Working Paper Series, N. 2, February 2009

Combinations of covariance selections for graphical modelling

M. Sofia Massa, Monica Chiogna
Department of Statistical Sciences
University of Padua
Italy

Abstract: We explore the possibility of composing the results of a fixed number of Gaussian graphical model selections on some partially overlapping variables. This appears to be an useful approach in all the research areas where a large amount of data from different sources and types of experiments is available. Therefore the focus is in binding together information coming from heterogeneous studies to improve the understanding of a particular phenomenon of interest. The proposed approach relies on numerical results on artificial and real data.

Keywords: Gaussian graphical models, model selection, composition of graphical models.

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
M. Sofia Massa
tel: +39 049 827 4124
massa@stat.unipd.it
http://www.stat.unipd.it/~massa

Combinations of covariance selections for graphical modelling

M. Sofia Massa, Monica Chiogna
Department of Statistical Sciences
University of Padua
Italy

Abstract: We explore the possibility of composing the results of a fixed number of Gaussian graphical model selections on some partially overlapping variables. This appears to be an useful approach in all the research areas where a large amount of data from different sources and types of experiments is available. Therefore the focus is in binding together information coming from heterogeneous studies to improve the understanding of a particular phenomenon of interest. The proposed approach relies on numerical results on artificial and real data.

Keywords: Gaussian graphical models, model selection, composition of graphical models.

1 Introduction and aim

Motivated by the idea of performing statistical inference on some phenomenon by taking into account information coming from different sources, Massa (2008) and Massa and Lauritzen (2009) present a general framework for combining statistical models. Combining models allows to reach a deeper understanding of the phenomenon under study, especially when the sources of information are readily accessible.

Various applications motivate such theoretical developments. Recent advances in the field of genomics, for instance, have made available a large amount of different data and a great effort is concentrated in putting such information together. See, for example, Garret-Mayer *et al.* (2008), where three different types of datasets are combined together to improve the understanding of biological processes, and Goh *et al.* (2007), where a combination of networks representing different diseases with some recurrent genes is shown.

Combining models is a difficult task which involves many issues. Firstly, the models involved in a combination need to respect some form of compatibility. They should need some common elements (variables), but also the information that they provide must not be in conflict. Furthermore, one should be able to recover the initial information when restricting the attention (marginalizing) on the initial variables. Finally, one needs to exactly specify how the combination is performed. All these issues are addressed from a theoretical point of view in Massa and Lauritzen (2009).

Even though the approach presented in the above mentioned paper is quite general, the main application concerns the combination of Gaussian graphical models.

These are families of multivariate normal distributions respecting conditional independence relations between the variables that can be easily visualized by a graph G. As a simple example of combination, consider two Gaussian graphical models with some common variables which respect some form of compatibility and represent two different studies. Following Massa and Lauritzen (2009), one can find the simplest graphical model that incorporates the information (the conditional independence statements) brought by the two original sets of variables and can be considered the combination of them. Clearly, this is not always possible.

In this paper, we aim at exploring the potential of graphical models combination for model building. We will restrict our attention to the cases in which the combination of initial Gaussian graphical models is possible, i.e., it leads to a joint family of distributions that can itself be defined a Gaussian graphical model. In particular, we will try to build a joint model by combining two submodels representing two different studies involving some common variables. The approach that we will propose is based on performing composition of Gaussian graphical models selections (covariance selections) on the two studies. Adequacy and effectiveness of this approach will be discussed by analyzing simulated and real data.

The outline of the paper is as follows. Section 2 provides some essentials about Gaussian graphical models and briefly reviews some procedures for Gaussian graphical models selection. Section 3 describes the main idea of the paper and presents some numerical studies. Section 4 gives some concluding remarks.

2 Some approaches to model selection for Gaussian graphical models: a review

A Gaussian graphical model (Dempster, 1972; Whittaker, 1990; Lauritzen, 1996; Edwards, 2000) is the family of distributions $\mathcal{F} = \{Y \sim N_p(\mu, \Sigma), \Sigma^{-1} \in S^+(G)\}$, where $S^+(G)$ is the set of positive definite concentration matrices according to the graph G. If we let $\Sigma = \{\sigma_{ij}\}$ and $\Sigma^{-1} = \{\sigma^{ij}\}$, then $\sigma^{ij} = 0$ if and only if there is a missing edge between variables Y_i and Y_j on the graph G corresponding to the pairwise Markov property $Y_i \perp\!\!\!\perp Y_j | Y_{V\setminus\{i,j\}}$. An equivalent condition is that the partial correlation coefficient $\rho_{ij|V\setminus\{i,j\}}$ between Y_i and Y_j is null, since it is well known that

$$\rho_{ij|V\setminus\{i,j\}} = \frac{-\sigma^{ij}}{(\sigma^{ii}\sigma^{jj})^{0.5}}.$$

Note also that $\rho_{ij|V\setminus\{i,j\}}$ is the correlation coefficient of Y_i and Y_j computed from the conditional distribution of Y_i and Y_j given the remaining variables (see Lauritzen, 1996).

