A multiattribute Gaussian graphical model for inferring multiscale regulatory networks: an application in breast cancer

Julien Chiquet, Guillem Rigaill and Martina Sundqvist
AgroParisTech, INRA, Université Paris-Saclay

Abstract . This chapter addresses the problem of reconstructing regulatory networks in molecular biology by integrating multiple sources of data. We consider data sets measured from diverse technologies all related to the same set of variables and individuals. This situation is becoming more and more common in molecular biology, for instance when both proteomic and transcriptomic data related to the same set of "gene" are available on a given cohort of patients.

To infer a consensus network that integrates both proteomic and transcriptomic data, we introduce a multivariate extension of Gaussian graphical models (GGM), which we refer to as *multiattribute* GGM. Indeed, the GGM framework offers a good proxy for modeling direct links between biological entities. We perform the inference of our multivariate GGM with a neighborhood selection procedure that operates at a multiscale level. This procedure employs a group-Lasso penalty in order to select interactions which operate both at the proteomic and the transcriptomic level between two genes. We end up with a *consensus* network embedding information shared at multiple scales of the cell. We illustrate this method on two breast cancer data sets.

Key words: Multiscale Regulatory Network · Gaussian Graphical Model · Group-Lasso · Proteomic Data · Multi-Omic Data

Julien Chiquet
MIA Paris, AgroParisTech/INRA, 16 rue Claude Bernard, 75231 Paris CEDEX 05, France
e-mail: julien.chiquet@inra.fr

Martina Sundqvist
MIA Paris, AgroParisTech/INRA, 16 rue Claude Bernard, 75231 Paris CEDEX 05, France
e-mail: martina.sundqvist@agroparistech.fr

Guillem Rigaill
IPS2, Bâtiment 630, Rue de Noetzlin, Plateau du Moulon, 91405 Orsay, France
e-mail: guillem.rigaill@inra.fr

1 Introduction

Gaussian Graphical Models (GGMs): a canonical framework for network inference modeling. Gaussian Graphical Models (GGMs) (Lauritzen, 1996; Whittaker, 1990) are a very convenient tool for describing the patterns at play in complex data sets. Indeed, through the notion of partial correlation, they provide a well-studied framework for spotting direct relationships between variables, and thus reveal the latent structure in a way that can be easily interpreted. Application areas are very broad and include for instance gene regulatory network inference in biology (using gene expression data) as well as spectroscopy, climate studies, functional magnetic resonance imaging, etc. Estimation of GGMs in a sparse, high-dimensional setting has thus received much attention in the last decade, especially in the case of a single homogeneous data set (Meinshausen and Bühlmann, 2006; Friedman et al, 2008; Banerjee et al, 2008; Yuan, 2010; Cai et al, 2011).

However, this simple canonical setup is uncommon in real world data sets, because of both the complexity of the mechanism at play in biology and the multiplicity of the sources of data in omic. Hence, the need for variants of sparse GGM more adapted to recent omic data set is huge.

As a first example, the work developed in Ambroise et al (2009); Chiquet et al (2009) addresses the introduction of a possible special organization of the network itself to drive the reconstruction process. Indeed, while sparsity is necessary to solve the problem when few observations are available, biasing the estimation of the network towards a given topology can help us find the correct graph in a more robust way, by preventing the algorithm from looking for solutions in regions where the correct graph is less likely to reside. As a second example, Chiquet et al (2011) addresses the problem of sample heterogeneity which typically occurs when several assays are performed in different experimental conditions that potentially affect the regulations, but are still merged together to perform network inference as data is very scarce. We remedy heterogeneity among sample experiments by estimating multiple GGMs, each of which matches different modalities of the same set of variables, which correspond here to the different experimental conditions. This idea, coupled with the integration of biological knowledge, was further explored for application in cancer in Jeanmoungin et al (2014).

A deeper generalization of GGM comes by integrating multiple types of data measured from diverse platforms, what is sometimes referred to as *horizontal* integration: not only does this means a better treatment of the heterogeneity of the data, but it also makes the network reconstruction more robust. The model presented in this chapter gives an answer to this question by offering a solution to reconstruct a sparse, multiattribute GGM, recoursing on both proteomic and genomic data to infer a consensus network. Our main motivating application is a better understanding of heterogeneity in breast cancer, as detailed below.

Proteomic and Transcriptomic data integration in cancer. Protein deregulations, leading to abnormal activation of signaling pathways, contribute to tumorige-

nesis (Giancotti, 2014). Knowing the level of activation of the signaling pathways in any subgroup of tumors could therefore be a key indication to understand the biological mechanisms involved in tumorigenesis, and to identify some therapeutic targets. Therefore, the analysis of proteins is essential. However, measuring the expression of proteins is more difficult to implement than the measure of transcriptome (RNA) or genome (DNA). Several technologies have been developed to measure the proteome, but the number of samples and the number of proteins that can be studied simultaneously is, up to now, limited. A useful technique for this task is the RPPA (Reverse Phase Protein Arrays). It allows studying protein expression levels and the activity status of a few hundred proteins by analyzing their phosphorylated state, in several hundred of samples (Akbani et al, 2014).

To summarize, better understanding the proteome of tumors is essential to further our knowledge of cancer cells but proteome data are still small and rare. Integration of proteomic and transcriptomic data is a promising avenue to get the most of available proteomic datasets and better understand the relatives roles of transcriptome and proteome. To take into account the different levels of information, a solution is to use multivariate GGM. Since it is probable that the proteomic and transcriptomic heterogeneity in cancers is caused by some few major underlying alterations, the hypothesis of proximity in between networks seems reasonable. Identifying commonalities between the transcriptome and proteome networks should help the prediction of the proteome using the transcriptome for which several large public cancer data sets are available.

Chapter Outline. In the next section we give a quick overview of the literature of sparse GGM (models, basic theoretical results, inference and software). This provides the reader with the necessary material to approach the model at play in this chapter, dedicated to multiattribute GGM, introduced in Section 3. In Section 4 we perform some numerical studies: we demonstrate on simulated data the superiority of our approach in several scenarios. Then, two breast cancer data sets are used to illustrate the reconstruction of multiscale regulatory networks.

2 Background

This section provides an overview on the state-of-the-art ℓ_1-regularization methods for sparse GGM inference and their most recent striking variants, insisting on their computational and statistical properties.

