

Numerical studies on covariance selection problems

M. Sofia Massa

Abstract We investigate covariance selection problems in which joint data on a set of variables of interest are not available, though different datasets with overlapping variables are. In this context, the intent is in describing the overall statistical model by means of conditional independencies between all variables of interest. We show that, under particular assumptions of consistency, the combination of smaller Gaussian graphical models can be useful for problems of Gaussian graphical model selections. We present two simulation studies to compare the results of a covariance selection on all variables of interest and the combination of smaller covariance selections performed on fewer variables.

Key words: Gaussian undirected graphical models, model selection, structural meta-analysis.

1 Introduction

Structural meta-analysis, introduced by [9], can be defined as the combination of evidence about relationships between variables from studies or experiments carried out independently under partially comparable conditions. By relationships between variables we mean conditional independence relations, as defined by [1]. As an example, consider different investigations with some elements in common or different parts of the same experiment developed by different laboratories. In particular, the results of the investigations are in the form of conditional independence relations over some set of variables and some of these are overlapping between the studies. In more detail, we work with Gaussian undirected graphical models which are statistical models comprising conditional independencies between variables that are usu-

M. Sofia Massa

Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via Cesare Battisti, 241, 35121, Padova, Italy. e-mail: massa@stat.unipd.it

ally read off an undirected graph. In all these cases, methods to perform a structural meta-analysis are necessary. See [9] for an overview on combinations of statistical models and for all relevant definitions used in this paper.

Here, we intend to study the applicability of a structural meta-analysis in the context of model building. A Gaussian graphical model selection, also called covariance selection, is usually performed on a set of observations collected over some variables of interest with the intent of retrieving one possible model consistent with the data and its undirected graph. If the data come from different experiments, or from different parts of the same experiment, one could perform a combination of the results of different covariance selections. In this paper, we will compare the results of some model selection procedures for Gaussian graphical models over a set of variables with the composition of smaller model selections performed on fewer variables.

The paper is organized as follows. Section 2 gives some details about Gaussian graphical models and the particular type of combination that we exploit in this paper. Section 3 and 4 will present two simulation studies in different scenarios. Concluding remarks are given in Section 5.

2 Combinations of Gaussian graphical models

We start by giving some definitions that we will use throughout the paper. An *undirected graph* is a pair $G = (V, E)$, where V is a finite set of vertices and $E \subseteq V \times V$ is a finite set of edges between these vertices. If two vertices u and v are joined by an edge, u and v are said to be adjacent. The *boundary* of a subset $A \subset V$ is the set of all vertices in $V \setminus A$ that are adjacent to vertices of A . If $A \subset V$, an *induced subgraph* over A is $G_A = (A, E_A)$ where $E_A = E \cap (A \times A)$. A *path* is a sequence of vertices such that every vertex has an edge to the next vertex in the sequence. Two subsets $A \subset V$ and $B \subset V$ are *separated* by a set $C \subset V$ in G if every path from A to B intersects C . A *connected component* of a graph G is a subset of vertices $A \subset V$, such that any vertex in the subgraph G_A is connected by some path to any other vertex. A graph is *collapsible* onto a subset $A \subset V$ if and only if every connected component of $V \setminus A$ has a complete boundary.

Gaussian undirected graphical models are families of multivariate normal distributions that satisfy the global Markov property with respect to an undirected graph G . The global Markov property states that if two sets of vertices A and B are separated by a set of vertices C , then the variables Y_A and Y_B are conditionally independent given the variables Y_C . So if $G = (V, E)$ is an undirected graph and $|V| = p$, this is equivalent to define a Gaussian undirected graphical model as $Y_V \sim N_p(\mu, \Sigma)$ with $\Sigma^{-1} \in S^+(G)$, where $S^+(G)$ is the set of definite positive matrices that have null elements whenever there is a missing edge between the corresponding elements of V . The graph G is usually called dependence graph. In the following, we will only work with undirected graphs and corresponding Gaussian undirected graphical models, and will refer to them simply as graphs and Gaussian graphical models (see [7]).