Given some data, the procedure of model selection for Gaussian graphical models, also called covariance selection (Dempster, 1972), aims at choosing a Gaussian family of distributions and a dependence graph with the properties above described. More in detail, if $y = (y^1, \cdots, y^n)$ is a sample from an unknown multivariate normal distribution $Y \sim N_p(0, \Sigma)$, with Σ positive definite, we are interested in detecting the undirected graph G such that $\Sigma^{-1} \in S^+(G)$. Clearly, this is equivalent to finding an estimate of the concentration matrix and detect its structural null elements. If

the focus is only on the structure of the graph, model selection is called also structural learning, as it happens in artificial intelligence and machine learning research. Recently, this area of research has attracted more interest, especially to detect the structure of large amount of microarray data (see Section 3 of Castelo and Roverato (2006) for an accurate review), where the number of variables p is very small in comparison with the sample size n. In this paper, we will focus on the classical approach to model selection, i.e., we will work in the $n > p$ setting. This constraint ensures that the sample covariance matrix is a.s. non-singular (Buhl, 1993).

A frequentist view for graphical models selection is founded on a hypothesis testing approach (Edwards, 2000) based on backward stepwise selection. Given some data corresponding to p variables, one usually starts by imposing a complete graph (with all the edges present) and successively checks the inclusion of all the possible edges by testing $p(p-1)/2$ times the null hypothesis $H_0 : \rho_{ij|V\setminus\{i,j\}} = 0$ against the alternative $H_1 : \rho_{ij|V\setminus\{i,j\}} \neq 0$, at a specified level α. Then, the statistically not significant edges are removed and the procedure is repeated until a reduced graph with no edges removed is found. Note that, as pointed out by Edwards (2000) and remarked by Drton and Perlman (2004, 2007), this cannot be seen as a simultaneous testing procedure because the overall error rate cannot be controlled and there is no clear relation with the error level α. Recently, a multi-testing approach for model selection for graphical models was proposed by Drton and Perlman (2004, 2007). Their approach permits to recover the graph and also to control error rates for incorrect edge inclusion.

Another route to model selection is given by maximizing a goodness of fit score by searching through the space of all possible graphs. The goodness of fit scores are given, for example, by the Akaike Information Criterion (Akaike (1974), AIC in the following), the Bayesian Information Criterion (Schwarz (1978), BIC in the following) or the Takeuchi (1976) criterion. However, it is well known that these methods are unfeasible for large values of p, because the number of total graphs deducible from p nodes is in general too big, i.e., $2^{\binom{p}{2}}$. A method that reduces the number of graphs to be searched is the EH procedure of Edwards and Havránek (1985, 1987), which is a global search strategy among all possible models based on the coherence principle of Gabriel (1969). It is based on the assumption that rejection of a model implies rejection of all submodels and acceptance of a model implies acceptance of all models including it.

Recently, some quite new approaches using penalized likelihood methods have been proposed, focussed in particular to estimation of the structural zeros in large sparse concentration matrices. Meinshausen and Bühlmann (2006) introduce the idea of neighborhood selection, which is obtained by fitting a lasso model (Tibshirani, 1996) to each variable. For each variable (node) in the graph, a linear regression (lasso) is performed by using all the other variables as predictors. The non-null coefficients of the regression define the variables on the neighborhood of the initial node. In particular, they set $\hat{\sigma}^{ab} = 0$ if both the coefficients of Y_b and Y_a, obtained by regressing Y_a versus all the other variables and Y_b versus all the other variables, respectively, are zero. Note that, in this way, they only find the structure of the graph. However, once the graph is known, the estimation of the concentration matrix can be achieved in the usual way (see Lauritzen, 1996). Their method is also feasible

for the case $n < p$. Yuan and Lin (2007) propose a sparse and shrinkage estimator of the concentration matrix by imposing a penalization on its off-diagonal elements. Li and Gui (2006) propose a gradient descent algorithm and Banerjee *et al.* (2008) a block coordinate descent algorithm which is the starting point of the graphical lasso by Friedman *et al.* (2008).