2.1 Basics on Gaussian graphical models

Let $\mathscr{P} = \{1, \ldots, p\}$ be a set of fixed vertices and $X = (X_1, \ldots, X_p)^\intercal$ a random vector describing a signal over this set. The vector $X \in \mathbb{R}^p$ is assumed to be multivariate

Gaussian with unknown mean and unknown covariance matrix $\Sigma = (\Sigma_{ij})_{(i,j)\in\mathscr{P}^2}$. No loss of generality is involved when centering X, so we may assume that $X \sim \mathscr{N}(\mathbf{0}_p, \Sigma)$. The covariance matrix Σ, equal to $\mathbb{E}(XX^\intercal)$ under the assumption that X is centered, belongs to the set $\mathscr{S}_p^+$ of positive definite symmetric matrices of size $p \times p$.

Graph of conditional dependencies. GGMs endow Gaussian random vectors with a graphical representation $\mathscr{G}$ of their *conditional dependency structure*: two variables i and j are linked by an undirected edge (i, j) if, conditional on all other variables indexed by $\mathscr{P}\backslash\{i, j\}$, random variables X_i and X_j remain or become dependent. Thanks to the Gaussian assumption, conditional independence actually boils down to a zero conditional covariance $\mathrm{cov}(X_i, X_j|X_{\mathscr{P}\backslash\{i,j\}})$, or equivalently to a zero partial correlation which we denote by ρ_{ij}, the latter being a normalized expression of the former.

Concretely, the inference of a GGM is based upon a classical result originally emphasized in Dempster (1972) stating that partial correlations ρ_{ij} are actually proportional to the corresponding entries in the *inverse* of the covariance matrix $\Sigma^{-1} = \Theta$, also known as the *concentration matrix*. More precisely, we have

$$\rho_{ij} = -\Theta_{ij}/\sqrt{\Theta_{ii}\Theta_{jj}}, \qquad \Theta_{ii} = \mathrm{Var}(X_i|X_{\mathscr{P}\backslash i})^{-1}; \tag{1}$$

thus Θ directly describes the conditional dependency structure of X. Indeed, after a simple rescaling, Θ can be interpreted as the adjacency matrix of an undirected weighted graph representing the partial covariance (or correlation) structure between variables $X_1, \ldots, X_p$. Formally, we denote by $\mathscr{G} = (\mathscr{P}, \mathscr{E})$ this graph, the edges of which are characterized by

$$(i, j) \in \mathscr{E} \Leftrightarrow \Theta_{ij} \neq 0, \quad \forall (i, j) \in \mathscr{P}^2 \text{ such that } i \neq j.$$

In words, $\mathscr{G}$ has no self-loop and contains all edges (i, j) such that Θ_{ij} is nonzero. Therefore recovering nonzero entries of Θ is equivalent to inferring the graph of conditional dependencies $\mathscr{G}$, and the correct identification of nonzero entries is the main issue in this framework.

Maximum Likelihood inference. GGMs fall into the family of exponential models for which the whole range of classical statistical tools applies. As soon as the sample size n is greater than the number p of variables, the likelihood admits a unique maximum over $\mathscr{S}_p^+$, defining a maximum likelihood estimator (MLE): suppose we observe a sample $\left\{X^1, \ldots, X^n\right\}$ composed of n i.i.d. copies of X, stored row-wise once centered in a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ such that $(X^i)^\top$ is the ith row of $\mathbf{X}$. The empirical covariance matrix is denoted by $\mathbf{S}_n = \mathbf{X}^\intercal\mathbf{X}/n$. Maximizing the likelihood is equivalent to

$$\widehat{\Theta}^{\mathrm{mle}} = \underset{\Theta\in\mathscr{S}_p^+}{\arg\max} \quad \log\det(\Theta) - \mathrm{Tr}(\Theta\mathbf{S}_n). \tag{2}$$

When $n > p$, Problem (2) admits a unique solution equal to $\mathbf{S}_n$. The scaled empirical covariance matrix $\mathbf{S}_n$ follows a Wishart distribution while its inverse $\mathbf{S}_n^{-1}$ follows an inverse Wishart distribution with computable parameters.

There are two major limitations with the MLE regarding the objective of graph reconstruction by recovering the pattern of zeroes in Θ. First, it provides an estimate of the saturated graph: all variables are connected to each other; second, we need n to be larger than p to be able to even define this estimator, which is rarely the case in genomics. In any case, the need for regularization and feature selection is huge. A natural assumption is that the true set of direct relationships between the variables remains small, that is, the true underlying graph is sparse (say, of the order of p rather than the order of p^2). Sparsity makes estimation feasible in the case where $n < p$ since we can concentrate on sparse or shrinkage estimators with fewer degrees of freedom than in the original problem. Henceforth, the question of selecting the correct set of edges in the graph is treated as a question of variable selection.

High-dimensional inference of GGM. The different methods for the inference of sparse GGMs in high-dimensional settings fall into roughly three categories. The first contains constraint-based methods, performing statistical tests Castelo and Roverato (2006); Drton and Perlman (2007, 2008); Kiiveri (2011); Wille and Bühlmann (2006). However, they either suffer from the excessive computational burden Castelo and Roverato (2006); Wille and Bühlmann (2006) or strong assumptions Drton and Perlman (2007, 2008) that correspond to regimes never attained in real situations. The second of these categories is composed of Bayesian approaches, see for instance Dobra et al (2004); Jones et al (2005); Rau et al (2012); Schwaller et al (2015). However, constructing priors on the set of concentration matrices is not a trivial task and the use of MCMC procedures limits the range of applications to moderate-sized networks. The third category contains regularized estimators, which add a penalty term to the likelihood in order to reduce the complexity or degrees of freedom of the estimator and more generally regularize the problem: throughout this chapter we focus on ℓ_1-regularized procedures, which are freed from any test procedure – and thus multiple testing issues – since they directly perform estimation and selection of the most significant edges by zeroing entries in the estimator of Θ. The remainder of this section is dedicated to a quick review of the state-of-the-art methods of this kind.

2.2 Sparse methods for GGM inference

The idea underlying sparse methods for GGM is the same as for the Lasso in linear regression (Tibshirani, 1996): it basically uses ℓ_1-regularization as a convex surrogate of the ideal but computationally intensive ℓ_0-regularized problem:

$$\underset{\Theta \in \mathscr{S}_p^+}{\arg\max} \quad \log\det(\Theta) - \mathrm{Tr}(\Theta \mathbf{S}_n) - \lambda \, \|\Theta\|_{\ell_0}. \tag{3}$$

Problem (3) achieves a trade-off between the maximization of the likelihood and the sparsity of the graph within a single optimization problem. The penalty term can also be interpreted as a log prior on the coefficients in a Bayesian perspective. BIC or AIC criteria are special cases of such ℓ_0 regularized problems, except that the maximization is made upon a restricted subset of candidates $\{\tilde{\Theta}_1, \ldots, \tilde{\Theta}_m\}$ and the choice of λ is fixed ($\log(n)$ for BIC and $1/2$ for AIC). Actually solving (3) would require the exploration of all possible 2^p graphs. On the contrary, by preserving the convexity of the optimization problem, ℓ_1-regularization paves the way to fast algorithms. For the price of a little bias on all the coefficients, we get to shrink some coefficients to exactly 0, operating selection and estimation in one single step as hoped in Problem (3).

