Inferring gene networks using a sparse factor model approach

Anne Blum, Magalie Houee, Chloé Friguet, Sandrine Lagarrigue, David Causeur

Inferring gene networks using a sparse factor model approach, Statistical Learning and Data Science

Yuna Blum, Magali Houée, Chloé Friguet, Sandrine Lagarrigue, David Causeur

Inferring gene networks using a sparse factor model approach

Yuna Blum[1,2,3], Magalie Houée[2,3] Chloé Friguet[4], Sandrine Lagarrigue[2,3] & David Causeur[1]

[1] *Laboratoire de Mathématiques Appliquées - Agrocampus Ouest, Rennes*
[2] *INRA, UMR598 Génétique Animale, Rennes*
[3] *Agrocampus Ouest, UMR598 Génétique Animale, Rennes*
[4] *UBS / Lab-STICC, Vannes*

Abstract. The availability of genome-wide expression data to complement the measurements of a phenotypic trait opens new opportunities for identifying biologic processes and genes that are involved in trait expression. Usually differential analysis is a preliminary step to identify the key biological processes involved in the variability of the trait of interest. However, this variability shall be viewed as resulting from a complex combination of genes individual contributions. In other words, exploring the interactions between genes viewed in a network structure which vertices are genes and edges stand for inhibition or activation connections gives much more insight on the internal structure of expression profiles. Many currently available solutions for network analysis have been developed but an efficient estimation of the network from high-dimensional data is still a questioning issue. Extending the idea introduced for differential analysis by Friguet *et al.* (2009) [1] and Blum *et al.* (2010) [2], we propose to take advantage of a factor model structure to infer gene networks. This method shows good inferential properties and also allows an efficient testing strategy for the significance of partial correlations, which provides an interesting tool to explore the community structure of the networks. We illustrate the performance of our method comparing it with competitors through simulation experiments. Moreover, we apply our method in a lipid metabolism study that aims at identifying gene networks underlying the fatness variability in chickens.

Keywords. gene networks, high dimensions, factor model, partial correlations

1 Introduction

Inference on gene networks from high throughput expression data is one of the most challenging issues in systems biology. Rather than uncovering single genes for complex traits, a system-based perspective is interesting in elucidating the interactions of genes and environment operating on a complex multicellular biological system [3]. Such a "system" approach involves modeling the relationship among elements of the system such as transcript levels in the form of a network.
Mathematical models for interaction graphs have recently been proposed for gene networks

and inference procedures have been derived to estimate networks from transcription datasets [4]. Linear modeling of gene regulatory networks often appears as a simple and efficient solution, especially for inference from a large transcription profile observed on a few number of samples. In this context, partial correlations between two gene expressions given the remaining profile are viewed as measures of the interaction or co-expression between those two genes. Many currently available solutions for network analysis are therefore based on the so-called Gaussian Graphical Model but an efficient estimation of the network from high-dimensional data is still a questioning issue.

2 Methods

A convenient model of multivariate dependence patterns is Gaussian Graphical Modeling (GGM) [5]. In this framework, a multidimensional Gaussian variable is characterized by the so-called concentration matrix, where conditional independence between pairs of variables is characterized by a zero entry. This matrix may be represented by an undirected graph, where each vertex represents a variable, and an edge connects two vertices if the corresponding pair of random variables are dependent, conditional on the remaining variables. Thus, partial correlations allow the detection of direct genes interactions only.

2.1 GGM: general settings

Let Y be the observed data matrix with n rows, corresponding to the number of samples, and m columns, corresponding to the number of genes. Y is supposed to follow a multivariate normal distribution $Y \sim \mathcal{N}_m(\mu, \Sigma)$ with mean vector $\mu = (\mu_1, ..., \mu_m)'$ and positive-definite covariance matrix $\Sigma = \sigma(ij)$ with $i, j \in [1, m]$.

Let Π be the partial correlation matrix and $\pi_(i, j)$ the correlation between two genes i and j conditionally on all the other genes. It can be shown [6] that partial correlation matrix Π is related to the inverse of the covariance matrix Σ as follows:

$$\pi_{i,j} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

with $\Sigma^{-1} = (\omega_{i,j})$for $i, j \in [1, m]$
After a simple rescaling, the matrix Σ^{-1}, also known as the concentration matrix, can be interpreted as the adjacency matrix of an undirected weighted graph $\mathcal{G}$ representing partial correlation structure between variables $Y_1, ..., Y_m$

A challenging issue remains with Σ^{-1} estimation, as in high dimensions ($n << m$), the estimated covariance matrix is not positive-definite and thus not invertible.