3 Combination of covariance selections: numerical studies

Suppose to have some observations from two experiments performed independently from two laboratories and suppose that the intent is to build a reasonable simple model embracing information coming from both the studies. We propose to achieve such construction by composing the results of model selections performed on the two studies. By this, we simply mean that the two graphs obtained through the selection procedures are connected to form a new graph. We tacitly assume that this is admissible by considering only graphs with the same induced subgraphs on the common variables (therefore the conditional independence constraints involving only these variables are the same). In this section, we explore the effectiveness of this approach through some simulation work.

In all the numerical studies, we fix a graph G with p vertices and choose two induced subgraphs G_A and G_B such that if we connect them we obtain G. The two induced subgraphs have q and r vertices, respectively, with $q, r < p$, the intersection of the vertex sets of G_A and G_B is non-empty, and they have the same edges between the common variables. We interpret these three graphs as the true graphs. Then, we generate a sample of size n, $y = (y^1, \ldots, y^n)$, from the Gaussian graphical model corresponding to G, $Y \sim N_p(0, \Sigma)$ with $\Sigma^{-1} \in S^+(G)$. From this sample, we retrieve two subsamples y_A, of size n_A, and y_B, of size n_B, for the variables that are in the vertex set of G_A and G_B. From the theory of multivariate normal distributions, these samples are realizations of the corresponding random vectors $Y_A \sim N_q(0, \Omega)$ and $Y_B \sim N_r(0, \Phi)$, respectively.

The main interest is to identify some reasonable Gaussian graphical model for Y. In this simulation setting, this can be achieved in two ways: 1) by selecting a Gaussian graphical model $\hat{G}$ starting from y; 2) by selecting a Gaussian graphical model $\hat{G}_A$ from y_A and a Gaussian graphical model $\hat{G}_B$ from y_B and then by combining them to obtain $\hat{G}_{\text{comb}}$. The final models obtained in the two ways can then be compared to check whether their structures coincide.

In more detail, in each simulation, we record (i) the number of times that the selection procedure finds a graph $\hat{G}$ with the same structure of G, (*cont1*); (ii) the number of times that the selection procedure finds a graph $\hat{G}_A$ with the same structure of G_A, (*cont2*); (iii) the number of times that the selection procedure finds a graph $\hat{G}_B$ with the same structure of G_B, (*cont3*); (iv) the number of times that the selection procedure finds simultaneously a graph $\hat{G}_A$ and $\hat{G}_B$ with the same structures of G_A and G_B, (*cont4*). Note that if a selection procedure on the subsamples y_A and y_B retrieves both the induced subgraphs G_A and G_B, it also finds the initial one, because their composition is G. When *cont4* (which depends on *cont1* and *cont2*) is greater than *cont1*, the combination of two models may be

considered more convenient than the model selection procedure on all the variables.

To broaden the perspective of the simulation exercise, experiments are divided into balanced and unbalanced. In the first case, it is assumed that the number of observations available for each study is the same. In this case, we set $n_A = n_B = n$. This corresponds to the situation in which model selection on a large set of variables is broken down in selections on smaller subsets of variables followed by a composition of the partial selections. In the unbalanced case, we assume that the number of observations in the two studies is different. As the aim of such unbalanced studies is to test the efficiency of the combination in more general and less favorable contexts, it is assumed that n_A and n_B are smaller than n. For simplicity, we also assume that $n_A + n_B = n$. All the simulations have been performed using R.

3.1 Balanced samples

For the balanced case, we consider two different scenarios, pictured in Figure 1 ($p = 4, q = r = 3$), and Figure 2 ($p = 7, q = r = 5$). To estimate graphs $\hat{G}$, $\hat{G}_A$, $\hat{G}_B$ from samples y, y_A and y_B, we resort on three different selection procedures: (1) unrestricted stepwise backward model selection (Edwards, 2000); (2) model selection based on BIC criterion (Schwarz, 1978); (3) graphical lasso (Friedman *et al.*, 2008). The stepwise and BIC procedures are implemented in the package 'mimR', whereas graphical lasso is implemented in the package 'glasso'. 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. In this way, we tried to let glasso work at its best.