Graphical-Lasso. The criterion optimized by the graphical-Lasso was simultaneously proposed in Yuan and Lin (2007) and Banerjee et al (2008). It corresponds to the estimator obtained by fitting the ℓ_1-penalized Gaussian log-likelihood, *i.e.* the tightest convex relaxation of (3):

$$\widehat{\Theta}_\lambda^{\text{glasso}} = \arg\max_{\Theta \in \mathscr{S}_p^+} \log\det(\Theta) - \text{Tr}(\Theta \mathbf{S}_n) - \lambda \, \|\Theta\|_{\ell_1}. \tag{4}$$

In this regularized problem, the ℓ_1-norm drives some coefficients of Θ to zero. The non-negative parameter λ tunes the global amount of sparsity: the larger the λ, the fewer edges in the graph. A large enough penalty level produces an empty graph. As λ decreases towards zero, the estimated graph tends towards the saturated graph and the estimated concentration matrix tends towards the usual MLE (2). By construction, this approach guarantees a well-behaved estimator of the concentration matrix *i.e.* sparse, symmetric and positive-definite, which is a great advantage of this method.

Ever since Criterion (4) was proposed, many efforts have been dedicated to developing efficient algorithms for its optimization. In the original proposal of Banerjee et al (2008), it is shown that solving for one row of matrix Θ in (4) while keeping other rows fixed boils down to a Lasso problem. The global problem is solved by cycling over the matrix rows until convergence. Thus, if one considers that L passes over the whole matrix are needed to reach convergence, a rough estimation of the overall cost is of the order of $Lp \times$ (cost for solving for one row). With a block-coordinate update each iteration over a row has $\mathscr{O}(p^3)$ complexity and their implementation is $\mathscr{O}(Lp^4)$ for L sweeps over the whole matrix $\widehat{\Theta}$. In Banerjee et al (2008) again, a rigorous analysis is conducted in Nesterov's framework Nesterov (2005) showing that the complexity for a single λ reaches $\mathscr{O}(p^{4.5}/\varepsilon)$ where ε is the desired accuracy of the final estimate.

The *Graphical-Lasso* algorithm of Friedman et al (2008) follows the same line but builds on a coordinate descent algorithm to solve each underlying Lasso problem. While no precise complexity analysis is possible with these methods, empirical results tend to show that this algorithm is faster than the original proposal of Banerjee et al (2008). Additional insights on the convergence of the graphical-Lasso are

provided in Mazumder and Hastie (2012), simultaneously with Witten et al (2011), showing how to take advantage of the problem sparsity by decomposing (4) into block diagonal problems depending on λ: this considerably reduces the computational burden in practice. Implementations of the graphical-Lasso algorithm are available in the R-packages **glasso**, **huge** (Zhao et al, 2014), or **simone** (Chiquet et al, 2009). The most recent notable efforts related to the optimization of (4) are due to Hsieh et al (2014, 2013) and the QUIC (then BIG&QUIC) algorithm, a quadratic approximation which allows (4) to be solved up to $p = 1,000,000$ with a superlinear rate of convergence and with bounded memory. The R-package **quic** implements the first version of this algorithm.

On the statistical side, the most striking results are due to Ravikumar et al (2011): they show that selection consistency of the estimator defined by (4) – that is, recovery of the true underlying graphical structure –, is met in the sub-Gaussian case when, for an appropriate choice of λ, the sample size n is of the same order as $\mathcal{O}(d^2 \log(p))$, where d is the highest degree in the target graph. Additional conditions on the empirical covariance between relevant and irrelevant features are required, known as the "irrepresentability conditions" in the Lasso case. Such statistical results are important since they provide insights on the "data" situations where such methods may either be successful or completely hopeless. More on this is discussed in Verzelen (2012). For instance, this should prevent blindly applying the graphical-Lasso in situations where the sample size n is too small compared to p. Similarly, when the presence of hub nodes with high degree is suspected, the estimated graph should be interpreted with care.

Neighborhood selection. This approach, proposed in Meinshausen and Bühlmann (2006), determines the graph of conditional dependencies by solving a series of p independent Lasso problems, successively estimating the neighborhoods of each variable and then applying a final reconciliation step as post-treatment to recover a symmetric adjacency matrix. Concretely, a given column $\mathbf{X}_j$ of the data matrix is "explained" by the remaining columns $\mathbf{X}_{\setminus j}$ corresponding to the remaining variables: the set $\mathrm{ne}(j)$ of neighbors of variable j in the graph $\mathscr{G}$ is estimated by the support of the vector solving

$$\hat{\beta}_j = \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min} \frac{1}{2n} \left\| \mathbf{X}_j - \mathbf{X}_{\setminus j} \beta \right\|_{\ell_2}^2 + \lambda \, \|\beta\|_{\ell_1}. \tag{5}$$

Indeed, if each row of $\mathbf{X}$ is drawn from a multivariate Gaussian $\mathscr{N}(\mathbf{0}, \Theta^{-1})$, then the best linear approximation of $\mathbf{X}_j$ by $\mathbf{X}_{\setminus j}$ is given by

$$\mathbf{X}_j = \sum_{k \in \mathrm{ne(j)}} \beta_{jk} \mathbf{X}_k = - \sum_{k \in \mathrm{ne(j)}} \frac{\Theta_{jk}}{\Theta_{jj}} \mathbf{X}_k, \tag{6}$$

thus coefficients β_j and column Θ_j – once its diagonal elements are removed – share the same support. By support, we mean the set of nonzero coefficients. Adjusting (5) for each $j = 1, \ldots, p$ allow us to reconstruct the full graph $\mathscr{G}$. Because

the neighborhoods of the p variable are selected separately, a post symmetrization must be applied to manage inconsistencies between edge selections; Meinshausen and Bühlmann (2006) suggests AND or OR rules.