2.2 Existing methods based on GGM

2.2.1 Shrunk estimation of the concentration matrix

Recently, a number of methods have been introduced to address this issue. One of them introduced by Schäfer and Strimmer (2005) [7] and very popular in the biologist community, propose a shrinkage covariance estimator:

$$\Sigma_{shrink.} = \lambda T + (1 - \lambda)S$$

where $\lambda \in [0, 1]$, S is the empirical covariance matrix and T a specified target matrix.

This approach allows to construct a well conditioned positive-definite matrix $\Sigma_{shrink.}$ so that the matrix has full rank and can easily be inverted. The method has been implemented in an R package called GeneNet (ref).

2.2.2 Regularized regression

An alternative route is offered by using regularized regression. Considering the regression for each gene against the others:

$$y_i = \sum_{j \neq i} \beta_{i,j} y_j + \epsilon_i,$$

the regression coefficients can be used to obtain the partial correlations as follow (ref):

$$\pi_{i,j} = sign(\beta_{i,j}) \sqrt{\beta_{i,j} \beta_{j,i}}$$

Therefore, partial correlation estimation can be convert to a regression problem.

As high dimensional problems are supposed to be intrinsically sparse, related methods employe regularized regression techniques [8-10].
For the comparative analysis, we focus on a recent regularized regression method called Sparse PArtial Correlation Estimation (SPACE) which uses a symmetric constraint and an L1 penalization [9].

2.3 Our proposal: sparse factor model

2.3.1 Factor gaussian graphical model

Extending the idea introduced for differential analysis by Friguet *et al.* (2009) [1] and Blum *et al.* (2010) [2], we propose to take advantage of a factor model structure to infer gene networks.

The dependence within the expression profile Y_i is assumed to be modeled by a latent factor structure: there exists q, with $q \leq \min(n, m)$, unobserved factors $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ independently and identically normally distributed with mean 0 and variance I_q, such that:

$$\mathbb{E}(Y_i \mid Z_i) \;=\; \mu + BZ_i \text{ and } \mathrm{Var}(Y_i \mid Z_i) = \Psi$$

where B is the $m \times q$ matrix of loadings with rank q and Ψ is a diagonal matrix which diagonal elements ψ_j^2 are positive. Note that the latent factors can be viewed as sources of dependence across the expression profile in the sense that, conditionally on Z_i, the components of Y_i are independent. It is straightforward checked that the above factor model assumption leads to the following structure for Σ: $\Sigma = \Psi + BB'$, where Ψ is referred to as the specific variance component and BB' as the common variance component. The term BZ_i is also referred to as the kernel of dependence by Leek and Storey (2008) [11].

The resulting log-likelihood is therefore given by

$$-\frac{n}{2}\log\det\big(\Psi + BB'\big) - \frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)'\big(\Psi + BB'\big)^{-1}(Y_i-\mu).$$

It is well-known that the above log-likelihood is maximized with respect to μ at $\hat{\mu} = \bar{Y}$, for all Ψ and B. As we focus on the maximum likelihood (ML) estimation of the variance parameters, μ will hereafter be replaced by its $\bar{Y}$ in deviance model expression:

$$\mathcal{D}(\Psi, B) \;=\; n\log\det\big(\Psi + BB'\big) + n\,\mathrm{trace}\big[S\big(\Psi + BB'\big)^{-1}\big],$$

where $S = (1/n)\sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})'$. Profiles Y_i will hereafter be considered as centered.