Let $\mathcal{A}$ and $\mathcal{B}$ be two Gaussian graphical models with dependence graphs G_A and G_B , respectively, defined over two sets of variables A and B , such that $A \cap B \neq \emptyset$. In particular, consider only graphical models (i) with the same induced subgraphs over $A \cap B$ and (ii) such that they are both collapsible onto $A \cap B$. Here, the attention is only directed towards their meta-Markov combination [2]. If (i) and (ii) are both satisfied, [9] show that the meta-Markov combination of $\mathcal{A}$ and $\mathcal{B}$ is the graphical model with dependence graph given by the union of the two starting graphs.

Consider the following example. Let $\mathcal{A} = \{Y_A \sim N_5(0, \Omega), \Omega^{-1} \in S^+(G_A)\}$ and $\mathcal{B} = \{Y_B \sim N_3(0, \Phi), \Phi^{-1} \in S^+(G_B)\}$, be two Gaussian graphical models with graphs G_A and G_B , respectively, as in Figure 1. The two graphs may represent two

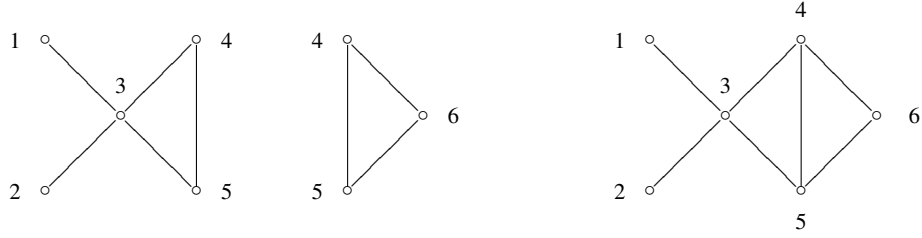


Fig. 1 On the left, two Gaussian graphical models with graphs G_A and G_B , respectively. On the right, the graph corresponding to their meta-Markov combination.

different studies with some variables in common, or parts of the same study, performed by two laboratories. Their meta-Markov combination has dependence graph shown on the right of Figure 1. Of course, the combination includes all the variables under examination and preserves the original conditional independence relationships between the subsets of starting variables.

In the above example, the two starting Gaussian graphical models are fixed and known. Suppose now to have different sets of observations on some variables of interest and that the intent is to build a reasonable model consistent with these data in terms of conditional independencies. We do not know anything about the statistical models by which the data were generated and we want to select the appropriate Gaussian graphical model and its dependence graph, consistent with all datasets at hand. Note that a model selection procedure designed for Gaussian graphical models permits to recover the estimate of the mean vector, the covariance matrix, and hence the edges of the dependence graph. We propose to achieve such construction by composing, through meta-Markov combinations, the results of Gaussian graphical models selections performed on the sets of observations provided by the studies. In order to be able to do so, both conditions (i) and (ii) need to hold. In this paper, we will restrict our attention to situations where this happens. More general situations will be addressed elsewhere.

In the following sections, we perform a covariance selection on a set of variables by dividing the problem into smaller covariance selections and by combining, the

results, or, in simpler words, by combining the dependence graphs given by each covariance selection. We touch this possibility upon by means of small examples. In the first study, we do not know the true graphical model underlying a set of observations provided for some variables. We compare a Gaussian graphical models combination on fewer variables and a direct Gaussian graphical models selection over all the variables by using a stepwise unrestricted model selection [3] and the graphical lasso [4]. In the second example, we fix the true graph and we look at the results of a combination given by a stepwise unrestricted model selection and a graphical model selection based on the BIC criterion [10]. Some extensive simulation studies related to the problem addressed in this paper and a direct comparison on the three model selection procedures can be found in [8].

3 Simulation study on real data

In this study, we examine the selection of a Gaussian graphical model over a set of variables and the composition of smaller Gaussian graphical models selections over fewer variables. In order to do so, we compare the dependence graph of all the variables and the dependence graph of the composition of the smaller model selections.