Let us fill the gap with Criterion (4). First, note that the p regression problem can be rewritten as a unique matrix problem, where $\mathbf{B}$ contains p vectors $\beta_j, j = 1, \ldots, p$:

$$\hat{\mathbf{B}}^{\text{ns}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times p}, \text{diag}(\mathbf{B})=\mathbf{0}_p}{\arg\min} \frac{1}{2}\text{Tr}(\mathbf{B}^\top \mathbf{S}_n \mathbf{B}) - \text{Tr}(\mathbf{B}^\top \mathbf{S}_n) + \lambda \|\mathbf{B}\|_{\ell_1}. \tag{7}$$

In fact, it can be shown Rocha et al (2008); Ambroise et al (2009); Ravikumar et al (2010) that the optimization problem (7) corresponds to the minimization of a penalized, negative *pseudo*-likelihood: the joint distribution of X is approximated by the product of the p distributions of the p variables conditional on the other ones, that is

$$\log \mathbb{P}(\mathbf{X}; \Theta) = \sum_{j=1}^{p} \sum_{i=1}^{n} \log \mathbb{P}(X_j^i | X_{\setminus j}^i; \Theta_j).$$

This pseudo-likelihood is based upon the (false) assumption that conditional distributions are independent. Moreover, all variables are assumed to share the same variance in this formulation. Building on these remarks, Rocha et al (2008) amend criterion (7) by the adjunction of an additional symmetry constraint, and introduce additional parameters to account for different variances between the variables.

Concerning the computational aspect, this approach has very efficient implementation as it basically boils down to solving p Lasso problems. Suppose for instance that the target neighborhood size is k per variable: fitting the whole solution path of a Lasso problem using the Lars algorithm can be done in $\mathcal{O}(npk)$ complexity Bach et al (2012). This must be multiplied by p for the whole network, yet we underline that a parallel implementation is straightforward in this case. This makes this approach quite competitive, especially when coupled with additional bootstrap or resampling techniques Meinshausen and Bühlmann (2010).

On the statistical side, neighborhood selection has been reported to be sometimes empirically more accurate in terms of edge detection than the graphical-Lasso Villers et al (2008); Rocha et al (2008) on certain types of data. This is somewhat supported by the statistical analysis of Ravikumar et al (2011), who show that under the classical irrepresentability conditions for the Lasso Zhao and Yu (2006); Meinshausen and Bühlmann (2006) and for an appropriate choice of λ, neighborhood selection achieves selection consistency with high probability when the sample size n is of the order of $\mathcal{O}(d \log(p))$ with d the maximal degree of the target graph $\mathcal{G}$. This is to be compared with the $\mathcal{O}(d^2 \log(p))$ required by the graphical-Lasso (even if the corresponding "irrepresentability conditions" are not strictly comparable). A rough explanation for this difference on the asymptotic is that the graphical-Lasso intends to estimate the concentration matrix on top of selecting the nonzero entries, while neighborhood selection focuses on the selection problem.

Model selection issues. Up to this point, we have completely avoided the fundamental model selection issue, that is, the choice of the tuning parameter λ, which is at play in all the sparse methods mentioned thus far. The first possibility is to rely on information criteria of the form

$$\mathrm{IC}_\lambda = -2\mathrm{loglik}(\widehat{\Theta}_\lambda; \mathbf{X}) + \mathrm{pen}(\mathrm{df}(\widehat{\Theta}_\lambda)),$$

where "pen" is a function penalizing the model complexity, described by df, the degrees of freedom of the current estimator. We meet the AIC by choosing $\mathrm{pen}(x) = 2x$ and the BIC by choosing $\mathrm{pen}(x) = \log(n)x$. However, AIC and BIC are based upon assumptions which are not suited to high-dimensional settings (see Giraud et al, 2012b). Moreover, the notion of degrees of freedom for sparse methods has to be specified, not to mention that one has to adapt these criteria to the case of GGMs. An example of a criterion meeting these prerequisites is the extended BIC for sparse GGMs (Foygel and Drton, 2010):

$$\mathrm{EBIC}_\gamma(\widehat{\Theta}_\lambda) = -2\mathrm{loglik}(\widehat{\Theta}_\lambda; \mathbf{X}) + |\mathscr{E}_\lambda|(\log(n) + 4\gamma\log(p)), \tag{8}$$

where the function df is equal to $|\mathscr{E}|$, the total number of edges in the inferred graph. The parameter $\gamma \in [0,1]$ is used to adjust the tendency of the usual BIC – recovered for $\gamma = 0$ – to choose overly dense graphs in the high-dimensional setting. Further justification can be found in Foygel and Drton (2010). A competing approach, designed to compare a family of GGM – possibly inferred with different methods –, is GGMSelect (Giraud et al, 2012a; Giraud, 2008).

Another possibility is to rely on resampling/subsampling procedures to select a set of edges which are robust to small variations of the sample. The most popular approach is the *Stability Selection* procedure proposed in Meinshausen and Bühlmann (2010), also related to the bootstrapped procedure of Bach (2008). A similar approach, called StaRS (Stability approach to Regularization Selection) is developed specifically in the context of GGM in Liu et al (2010). The basic idea is as follows: for a given range of the tuning parameter $\Lambda = [\lambda_{\min}, \lambda_{\max}]$, the same method is fitted on many subsamples (with or without replacement) with size, say $n/2$. The idea is then to construct a score indexed on Λ that measures stability – or instability – of the selected variables. The selected edges are those matching a given score, for which the probability of false discovery is controlled. This requires an additional threshold in place of a choice of λ, but the authors in Meinshausen and Bühlmann (2010); Liu et al (2010) claim that such a threshold is typically much less sensitive than is the tuning parameter λ. An application of such resampling techniques to the inference of biological networks has been pursued with success in Haury et al (2012), advocating for the use of stability methods on real problems.

3 Accounting for multiscale data: multiattribute GGM

We now place ourselves in the situation where, for our collection of features $\mathscr{P}$, we observe not one but several attributes. The question at hand remains the same, that is to say, unraveling strong interactions between these features according to the observation of their attributes. Such networks are known as "association networks", which are systems of interacting elements, where a link between two different elements indicates a sufficient level of similarity between element attributes. In this section, we are interested in reconstructing such networks based upon n observations of a set of K attributes of the p elements composing the vertices of the network. To this end, we propose a natural generalization of sparse GGM to sparse *multiattribute* GGMs.

