Reconstruction quality of a biological network when its constituting elements are partially observed

Victor Picheny, Jimmy Vandel, Matthieu Vignes and Nathalie Villa-Vialaneix
(authors listed in alphabetical order)
MIA-T unit, INRA Toulouse, chemin de Borderouge, 31326 Castanet-Tolosan Cedex, France

Motivation Unravelling regulatory regulations between biological entities is of utmost importance to understand the functioning of living organisms. As the number of available samples is often very low (often less than one hundred), inference methods are frequently performed on a subset of variables which make sense in the mechanisms under study. Classical remedies are either data driven (*e.g.*, differentially expressed genes) or knowledge driven (*e.g.*, using ontology information). However, whatever the chosen solution, important variables are very likely missed by the selection process, which is the issue at stake in the present paper.

Dealing with partially available variables A Gaussian graphical framework is considered, which is specified by its concentration matrix $\mathbf{K} = \Sigma^{-1}$ (Σ is the covariance matrix of the Gaussian model). Because of high-dimensional issues, $\mathbf{K}$ (from which the underlying network can be deduced), is frequently estimated by adding an ℓ_1-penalty to the model likelihood (the method is known under the name *glasso*, see Friedman et al. 2007). But, if the gene expressions are represented by the variable $X = (X_o, X_h)$, where the first p nodes X_o are observed and the remaining nodes are hidden, directly applying the glasso to the observed variables leads to a spurious estimation of the network topology: $\Sigma_{o,o}^{-1} = K_{o,o} - K_{o,h}K_{h,h}^{-1}K_{h,o}$. In less formal terms, this simply means that estimating the true network topology on observed nodes need to access some information on the structure of hidden nodes. Chandrasekaran et al. (2012) gives identifiability conditions for this model and proposed a consistent convex formulation of the problem which allows them to solve it (hereafter denoted by "CPW-S+L"). However, whether accounting or not for unobserved variables improves the inference quality, prediction capability and global understanding of the system remains an open question in practice.

Aim and scope The objective of this work is twofold. Firstly, we conduct an empirical study to evaluate and compare the use of (blindly used) glasso and CPW-S+L for network reconstruction, as regards to: (i) the sample size n, (ii) the ratio of missing variables r, (iii) the missing node context (random or peculiar nodes), (iv) several methods for simulating the data. The overall objective is to provide some guidelines for practicioners for the choice of the methods and for the analysis of their results.

Secondly, we aim at discussing the concept of "natural graph" inherited from the complete graph. Three different possibilities to account for the reference graph (i.e., graph projections) were studied: (i) the graph whose links reflect path existence in the complete graph, (ii) the graph inherited from edges of the complete graph only, or (iii) the graph learnt from complete data. This questions whether inference algorithms purely recover actual edges or whether they reflect dependencies either direct or of higher orders. The concept of reference graph is also closely linked to the user's objective, in particular, whether inference or prediction is sought.

Preliminary results Our first experiments show that, as expected, the overall reconstruction quality decreases as n decreases and as r increases. Moreover, the reconstructed network quality is worse when central nodes are missing than when nodes are missing at random. Also, glasso obtains worse performance than CPW-S+L but a perhaps more surprising conclusion is that predicted edges do not replace links which cannot be retrieved because intermediate nodes are missing.

References

J. Friedman, T. Hastie, and R. Tibshirani (2007) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432-441.

V. Chandrasekaran, P.A. Parillo, and A.S. Willsky (2012). Latent variable graphical model seletion via convex optimization. *Annals of Statistics* **40**:1935–1967.