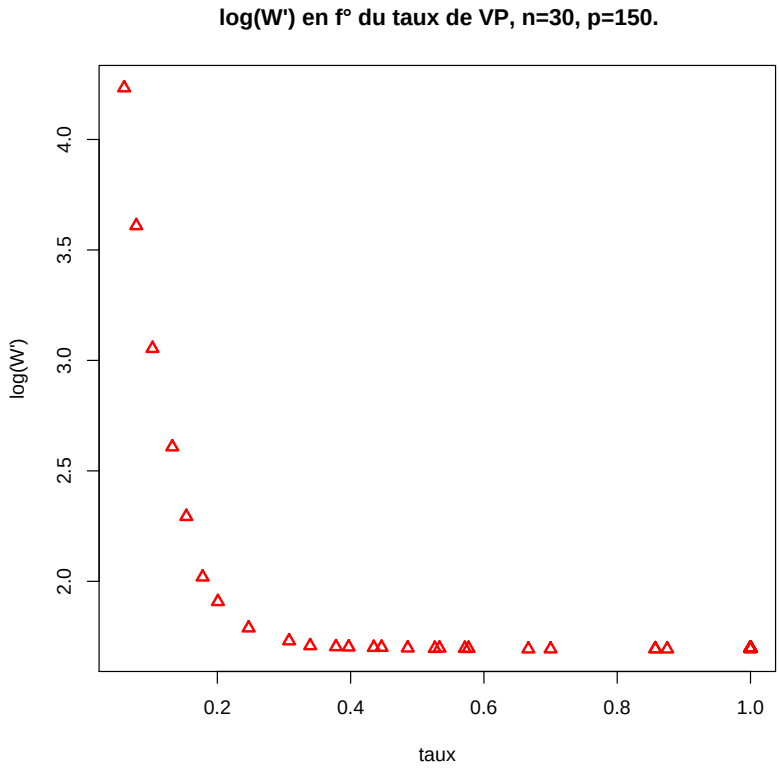
5 Conclusions et perspectives

La statistique W' présentée permet de comparer plusieurs graphes inférés sur un jeu de données transcriptomiques, ce classement ayant été validé sur des données simulées gaussiennes multivariées.

Il faut maintenant travailler sur les résultats asymptotiques présentés par Ledoit et Wolf pour déterminer automatiquement un seuil sur W' permettant de décider au rejet de $\mathcal{H}_0$.



(a) Pour des jeux de données indépendants, $\log(W')$ en fonction de τ.



(b) Pour le même jeu de données, $\log(W')$ en fonction de τ.

Fig. 1. W' en fonction du taux d'arêtes correctement inférées par *glasso*. n et p sont fixés.

Références

[1] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5 :17-61, 1960.

[2] Theodore W. Anderson. *An introduction to multivariate statistical analysis*, third edition, Wiley series in probability and statistics, 2003.

[3] Michael I. Jordan. *Learning in Graphical Models*, The MIT Press, 1998.

[4] Nicolas Verzelen and Fanny Villers. Goodness-of-fit tests for high-dimensional gaussian linear models. *Ann. Statist. (to appear)*, 2008.

[5] Nicolas Verzelen and Fanny Villers. Tests for gaussian graphical models. *Computational Statistics & Data Analysis*, 2008.

[6] Olivier Ledoit and Michael Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, 30 :1081-1102, 2002.

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432-441, 2008.