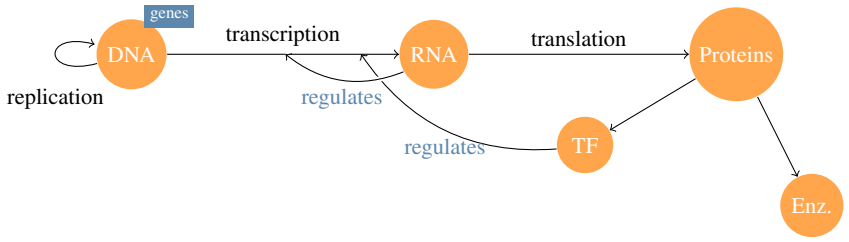

Fig. 1 Basic example of a multiattribute network in genomics: activity of a gene can be measured at the transcriptomic and proteomic levels, and gene regulation affected accordingly (TF = Transcription Factor; Enz. = Enzyme).

Why multiattribute networks? The need for multiattribute networks is relevant in many application fields, but seems particularly applicable in genomics. Indeed, with the plurality of emerging technologies and sequencing techniques, it is possible to record many signals related to the same set of biological features at various scales or locations of the cell. Consider for instance the simplifying – still hopefully didactic – central dogma of molecular biology, sketched in Figure 1: basically, expression of a gene encoding for a protein can be measured either at the transcriptome level, in terms of its quantity of RNA, or at the protein level, in terms of the concentration of the associated protein. Still, different technologies are used to measure either the transcriptome or the proteome, typically, microarray or sequencing technology for gene expression levels and cytometric or immunofluorescence experiments for protein concentrations. Although these signals are very heterogeneous (different levels of noise, count vs. continuous data, etc.), they do share commonality as they undergo common biological processes. We then put an edge in the network if it is supported in both spaces (gene and protein spaces). Our hope is that molecular profiles combined on the same set of biological samples can be *synergistic*, in order to identify a "consensus" and hopefully more robust network.

Multiattribute GGM. Let $\mathscr{P} = \{1,\dots,p\}$ be a set of variables of interest, each of them having some K attributes. Consider the random vector $X = (X_1,\dots,X_p)^\intercal$ such as $X_i = (X_{i1},\dots,X_{iK})^\intercal \in \mathbb{R}^K$ for $i \in \mathscr{P}$. The vector $X \in \mathbb{R}^{pK}$ describes the K recorded signals for the p features. We assume that X is a multivariate centered Gaussian vector, that is, $X \sim \mathscr{N}(\mathbf{0},\Sigma)$, with covariance and concentration matrices defined block-wise

$$\Sigma = \begin{bmatrix} \Sigma_{11} & & \Sigma_{1p} \\ & \ddots & \\ \Sigma_{p1} & & \Sigma_{pp} \end{bmatrix}, \; \Theta = \begin{bmatrix} \Theta_{11} & & \Theta_{1p} \\ & \ddots & \\ \Theta_{p1} & & \Theta_{pp} \end{bmatrix}, \quad \Sigma_{ij}, \Theta_{ij} \in \mathscr{M}_{K,K}, \; \forall (i,j) \in \mathscr{P}^2,$$

where $\mathscr{M}_{a,b}$ is the set of real-valued matrices with a rows, b columns. Such a multiattribute framework has been studied in Katenka and Kolaczyk (2012) with a reconstruction method based upon canonical correlations in order to test dependencies between pairs (i,j) at the attribute level using covariance. Here, we propose to rely on partial correlations in a multivariate framework rather than (canonical) correlations to describe relationships between the features, and thus extend GGM to a multiattribute framework. The objective is to define a "canonical" version of partial correlations. In our setting, the target network $\mathscr{G} = (\mathscr{P},\mathscr{E})$ is defined as the multivariate analog of the conditional graph for univariate GGM, that is

$$(i,j) \in \mathscr{E} \Leftrightarrow \Theta_{ij} \neq \mathbf{0}_{KK}, \quad \forall i \neq j. \tag{9}$$

In words, there is no edge between two variables i and j when their attributes are all conditionally independent.

A multivariate version of neighborhood selection. Our idea for performing sparse multiattribute GGM inference is to define a multivariate analog of the neighborhood selection approach Meinshausen and Bühlmann (2006) (see Section 2.2, Equations (5) and (7)). Indeed, it seems to be the most natural and convenient setup toward multivariate generalization. Nevertheless, other sparse GGM inference method like the graphical-Lasso (4) should have an equivalent multiattribute version. A possibility is explored in Kolar et al (2014) for instance.

To extend the neighborhood selection approach to a multiattribute version, we look at the multivariate analog of equation (6): in a multivariate linear regression setup, it is a matter of straightforward algebra to see that the conditional distribution of $X_j \in \mathbb{R}^K$ on the other variables is

$$X_j \,|\, X_{\setminus j} = x \sim \mathscr{N}(-\Theta_{jj}^{-1}\Theta_{j\setminus j}x, \Theta_{jj}^{-1}) \;\; .$$

Equivalently, letting $\mathbf{B}_j^T = -\Theta_{jj}^{-1}\Theta_{j\setminus j}$, one has

$$X_j \,|\, X_{\setminus j} = \mathbf{B}_j^T X_{\setminus j} + \varepsilon_j \quad \varepsilon_j \sim \mathscr{N}(\mathbf{0}, \Theta_{jj}^{-1}), \quad \varepsilon_j \perp X,$$

where $\mathbf{B}_j \in \mathscr{M}_{(p-1)K,K}$ is defined block-wise

$$\mathbf{B}_j = \begin{bmatrix} \mathbf{B}_j^{(1)} \\ \hline \vdots \\ \hline \mathbf{B}_j^{(p-1)} \end{bmatrix} = \Theta_{jj}^{-1} \times \begin{bmatrix} \Theta_j^{(1)} \\ \hline \vdots \\ \hline \Theta_j^{(p-1)} \end{bmatrix},$$

and where each $\mathbf{B}_j^{(i)}$ is a $K \times K$ matrix which links attributes of variables (i, j). We see that recovering the support of $\mathbf{B}_j$ block-wise is equivalent to reconstructing the network defined in (9). Estimation of $\mathbf{B}_j$ is thus typically achieved through sparse methods. To this end, we consider an i.i.d. sample $\{X^\ell\}_{\ell=1}^n$ of X such that each attribute is observed n times for the p variable, each X^n being a pK-size row vector staked in a $\mathscr{M}_{n,pK}$ data matrix $\mathbf{X}$, so that $\mathbf{X}_j \in \mathscr{M}_{n,K}$ is a real-value, $n \times K$ block matrix containing the data related to the jth variable:

$$\mathbf{X} \quad = \quad \begin{bmatrix} \mathbf{x}^1 \\ \hline \vdots \\ \hline \mathbf{x}^N \end{bmatrix} \quad = \quad \begin{bmatrix} \mathbf{X}^1 \ \ldots \ \mathbf{X}^p \end{bmatrix} \quad = \quad \left[\begin{array}{cc|c|cc} X_1^{11} & X_1^{1K} & \ldots & X_1^{p1} & \ldots X_1^{pK} \\ \hline \vdots & \vdots & \ldots & & \\ \hline X_n^{11} & X_n^{1K} & \ldots & X_n^{1K} & \ldots X_n^{pK} \end{array}\right].$$