For the first scenario (Figure 1), we set $n = n_A = n_B = 100$ and generate data from the multivariate normal distribution $Y \sim N_4(0, \Sigma)$, with $\Sigma^{-1} \in S^+(G)$. We consider three different choices for Σ, i.e., $\Sigma_1, \Sigma_2, \Sigma_3$ given as

$$\Sigma_1 = \begin{pmatrix} 2.0000 & -1.7807 & -0.9121 & 1.3406 \\ & 2.0000 & 0.6424 & -1.7697 \\ & & 2.0000 & 0.0593 \\ & & & 2.0000 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2.0000 & 1.4445 & 0.0032 & 1.1479 \\ & 2.0000 & 0.9234 & -0.0087 \\ & & 2.0000 & -1.1540 \\ & & & 2.0000 \end{pmatrix}$$

and

$$\Sigma_3 = \begin{pmatrix} 2.0000 & -1.8689 & -1.0958 & -0.3287 \\ & 2.0000 & 0.9283 & -0.0753 \\ & & 2.0000 & 1.1081 \\ & & & 2.0000 \end{pmatrix}.$$

In all the three cases, we set the penalty parameter equal to 0.13, for estimation of G, and to 0.04, for estimation of G_A and G_B. The results are shown in Table 1, which, for each considered model, reports the results for 1000 runs. In general, the results depend on the initial covariance matrix from which the data were generated, which, in turn, depends on the initial partial correlation matrix. For example, we can see that the behaviour with matrix Σ_3 is slightly worse than with the other two. The results of the three selection procedures are comparable, but, as one

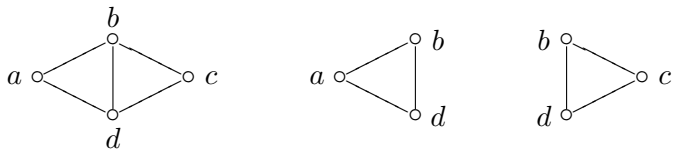

Figure 1: Graph G and induced subgraphs G_A and G_B for the first study with balanced samples.

would have expected, selections using BIC criterion slightly outperform results of the other methods. It is worth noting, however, that feasibility of the BIC procedure decreases rapidly when the dimension of the graph grows.
If Σ_2 is chosen as initial covariance matrix, the graphical lasso performs better than the competing selection procedures, whereas with the other two covariance matrices the graphical lasso is not able to detect the true structure of G. In more detail, if we consider matrix Σ_1, the graphical lasso correctly detects about 42% of the times the absence of only one edge, but this is different from (b,d); in the remaining cases it wrongly detects two or more missing edeges. If we consider matrix Σ_3, the same happens about 60% of the times. Therefore, it seems that, for Σ_1 and Σ_3, the graphical lasso recognizes one missing edge, but, unfortunately, this is not the wanted one. We remark that this behaviour does not depend on the chosen value for the penalty parameter as we found the same evidence in the analyses performed to select the values of λ (results not reported here).
For all three choices of Σ, *cont4* is bigger than *cont1*, showing that the combination is more capable to retrieve the true structure of G.

Matrix	*Method*	*cont1*	*cont2*	*cont3*	*cont4*
	Stepwise	945	1000	1000	1000
Σ_1	BIC	961	1000	1000	1000
	Glasso	0	977	1000	977
	Stepwise	920	1000	961	961
Σ_2	BIC	935	1000	940	940
	Glasso	948	1000	1000	1000
	Stepwise	811	1000	988	988
Σ_3	BIC	831	1000	977	977
	Glasso	0	980	995	975

Table 1: Results for the first study with balanced samples. They refer to the graphs in Figure 1.

In the second scenario (Figure 2), we set $n = n_A = n_B = 200$ and generate data

from the multivariate normal distribution $Y \sim N_7(0, \Sigma)$, with $\Sigma^{-1} \in S^+(G)$ and

$$\Sigma = \begin{pmatrix} 2.0000 & 1.2000 & 0.9622 & 1.2000 & 0.8401 & 0.7727 & 0.7909 \\ & 2.0000 & 1.2000 & 0.99331 & 0.8955 & 0.9931 & 1.2000 \\ & & 2.0000 & 1.2000 & 1.2000 & 1.2000 & 0.9622 \\ & & & 2.0000 & 1.2000 & 0.9342 & 0.7727 \\ & & & & 2.0000 & 1.2000 & 0.8401 \\ & & & & & 2.0000 & 1.2000 \\ & & & & & & 2.0000 \end{pmatrix}.$$

We set the penalty parameter for the graphical lasso equal to 1.0 for G, G_A and G_B.

The results for 1000 runs are shown in Table 2. Even if we use more observations in comparison with the previous example, the results show the difficulty of all the procedures in reconstructing the true graph G. On the contrary, the true graphs G_A and G_B are found about 60% of times. The BIC procedure provides better results but in this case it is extremely slow because it has to test more than two millions of models. The graphical lasso procedure does not provide comparable results. As in previous experiment, the procedure finds the right number of missing edges, but they are not in the right positions (results not shown). Also in this case, values of *cont4* are bigger than *cont1*.