Direct maximization of $\mathcal{D}(\Psi, B)$ is described by Jøreskog (1969) [12] and is quite popular among users of factor analysis models. However, when the sample size n is small regarding the size m of the expression profile, this algorithm can be numerically unstable and cumbersome. An alternative EM algorithm is proposed by Rubin and Thayer (1982) [13], taking advantage of the factorization of the likelihood based on the conditional distribution of the profile given the latent factors:

$$\begin{aligned}
\mathcal{D}(\Psi, B; Z) &= n\sum_{j=1}^{m}\log\psi_j^2 + \Big[\sum_{i=1}^{n}(Y_i - BZ_i)'\Psi^{-1}(Y_i - BZ_i)\Big] + \sum_{i=1}^{n} Z_i'Z_i, \qquad (1)\\
&= n\sum_{j=1}^{m}\log\psi_j^2 - n\Big[\mathrm{trace}\big(\Psi^{-1}S\big) - 2\,\mathrm{trace}\big(\Psi^{-1}BS_{yz}'\big) + \mathrm{trace}\big(B'\Psi^{-1}BS_{zz}\big)\Big]\\
&\quad + n\,\mathrm{trace}\big(S_{zz}\big), \qquad (2)
\end{aligned}$$

where $S = (1/n)\sum_{i=1}^{n} Y_iY_i'$, $S_{yz} = (1/n)\sum_{i=1}^{n} Y_iZ_i'$ and $S_{zz} = (1/n)\sum_{i=1}^{n} Z_iZ_i'$. Therefore, the log-likelihood is a linear combination of the sufficient statistics S, S_{yz} and S_{zz}.

The straightforward adaptation of the EM algorithm introduced by Dempster *et al.* (1977) [14] to the present issue is obtained by considering the factors Z_i as missing data. The E step consists in calculating the conditional expectation of $\mathbb{E}_y\mathcal{D}(\Psi, B; Z)$ given Y in expression (2) using the following intermediate results:

$$
\begin{aligned}
C_{yz} &= \mathbb{E}_y(S_{yz}) = S(\Psi + BB')^{-1}B, \\
C_{zz} &= \mathbb{E}_y(S_{zz}) = B'(\Psi + BB')^{-1}S(\Psi + BB')^{-1}B + I_q - B'(\Psi + BB')^{-1}B.
\end{aligned}
$$

Thanks to the Woodbury's identity,

$$
(\Psi + BB')^{-1} = \Psi^{-1} - \Psi^{-1}BG^{-1}B'\Psi^{-1},
$$

where $G = I_q + B'\Psi^{-1}B$, this step only requires the inversion of a $q \times q$ matrix.

The M-step consists in minimizing $\mathbb{E}_y\mathcal{D}(\Psi, B; Z)$ with respect to B and Ψ.

$$
\begin{aligned}
\hat{B} &= C_{yz}C_{zz}^{-1}, \\
\hat{\Psi} &= \operatorname{diag}\Big[S - C_{yz}C_{zz}^{-1}C_{yz}'\Big].
\end{aligned}
$$

This algorithm also described by [1] is used hereafter for the partial correlation estimation under the network sparsity assumption [9].

2.3.2 Partial correlation significance testing

First, lut us introduce the following new parameterization:

$$
\begin{aligned}
\varphi &= \Psi^{-\frac{1}{2}}, \\
\theta &= \Psi^{-\frac{1}{2}}B(I + B'\Psi^{-1}B)^{-\frac{1}{2}}.
\end{aligned} \tag{3}
$$

The above parameterization is reversible:

$$
\begin{aligned}
\Psi &= \varphi^{-2}, \\
B &= \varphi^{-1}\theta(I_q - \theta'\theta)^{-\frac{1}{2}}.
\end{aligned}
$$

It is straightforward checked that $\Sigma^{-1} = \varphi(I_m - \theta\theta')\varphi$. Sparsity of Σ^{-1} means that there exists pairs (i, j) such that $\left[\Sigma^{-1}\right]_{ij} = 0$. For those pairs (i, j), using the above expression of Σ^{-1} leads to the following orthogonality condition on the ith and jth rows (resp. θ_i and θ_j) of θ: $\theta_i'\theta_j = 0$.

Using the functional invariance property of the maximum likelihood, the estimators ML $\hat{\varphi}$ and $\hat{\theta}$ of φ and θ are now derived by replacing Ψ and B by $\hat{\Psi}$ and $\hat{B}$ respectively using (3).

The parameters of the factor model are hereafter expressed as $(\text{vec}(\theta)', diag(\varphi)')'$, where vec(.) is the matrix operator transforming a $m \times q$ matrix into a vector of length mq obtained by the superposition of its row vectors and $diag(.)$ is the operator transforming a diagonal matrix into a vector of length m of its diagonal elements.