Using these notations, a direct generalization of the neighborhood selection is to predict for each $j = 1, \ldots, p$ the data block $\mathbf{X}_j$ by regressing on $\mathbf{X}_{\setminus j}$. In matrix form, this can be written as the optimization problem

$$\arg\min_{\mathbf{B}_j \in \mathbb{R}} J(\mathbf{B_j}), \qquad J(\mathbf{B}_j) = \frac{1}{2n} \left\| \mathbf{X}_j - \mathbf{X}_{\setminus j} \mathbf{B}_j \right\|_F^2 + \lambda \ \Omega(\mathbf{B}_j), \tag{10}$$

where $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$ is the Frobenius norm of matrix $\mathbf{A}$ and Ω is a penalty which constrains $\mathbf{B}_j$ block-wise.

Choosing a penalizer. Various choices for Ω in (10) seem relevant: by simply setting $\Omega_0(A) = \sum_{i,j} |A_{i,j}|$, we just encourage sparsity among the $\mathbf{B}_i$ and thus do not couple the attributes. A clever choice would be to activate a set of attributes all together: hence, the group is defined by all the K attributes between variables i and j, therefore the penalizer turns to a group-Lasso like penalty

$$\Omega_1(\mathbf{B}_j) = \sum_{i \in \mathscr{P} \setminus j} \|\mathbf{B}_j^{(i)}\|_F, \tag{11}$$

in which case convex analysis and subdifferential calculus (see Boyd and Vandenberghe (2006)) can be used to show that a $\mathbf{B}_i$ is optimal for Problem (10) iff

$$\begin{cases} \forall i : \mathbf{B}_j^{(i)} \neq 0, & \left(\mathbf{S}_{ij} + \frac{\lambda}{\|\mathbf{B}_j^{(i)}\|_F} I \right)^{-1} \mathbf{S}_{ij} = \mathbf{B}_j^{(i)} \\ \forall i : \mathbf{B}_j^{(i)} = \mathbf{0}_{KK}, & \|\mathbf{S}_{ij}\|_F \leq \lambda \end{cases}, \tag{12}$$

where $\mathbf{S}_{ij} \in \mathscr{M}_{KK}$ is a $K \times K$ block in the empirical covariance matrix $\mathbf{S}_n = n^{-1}\mathbf{X}^\top\mathbf{X}$, which shows the same block-wise decomposition as Σ or Θ. This paves the way for an optimization algorithm like block-coordinate descent which we implemented, although we omit details here.

4 Numerical experiments

Simulation study. We propose a simple simulation to illustrate the interest of using multiattribute networks. The simulations are set up as follows:

1. Draw a random undirected network with p nodes from the Erdös-Renyi model with adajacency matrix $\mathbf{A}$;
2. Expand the associated adjacency matrix to multivariate space with

$$\mathbf{M} = \mathbf{A} \otimes \mathbb{S} + \mathbf{I}_{p \times K}$$

   where $\otimes$ is the Kronecker product. The $K \times K$ matrix $\mathbb{S}$ is used to consider different scenarios of agreement across the attributes of two genes. We consider three cases

   a. $\mathbb{S} = \mathbf{I}_{K,K}$ the $K \times K$ identity matrix: same intra-attribute network and no inter-attribute interactions;
   b. $\mathbb{S} = \mathbf{I}_{K,K} - \mathbf{1}_{K,K}$, same inter-attribute interactions and no intra-attribute interactions;
   c. $\mathbb{S} = \mathbf{1}_{K,K}$ a matrix full of one: full agreement between attributes.

3. Compute Θ a positive definite approximation of $\mathbf{M}$ by replacing null and negative eigenvalues by a small constant;
4. Control the difficulty of the problem with $\gamma > 0$ such that $\Theta = \Theta + \gamma I$;
5. Draw an i.i.d. n-size sample $\mathbf{X} \in \mathbb{R}^{n \times pK}$ of $X \sim \mathscr{N}\left(0, \Theta^{-1}\right)$.

We choose small networks with $p = 40$, with 40 edges on average and vary n from $p/2$ to $2p$. We fix γ to 0.1 and consider cases where the number of attributes is $K = 2, 3$ and 4. We compare our multiattribute approach of neighborhood selection to two baselines:

1. the standard neighborhood selection procedure applied on the data related to each attribute separately: to do so, we separate $\mathbf{X}$ in K data sets $\mathbf{X}^{(1)}, \ldots \mathbf{X}^{(K)}$ all with size $\mathbb{R}^{n \times p}$ and reconstruct one network per attribute. We refer to this method as the *separate* variant.
2. the standard neighborhood selection approach applied on a merge data set, obtained by stacking the data sets $\mathbf{X}^{(1)}, \ldots \mathbf{X}^{(K)}$ of each attribute into a single $\tilde{\mathbf{X}}$ data set in $\mathbb{R}^{nK \times p}$. We refer to this method as the *merge* variant. This method is the exact opposite of the *separate* variant.

We assess the performances of each method in reconstructing the original adjacency matrix $\mathbf{A}$ with the area under ROC curve (AUC). For the *separate* variant, the re-

tained AUC is the AUC averaged over all attributes. We replicate the experiment 100 times.

On Figure 2, it is clear that aggregation (either by merging data sets or with multiattribute network inference) improves upon a single-attribute approach. Even when there is no inter-attributes interactions, (which is barely meaningful towards application to regulatory networks), in which case it is a very good idea to merge the problems together to increase the sample size, our multiattribute approach remains quite competitive and robust. In all other cases, it outperforms the competing approaches.