We use microarray data on rhabdomyosarcoma, a fast-growing, highly malignant soft tissue sarcoma in children, collected at the CRIBI biotechnology Center (University of Padova) in a study involving 1200 genes and 138 children. Here, we focus our attention on six genes (variables) selected by the biologists as relevant to address a particular research question. The aim is to build a reasonable network relating such genes.

We consider all six genes and also two subsets of three and four genes, respectively, with one gene in common. The dependence graphs are built by using two graphical models selection procedures, i.e., the unrestricted stepwise backward model selection [3] and the graphical lasso [4]. The former is implemented in the R package 'mimR' [6], whereas the latter is implemented in the R package 'glasso' [4]. As we said before, we will compare the dependence graph $\hat{G}$ corresponding to the graphical model selected over all six genes and the combination of the dependence graphs $\hat{G}_A$ and $\hat{G}_B$ corresponding to the two graphical models selected over the set of three and four genes, respectively.

It is worth noting that the use of graphical lasso requires the definition of a penalty parameter λ . The choice of the penalty to be used for estimating $\hat{G}$, $\hat{G}_A$ and $\hat{G}_B$ was made by performing a preliminary *ad hoc* analysis. We run the graphical lasso for the selection of the true graphs for values of the penalty parameter going from 0.01 to 2; then, we chose the value of λ that maximized the number of times that the true conditional independencies were identified. The penalization parameter of the graphical lasso for the estimation of G is chosen by looking at the graph in Figure 2, representing the number of edges selected for the penalty parameter varying from 0.01 to 1. The same figures are obtained for the selection of the

penalty parameters for G_A and G_B (figures not shown). Here, for the selection of G we set $\lambda = 0.09$, corresponding to a graph with six edges (about 46% of total edges), for the selection of G_A we set $\lambda = 0.1$, corresponding to a graph with four edges, and for the selection of G_B we set $\lambda = 0.12$, corresponding to a graph of two edges.

With the graphical lasso, the composition of the two smaller graphs gives the same results as the selection on all the variables (see Fig. 3), but this is not the case for the stepwise procedure (see Fig. 4). However, we remark that for the stepwise procedure, the edge between genes 3727 and 4086 can never be reconstructed by the composition of the two subnetworks. A selection procedure finds this edge only when both genes belong to the same set of variables. This suggests that the final network selected depends strongly on the chosen subsets of genes. One may think of using subject matter information to overcome this drawback. For example, prior knowledge about interactions between some of the genes, as well as evidence in favour of the presence or absence of an edge may be exploited.

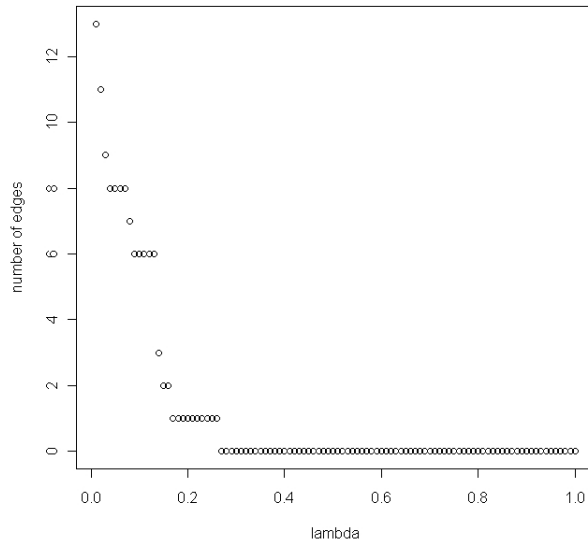


Fig. 2 Penalty parameters versus number of edges for the estimation of the graph G with 6 genes.