The asymptotic variance of the ML $(\text{vec}(\theta)', diag(\varphi)')'$ is given in the following proposition:

Proposition 1 *Let $\mathcal{H}$ denote the $m(q+1) \times m(q+1)$ Fisher's information matrix of the factor model. Then, $\mathcal{H}$ can be expressed as follow:*

$$\mathcal{H} = \begin{pmatrix} \mathcal{H}_\theta & \mathcal{H}_{\theta,\varphi} \\ \mathcal{H}'_{\theta,\varphi} & \mathcal{H}_\varphi \end{pmatrix},$$

where

$$\begin{aligned} n^{-1}\mathcal{H}_\theta &= \Big[I_m \odot (I_q - \theta'\theta)^{-1}\Big] \operatorname{vec}(\theta) \operatorname{vec}(\theta)' \Big[I_m \odot (I_q - \theta'\theta)^{-1}\Big] + (I_m - \theta\theta')^{-1} \odot (I_q - \theta'\theta)^{-1} \\ &\quad -(I_m - \theta\theta')^{-1} \odot I_q, \\ n^{-1}\mathcal{H}_{\theta,\varphi} &= -\varphi^{-1} D'_\theta \Big[I_m \odot (I_q - \theta'\theta)^{-1} + (I_m - \theta\theta')^{-1} \odot I_q\Big], \\ n^{-1}\mathcal{H}_\varphi &= 2\varphi^{-2} + \varphi^{-1} D'_\theta \Big[I_m \odot (I_q - \theta'\theta)^{-1}\Big] D_\theta \varphi^{-1} - \varphi^{-1} D'_\theta \Big[(I_m - \theta\theta')^{-1} \odot I_q\Big] D_\theta \varphi^{-1}, \end{aligned}$$

where D_θ is the following $mq \times m$ matrix:

$$D_\theta = \begin{pmatrix} \theta_1 & 0 & \dots & 0 \\ 0 & \theta_2 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \theta_m \end{pmatrix}$$

The Information matrix is now used to test the significance of partial correlation. Indeed, it is deduced from the standard ML theory that the asymptotic distribution of $\sqrt{n}(\operatorname{vec}(\hat{\theta}) - \operatorname{vec}(\theta))$ is normal with mean 0 and variance V_θ given by the $mq \times mq$ left upper block of $\mathcal{H}^{-1}$. Standard results on inversion of partitioned matrices can be used:

$$V_\theta = \mathcal{H}_\theta^{-1} + \mathcal{H}_\theta^{-1}\mathcal{H}_{\theta,\varphi}(\mathcal{H}_\varphi - \mathcal{H}'_{\theta,\varphi}\mathcal{H}_\theta^{-1}\mathcal{H}_{\theta,\varphi})^{-1}\mathcal{H}'_{\theta,\varphi}\mathcal{H}_\theta^{-1}.$$

Moreover:

$$\mathcal{H}_\theta^{-1} = K_\theta - \frac{K_\theta \operatorname{vec}(\theta) \operatorname{vec}(\theta)' K_\theta}{1 + \operatorname{vec}(\theta)' K_\theta \operatorname{vec}(\theta)},$$

where $K_\theta = (I_m - \theta\theta') \odot \Big[(I_q - \theta'\theta)^{-1} - I_q\Big]^{-1}$.

3 Simulations

3.1 Network and data generation

We generated a true partial correlation matrix Π with a density of 25% and comprised of 5 equivalent blocks of non-zero entries representing 5 modules of genes highly connected

(see Figure 1). The number of variables is fixed at m = 200, and 3 different sample sizes are considered: 200, 100 and 50. For each scenario, a total of 100 replications are performed and for each replication, the data are drawn randomly from a multivariate normal distribution with correlation structure derived from Π using the function *ggm.simulate.data* from the GeneNet package [7].