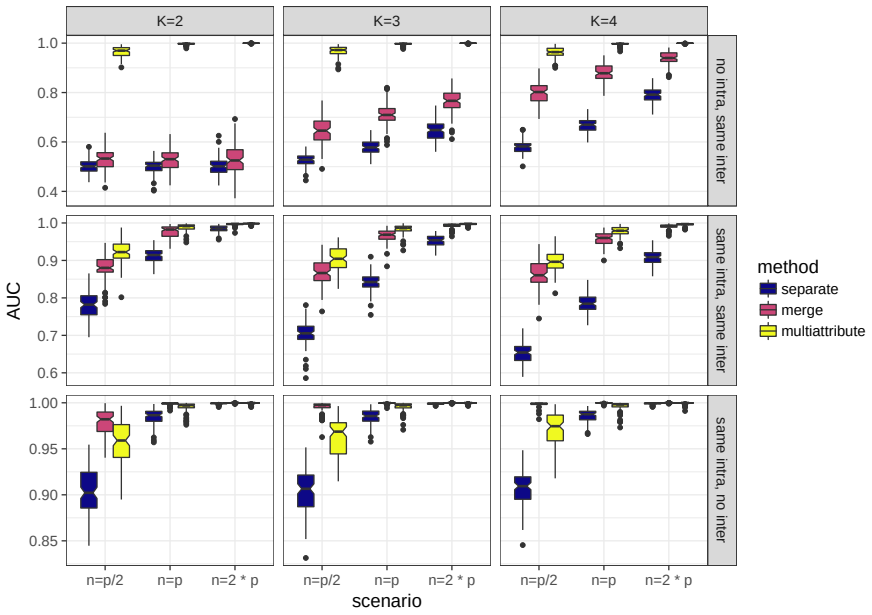

Fig. 2 Simple simulation study for the multiattribute network inference problem: the multiattribute procedure improves over the univariate procedures in every situation when networks are close for each attribute.

Illustration: Gene/Protein regulatory network inference. As an illustration, we applied our sparse multiattribute GGM approach to reconstruct networks on two large breast cancer data sets from the National Cancer Institute[1] and the Rational Therapy for Breast Cancer consortium[2], respectively referred to as NCI-60 and RATHER hereafter. These data sets contain both proteomic and transcriptomic profiles, respectively measured with reverse-phase protein arrays (RPPA) and RNA

[1] https://www.cancer.gov/

[2] http://www.ratherproject.com/

affymetrix array. We infer the multiattribute network between the subset of molecular entities which is common to the proteins measured by RPPA and the genes measured by RNA array, that we call the *consensus set*: in the NCI-60 cancer line data set (Pfister et al, 2009), a consensus set composed of $p = 91$ protein and corresponding gene profiles is retained, for the $n = 60$ samples. The RATHER data set (Michaut et al, 2016) contains proteomic and transcriptonic data from $n = 100$ patients for a consensus set of $p = 117$ entities[3].

We infer a sparse GGM for each attribute (gene expression and protein profile), separately to start with, and then get its multiattribute version. We do this on a large grid of the tuning parameter and thus have three families of networks indexed by their number of edges.

Figure 3 demonstrates that our sparse multiattribute method captures the characteristics of both univariate networks, as the Jaccard similarity index is high between each uni-attribute network and the multiattribute network, while it remains low when comparing uni-attribute networks together.

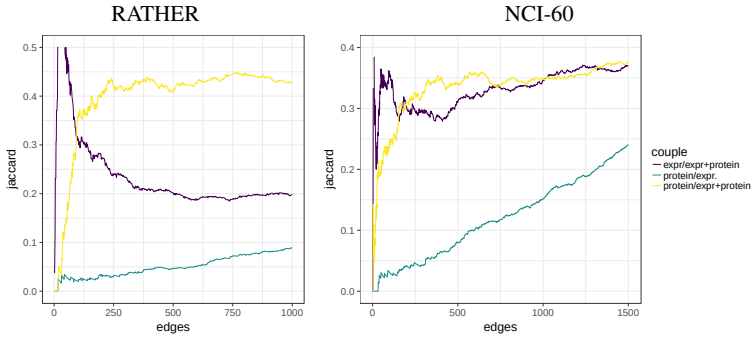

Fig. 3 Jaccard's similarity index $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ between uni-attribute and multiattribute networks, for RATHER and NCI60 data set: multiattribute networks share a high Jaccard index with both uni-attribute networks.

Figure 4 shows the finally retained networks, where the number of edges is controlled by the tuning parameter λ chosen by 10-fold cross-validation. It is clear that some motifs only present in each uni-attribute networks are caught in their multiattribute counterparts. This tends to prove that the multiattribute version proposes a consensus version of the interactions at hand in the cell, and one which is hopefully more robust to noise.

[3] The data can be downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66647.

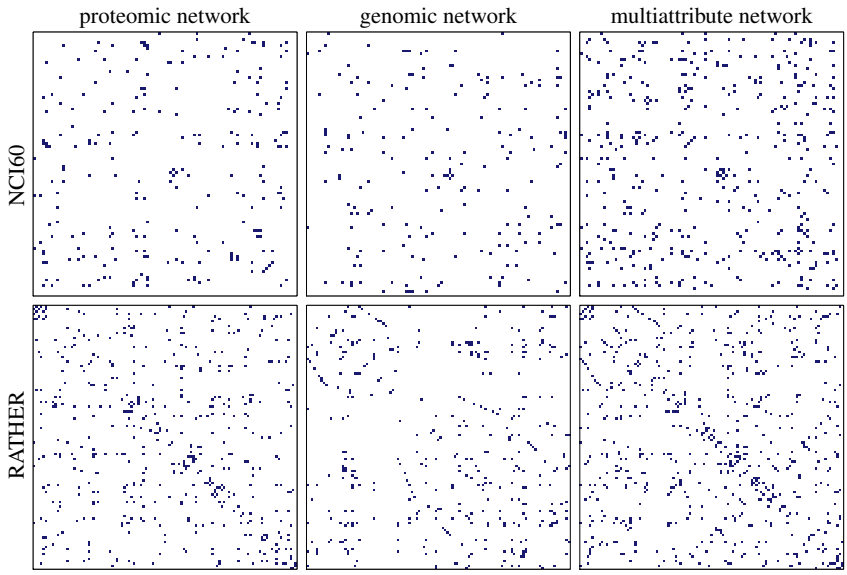

Fig. 4 Uni-attribute and multiattribute networks inferred on both NCI60 and RATHER dataset. The number of neighbors of each entity is chosen by cross-validation. Multiattribute networks catch motif found in the uniattribute counterparts.