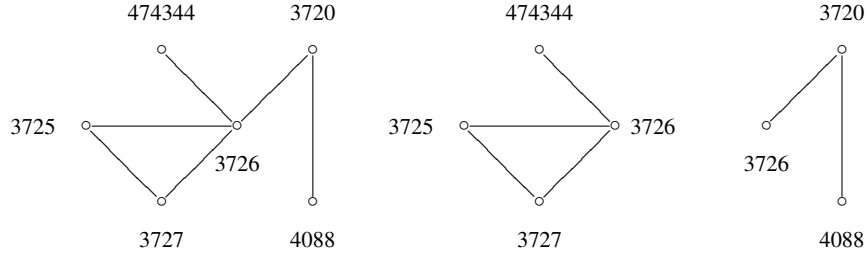


Fig. 3 Gaussian graphical models for 6 genes (on the left), for 4 genes (on the center), for 3 genes (on the right) selected using graphical lasso with $\lambda = 0.09$, $\lambda = 0.1$, $\lambda = 0.12$, respectively.

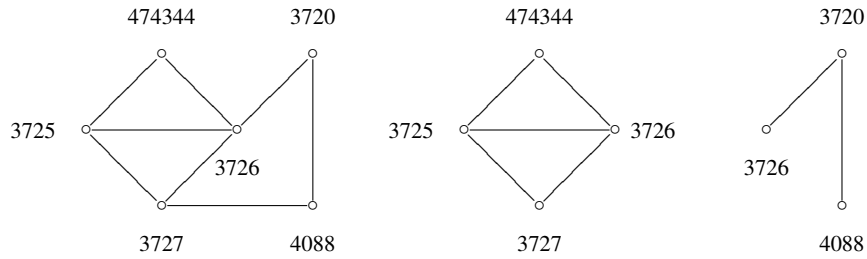


Fig. 4 Gaussian graphical models for 6 genes (on the left), for 4 genes (on the center), for 3 genes (on the right) selected using the stepwise procedure.

4 Simulation study on a real graph

In the previous example, we had data for all the variables of interest but nothing was known about the true graph. In this example, we fix the true graph for a certain number of variables, we generate a random sample from the Gaussian graphical model with such dependence graph, and we compare this graph with that retrieved by a composition of Gaussian graphical model selections on smaller sets of variables. We also increase the number of variables to investigate further their impact on the different model selection procedures.

To be more realistic on the choice of the graphical structure to study, we choose a subset of the human disease network presented in Fig. 2a of [5]. The chosen portion is based on the graph G shown in Fig. 5.

In the following, we take such structure to be the true structure to be learnt from the data. Now the graph G may be retrieved 1) by considering all variables at the same time; 2) by breaking down the problem into smaller problems.

With regard to the first point, the stepwise procedure detects a model with a highly connected graph, and the model selection procedure based on the BIC criterion cannot be used for its well known computational burden when searching among

2^{91} graphs. The glasso procedure performs very poorly, detecting the right amount of edges, but not the wanted one, as pointed out also in [8].

To the second aim, we divide the variables into 4 subsets of 5 elements that, pairwise, have some common elements. We consider therefore the following subsets: $\{a, b, c, d, h\}$, $\{c, d, e, f, g\}$, $\{c, h, i, l, m\}$, $\{l, m, n, o, p\}$. The true structures corresponding to these subsets are the subgraphs of G induced on the 4 subsets and are indicated as G_1 , G_2 , G_3 , G_4 , respectively. Their composition - obtained by combining two of them at a time - is equal to G .

We repeat 1000 times the following steps. We simulate 200 multivariate normal data coming from the Gaussian graphical model $\mathcal{M}(G) = \{Y \sim N_{14}(0, \Sigma), \Sigma \in S^+(G)\}$ and we record whether a model selection procedure finds the graphs G_1 , G_2 , G_3 , G_4 and whether their combination is exactly G . We count how many times a model selection procedure individuates the correct model and we indicate the number of time as $\#\{G_1\}$, $\#\{G_2\}$, $\#\{G_3\}$, $\#\{G_4\}$, $\#\{G_{\text{comb}}\}$, respectively. Only the stepwise [3] and model selection based on BIC criterion [10] procedures are taken into account. The BIC procedure is implemented in the R package 'mimR' [6].