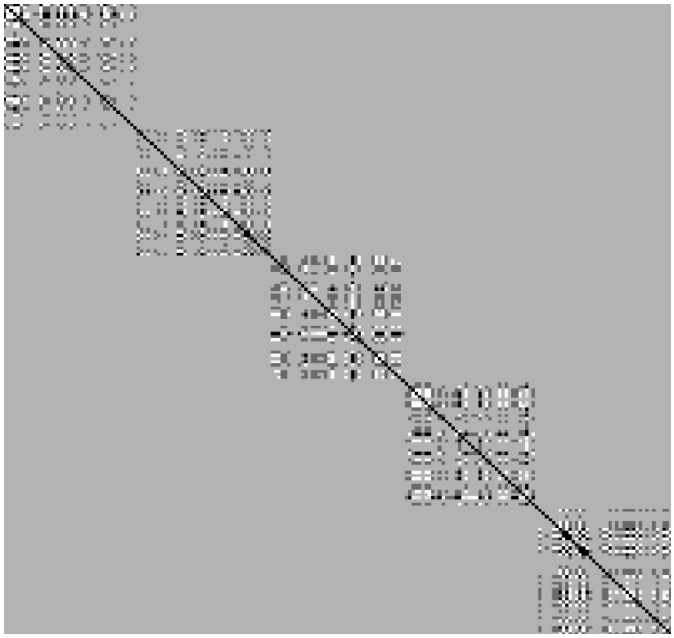

FIGURE 1 – True partial correlation matrix generated

3.2 Validation

For fair comparisons, the default parameters where used for each algorithm without additional tuning. For our method FA, we choose 10 as the default number of factors in the factor gaussian graphical model which is consistent with previous real life studies (ref,ref). Furthermore, for GeneNet and FA, an FDR cutoff of 5% was chosen.
Performance of methods is assessed by considering both partial correlation estimation and edge detection accuracy. An edge detected by a method is considered as a true positive (TP) or as a false positive (FP) depending on the presence or not of the corresponding edge in the underlying true network, respectively. Analogously, a zero entry (no edge) is considered as a true negative (TN) or a false negative (FN) depending on whether the corresponding edge is present or not in the underlying true network, respectively. The confusion matrix was calculated for each method and as in [15] we also calculated the F-scores, which allows a compact representation of the precision (p) - recall (r) diagram.

F-score quantity is expressed as follow:

$$F = \frac{2pr}{r+p}$$

with $p = \frac{TP}{TP+FP}$ and $r = \frac{TP}{TP+FN}$

4 Results

We first compare the methods in terms of partial correlation estimation. Figure 2 shows for each scenario and methods, the estimated partial correlations vs. the true partial correlations. n corresponds to the number of sample and MSE to the mean squared error. First of all, one can see that as sample size increased, the performance of the 3 methods tends to improve which is expected and consistent with previous research (ref compa). SPACE tends to overestimate the extreme partial correlations in comparison to GeneNet and FA method. In terms of MSE, all 3 methods show good results for sample size 200. However, when sample size is smaller, the estimation accuracy becomes very poor for GeneNet. For lower sample sizes, FA gives the best results in terms of partial correlation estimation.
Concerning the edge detection performance, we report for each case and method, the confusion matrix, the sensibility and specificity in Table 1. For GeneNet method, we consider both significant testing with and without FDR control. Except for GeneNet without FDR control, all other methods show an high specificity for edge detection. However the sensibility is lower and particularly for GeneNet method with FDR control. This results is consistent with a recent study [16] showing that GeneNet and Space perform well in identifying a few connections with high specificity. For each different sample size, the best sensibility/specificity ratio is given by FA method. In Figure 3, we represent the F-score boxplots. The figure shows that GeneNet without FDR control and SPACE give comparable results. However, SPACE tends to be more affected by sample size decrease. Interestingly, FA method gives the highest F-scores in average but with a higher degree of dispersion for sample size 200.