References

Akbani R, Becker KF, Carragher N, Goldstein T, de Koning L, Korf U, Liotta L, Mills GB, Nishizuka SS, Pawlak M, et al (2014) Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: A workshop report the rppa (reverse phase protein array) society. Molecular & cellular proteomics 13(7):1625–1643

Ambroise C, Chiquet J, Matias C (2009) Inferring sparse Gaussian graphical models with latent structure. Electronic Journal of Statistics 3:205–238

Bach F (2008) Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th international conference on Machine learning, ACM, pp 33–40

Bach F, Jenatton R, Mairal J, Obozinski G (2012) Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning 4(1):1–106

Banerjee O, El Ghaoui L, d'Aspremont A (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J Mach Learn Res 9:485–516

Boyd S, Vandenberghe L (2006) Convex Optimization, 3rd edn. Cambridge University Press

Cai T, Liu W, Luo X (2011) A constrained l1 minimization approach to sparse precision matrix estimation. J Amer Statist Assoc 106:594–607

Castelo R, Roverato A (2006) A robust procedure for Gaussian graphical model search from microarray data with p larger than n. J Mach Learn Res 7:2621–2650

Chiquet J, Smith A, Grasseau G, Matias C, Ambroise C (2009) SIMoNe: Statistical Inference for MOdular NEtworks. Bioinformatics 25(3):417–418, URL http://dx.doi.org/10.1093/bioinformatics/btn637

Chiquet J, Grandvalet Y, Ambroise C (2011) Inferring multiple graphical models. Statistics and Computing 21(4):537–553, URL http://dx.doi.org/10.1007/s11222-010-9191-2

Dempster A (1972) Covariance selection. Biometrics, Special Multivariate Issue 28:157–175

Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M (2004) Sparse graphical models for exploring gene expression data. J Multivariate Anal 90(1):196–212

Drton M, Perlman M (2007) Multiple testing and error control in Gaussian graphical model selection. Statistical Science 22:430

Drton M, Perlman M (2008) A SINful approach to Gaussian graphical model selection. J Statist Plann Inference 138(4):1179–1200

Foygel R, Drton M (2010) Extended Bayesian information criteria for Gaussian graphical models. In: Advances in Neural Information Processing Systems (NIPS), pp 2020–2028

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

Giancotti FG (2014) Deregulation of cell signaling in cancer. FEBS letters 588(16):2558–2570

Giraud C (2008) Estimation of Gaussian graphs by model selection. Electronic Journal of Statistics 2:542–563

Giraud C, Huet S, Verzelen N (2012a) Graph selection with GGMselect. SAGMB 11(3):1–50

Giraud C, Huet S, Verzelen N (2012b) High-dimensional regression with unknown variance. Statist Sci 27(4):500–518

Haury AC, Mordelet F, Vera-Licona P, Vert JP (2012) Tigress: trustful inference of gene regulation using stability selection. BMC systems biology 6(1):145

Hsieh CJ, Sustik M, Dhillon I, Ravikumar PK, Poldrack R (2013) Big & quic: Sparse inverse covariance estimation for a million variables. In: Advances in Neural Information Processing Systems (NIPS), pp 3165–3173

Hsieh CJ, Sustik M, Dhillon I, Ravikumar P (2014) Quic: quadratic approximation for sparse inverse covariance estimation. J Mach Learn Res 15(1):2911–2947

Jeanmoungin M, Charbonnier C, Guedj M, Chiquet J (2014) Probabilistic graphical models dedicated to applications in genetics, genomics and postgenomics, Oxford University Press, chap Network inference in breast cancer with Gaussian graphical models and extensions. URL http://ukcatalogue.oup.com/product/9780198709022.do

Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M (2005) Experiments in stochastic computation for high-dimensional graphical models. Statist Sci 20(4):388–400

Katenka N, Kolaczyk E (2012) Inference and characterization of multi-attribute networks with application to computational biology. Ann Appl Stat 6(3):1068–1094

Kiiveri H (2011) Multivariate analysis of microarray data: differential expression and differential connection. BMC Bioinformatics 12(1):42, DOI 10.1186/1471-2105-12-42, URL http://www.biomedcentral.com/1471-2105/12/42

Kolar M, Liu H, Xing E (2014) Graph estimation from multi-attribute data. J Mach Learn Res 15(1):1713–1750

Lauritzen S (1996) Graphical models, Oxford Statistical Science Series, vol 17. Clarendon Press, New York, oxford Science Publications

Liu H, Roeder K, Wasserman L (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. In: Advances in neural information processing systems (NIPS), pp 1432–1440

Mazumder R, Hastie T (2012) The graphical lasso: New insights and alternatives. Electron J Statist 6:2125–2149, DOI 10.1214/12-EJS740, URL https://doi.org/10.1214/12-EJS740

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Statist 34(3):1436–1462

Meinshausen N, Bühlmann P (2010) Stability selection. Journal of the Royal Statistical Society, Series B 72:417–473

Michaut M, Chin SF, Majewski I, Severson TM, Bismeijer T, de Koning L, Peeters JK, Schouten PC, Rueda OM, Bosma AJ, et al (2016) Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. Scientific reports 6:18,517

Nesterov Y (2005) Smooth minimization of non-smooth functions. Mathematical programming 103(1):127–152

Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, Parchment RE, Tomaszewski JE, Doroshow JH, Pommier Y (2009) Topoisomerase i levels in the nci-60 cancer cell line panel determined by validated elisa and microarray analysis and correlation with indenoisoquinoline sensitivity. Molecular cancer therapeutics 8(7):1878–1884

Rau A, Jaffrézic F, Foulley JL, Doerge R (2012) Reverse engineering gene regulatory networks using approximate Bayesian computation. Statistics and Computing 22(6):1257–1271

Ravikumar P, Wainwright MJ, Lafferty J (2010) High-dimensional Ising model selection using ℓ_1-regularized logistic regression. Ann Stat 38:1287–1319

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

Rocha G, Zhao P, Yu B (2008) A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE)

Schwaller L, Robin S, Stumpf M (2015) Bayesian inference of graphical model structures using trees. arXiv preprint arXiv:150402723

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Statist Soc B 58(1):267–288

Verzelen N (2012) Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. Elec Journal Stat 6:38–90

Villers F, Schaeffer B, Bertin C, Huet S (2008) Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. Stat Appl Genet Mol Biol 7(2)

Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley series in probability and mathematical statistics: Probability and mathematical statistics, Wiley

Wille A, Bühlmann P (2006) Low-order conditional independence graphs for inferring genetic networks. Stat Appl Genet Mol Biol 5(1)

Witten D, Friedman J, Simon N (2011) New insights and faster computations for the graphical lasso. J Comput Graph Statist 20(4):892–900

Yuan M (2010) Sparse inverse covariance matrix estimation via linear programming. J Mach Learn Res 11:2261–2286

Yuan M, Lin Y (2007) Model selection and estimation in the Gaussian graphical model. Biometrika 94(1):19–35

Zhao P, Yu B (2006) On model selection consistency of Lasso. J Mach Learn Res 7:2541–2563

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L (2014) huge: High-dimensional Undirected Graph Estimation. URL http://CRAN.R-project.org/package=huge, r package version 1.2.6