From Table 1 we can see that the subgraphs are detected a high number of times. Moreover, the BIC procedure outperforms the stepwise procedure especially on the value of $\#\{G_{\text{comb}}\}$. However, if we consider only the stepwise procedure, even if the value of $\#\{G_{\text{comb}}\}$ is small, it is definitely better compared to the results that we would have collected by selecting a model on all the variables together. As we remarked before, the stepwise procedure applied on data consisting of many variables finds a highly connected graph. Therefore, even if the stepwise and BIC procedures cannot detect all the correct edges, the combination of small subgraphs may represent an helpful way of improving the selection of an appropriate model.

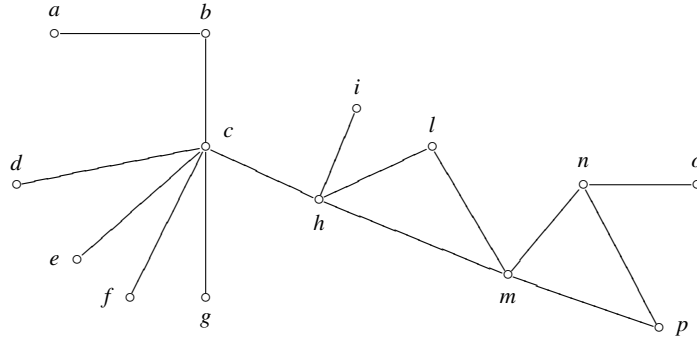


Fig. 5 Graph G considered in the study. It is a subset of the network of Figure 2a of [5].

Table 1 Results of the study that decomposes a model selection on many variables in smaller subselections. They refer to the graph G reported in Figure 5.

Method	$\#\{G_1\}$	$\#\{G_2\}$	$\#\{G_3\}$	$\#\{G_4\}$	$\#\{G_{\text{comb}}\}$
Stepwise	611	712	701	648	197
BIC	892	882	887	889	619

5 Conclusion

We performed some simulation studies in order to carry out a covariance selection by means of composition of smaller covariance selections. As combination of two Gaussian graphical models, we considered the meta-Markov combination [2] under the assumptions that the dependence graphs to be combined have the same induced subgraphs on the common variables and that they are both collapsible onto the common variables.

These small examples show that the combination may be considered as an useful approach to the problem of covariance selection. The results depended on the procedure used, in particular on its ability on detecting the most significant conditional independence relations, and on the chosen subsets of variables. For studies involving a small number of variables, a sensible strategy to be adopted before the combination may be that of applying more than one model selection procedure and compare the results obtained.

References

1. Dawid, A. P.: Conditional independence in statistical theory (with Discussion). *J. Roy. Statist. Soc. B* **41**, 1-31 (1979)
2. Dawid, A. P., Lauritzen, S. L.: Hyper-Markov laws in the statistical analysis of decomposable graphical models, *Ann. Statist.*, **21**, 1272–1317 (1993)
3. Edwards, D.: Introduction to graphical modelling. Springer-Verlag, New York (2000)
4. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441 (2008)
5. Goh, K-II., Cusick, M. E., Valle, D., Childs, B., Vidal, M., Barabasi, A-L.: The human disease network, *Proc. Nat. Acad. Sci.* **104**, 8685–8690 (2007)
6. Højsgaard, S.: The mimR package for graphical modelling in R. *J. Statist. Software*, **11**, 1–13 (2004).
7. Lauritzen, S.L.: Graphical models. Clarendon Press, Oxford (1996)
8. Massa, M.S., Chiogna, M.: Combinations of covariance selections for graphical modelling. Working Paper Series, n.2, Dipartimento di Scienze Statistiche, Università degli Studi di Padova (2009)
9. Massa, M.S., Lauritzen, S.L.: Combining Statistical models. In: Viana, M., Wynn, H.P. (eds.) *Algebraic methods in statistics and probability II*. AMS Contemporary Mathematics Series (2010). In corso di stampa
10. Schwarz, G.: Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464 (1978)