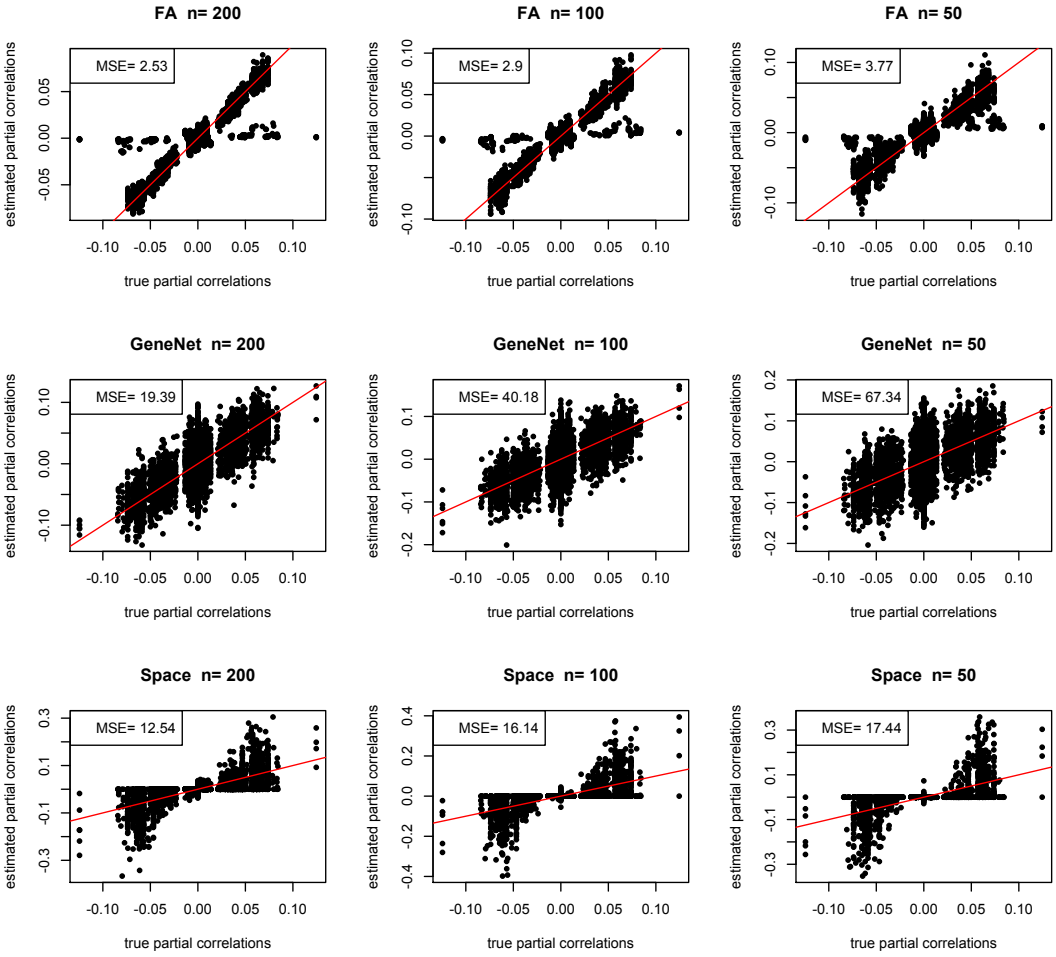

FIGURE 2 – Estimation of partial correlations

		total edge detected	TP	TN	FP	FN	specificity	sensibility
FA	n=200	3832.4	2858.4	14026.0	974.0	2141.6	0.935	0.572
FA	n=100	3194.4	2535.1	14340.7	659.3	2464.9	0.956	0.507
FA	n=50	2575.0	2314.0	14739.0	261.0	2686.0	0.983	0.463
GeneNet	n=200	10764.6	3027.0	7262.4	7737.6	1973.0	0.484	0.605
GeneNet	n=100	10552.9	2815.9	7263.0	7737.0	2184.1	0.484	0.563
GeneNet	n=50	10460.7	2715.2	7254.4	7745.6	2284.8	0.484	0.543
GeneNet + FDR	n=200	135.1	129.7	14994.6	5.4	4870.3	1.000	0.026
GeneNet + FDR	n=100	9.3	9.0	14999.7	0.3	4991.0	1.000	0.002
GeneNet + FDR	n=50	1.1	1.1	15000.0	0.0	4998.9	1.000	0.000
Space	n=200	1696.1	1469.2	14773.1	226.9	3530.8	0.985	0.294
Space	n=100	1138.1	1100.6	14962.5	37.5	3899.4	0.998	0.220
Space	n=50	785.5	781.1	14995.6	4.4	4218.9	1.000	0.156

Table 1 - Edge detection performance

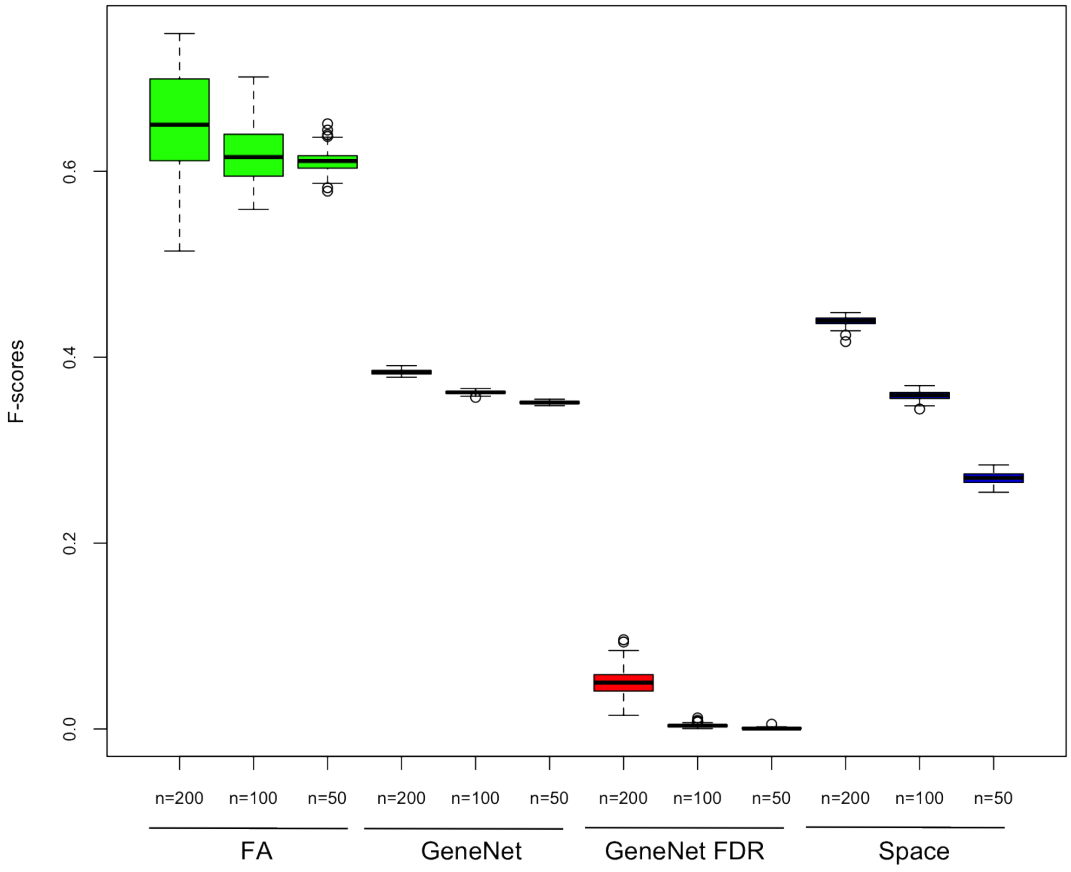

FIGURE 3 – F-score boxplots

5 Application

We apply our method in a lipid metabolism study that aims at identifying biological processes and genes involved in abdominal fatness variability in chickens [2,17] which is an economic trait of great interest. The biological purpose is here to better characterize a region of the genome known to controlled the abdominal fatness in chicken. Using the transcriptome profile of 45 chickens, 59 genes were found as being controlled by this region. Using our method, we infer the gene network of these 59 genes and highlight two functional modules (Figure 4). Using the available annotation, one of the two modules could be related to the lipid metabolism and more precisely to the cholesterol metabolism as it contains the gene encoding for the last enzyme of the cholesterol synthesis process. Thanks to the inferred gene network, we provide thus a new functional hypothesis about one of the causal mutations that affect abdominal fatness in chickens.

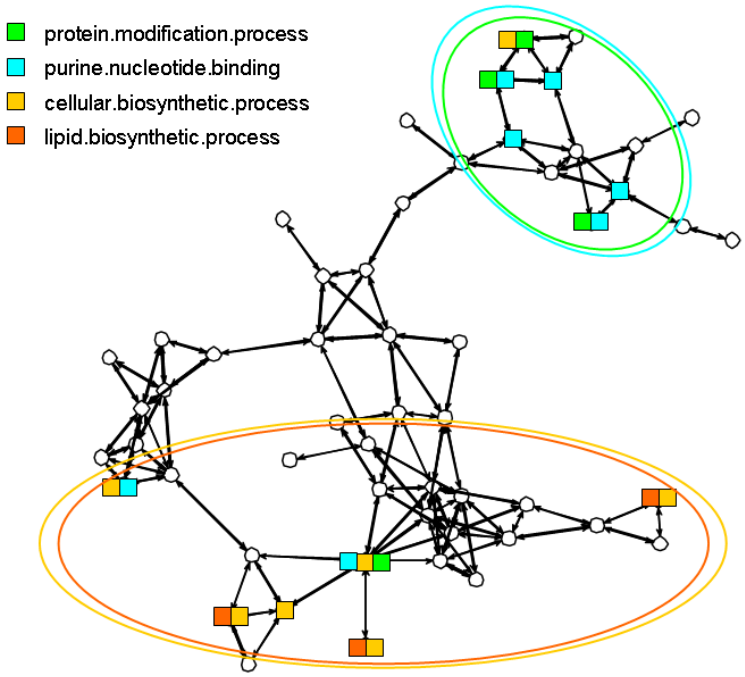

FIGURE 4 – Inferred gene network of the 59 genes

6 Conclusion

In this study, we investigate a new approach to infer gene networks based on a factor model structure. This method allows an efficient testing strategy for the significance of partial correlations. Through simulation studies, we show that our method achieves good performance in edge detection and partial correlation estimation in comparison with 2 competitors widely used by the biologist community. Moreover, we apply our method in a lipid metabolism and provide a new functional hypothesis about one of the causal mutations that affect abdominal fatness in chickens.

References

1. Blum, Y. Le Mignon, G., Lagarrigue, S. et Causeur, D. (2010). A factor model to analyze heterogeneity in gene expression, BMC Bioinformatics, 11:368.
2. Friguet, C., Kloareg, M. et Causeur, D. (2009). A Factor Model Approach to Multiple Testing Under Dependence, Journal of the American Statistical Association,

104:488, 1406-1415.

3. Sieberts, S.K. and Schadt, E.E. (2007). Moving toward a system genetics view of disease. Mammalian Genome, 6: 389-401
4. Goldenberg A, Zheng A. X., Fienberg S. E., Airoldi E. M. (2009). A survey of statistical network models, Foundations and Trends in Machine Learning 2, 1-117
5. Ambroise, C., Chiquet, J. and Matias. C. (2009) Inferring sparse Gaussian graphical models with latent structure. Electronic Journal of Statistics, 3:205-238
6. Whittaker J. (1990) Graphical Models in Applied Multivariate Statistics New York: Wiley.
7. Schäfer J, Strimmer K (2005). An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics 21: 754-64.
8. Friedman, T. Hastie, and R. Tibshirani. (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441.
9. Peng J., Wang P., Zhou N., Zhu J., (2009). Partial Correlation Estimation by Joint Sparse Regression Models. JASA. 104(486): 735-746.
10. Chiquet, J. and Smith, A. and Grasseau, G. and Matias, C. and Ambroise, C. (2009). Simone: Statistical inference for modular networks. Bioinformatics, 3: 417-418
11. Leek J.T. and Storey J.D. (2008) A general framework for multiple testing dependence. Proceedings of the National Academy of Sciences, 105: 18718-18723.
12. Jøreskog, K.-G. (1969) A general approach to confirmatory maximum likelihood actor analysis, Psychometrika, 34, 183–202.
13. Rubin, D.B. and Thayer, D.T. (1982). EM algorithms for ML factor analysis, Psychometrika 47, 1, 69–76
14. Dempster P, Laird N. M., and Rubin D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B, 34:1–38
15. Meyer, P. and Lafitte, F. and Bontempi, G.. (2008). minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information BMC bioinformatics, 9: 461
16. Allen, J.D. and Xie, Y. and Chen, M. and Girard, L. and Xiao, G. (2012). Comparing Statistical Methods for Constructing Large Scale Gene Networks. PloS one, 7: e29348
17. Le Mignon, G. and Désert, C. and Pitel, F. and Leroux, S. and Demeure, O. and Guernec, G. and Abasht, B. and Douaire, M. and Le Roy, P. and Lagarrigue, S. (2009). Using trancriptome profiling to characterize QTL regions on chicken chromosome 5. BMC Genomics. 1: 575