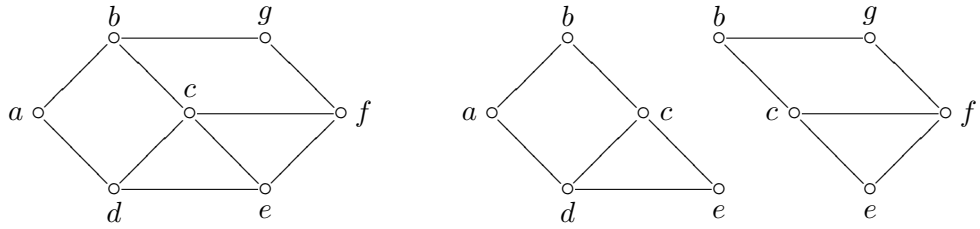

Figure 2: Graph G and induced subgraphs G_A and G_B for the second study with balanced samples.

Overall, taking into account the computational cost of the BIC selection procedure, the stepwise procedure seems to reasonably compete with its competitors. Moreover, combining covariance selections improves ability of the techniques in recognizing the true structures.

Matrix	*Method*	*cont1*	*cont2*	*cont3*	*cont4*
	Stepwise	267	608	612	443
Σ	BIC	373	602	494	361
	Glasso	55	196	173	63

Table 2: Results for the second study with balanced samples. They refer to the graphs in Figure 2.

3.2 Unbalanced samples

For the unbalanced case, we consider again two scenarios, but we substitute to the graph pictured in Figure 2 the graph pictured in Figure 3 ($p = 5, q = 4, r = 3$). For this case, we report only results relative to the stepwise procedure, as we wish to emphasize more on the effect of unbalancing than on the impact of selection procedure. This allows us also to increase the number of runs to 10000.

For the graph in Figure 1, we set $n = 100$, and make n_A and n_B vary from 20 to 80, with $n_A + n_B = 100$ (see Table 3). Then we generate data from the multivariate normal distribution $Y \sim N_4(0, \Sigma)$, with $\Sigma^{-1} \in S^+(G)$, where Σ is chosen as

$$\Sigma = \begin{pmatrix} 2.0000 & -1.7807 & -0.9121 & 1.3406 \\ & 2.0000 & 0.6424 & -1.7697 \\ & & 2.0000 & 0.0593 \\ & & & 2.0000 \end{pmatrix}.$$

The results of Table 3 show that, apart when n_A is small, that creates some difficulties in detecting the true model for G_A, the results of *cont4* always outperforms the results of *cont1*.

n	n_A	n_B	*cont1*	*cont2*	*cont3*	*cont4*
100	20	80	9457	7898	10000	7898
100	30	70	9419	9139	10000	9139
100	40	60	9437	9711	10000	9711
100	50	50	9418	9909	10000	9909
100	60	40	9465	9971	9998	9969
100	70	30	9467	9992	9961	9953
100	80	20	9451	9997	9667	9664

Table 3: Results of the first study with unbalanced samples. They refer to the graphs in Figure 1.

For the graph in Figure 3, we set $n = 100$ and n_A and n_B as before. Then we generate data from the multivariate normal distribution $Y \sim N_6(0, \Sigma)$, with $\Sigma^{-1} \in S^+(G)$. We consider two different choices for Σ, i.e., Σ_1 and Σ_2 given as

$$\Sigma_1 = \begin{pmatrix} 2.0000 & 1.2000 & 0.9698 & 1.2000 & 0.5819 & 0.5819 \\ & 2.0000 & 1.2000 & 0.9698 & 0.7200 & 0.7200 \\ & & 2.0000 & 1.2000 & 1.2000 & 1.2000 \\ & & & 2.0000 & 0.7200 & 0.7200 \\ & & & & 2.000 & 0.7200 \\ & & & & & 2.0000 \end{pmatrix},$$

$$\Sigma_2 = \begin{pmatrix} 2.0000 & 1.9185 & -0.3215 & 0.6645 & -0.1832 & -0.1944 \\ & 2.0000 & -0.3912 & 0.5939 & -0.2230 & -0.2366 \\ & & 2.0000 & 0.9179 & 1.1401 & 1.2095 \\ & & & 2.0000 & 0.5233 & 0.5551 \\ & & & & 2.0000 & 0.6895 \\ & & & & & 2.0000 \end{pmatrix}.$$

Results in Table 4 and Table 5 show that the composition of G_A and G_B is almost always more convenient. In fact, for $n_A \geq 40$, the two subgraphs G_A and G_B are

obtained a greater number of times compared to G (*cont2* $>$ *cont1*, *cont3* $>$ *cont1*) even though less observations are used. Of course, this increases the chances that their combination correctly retrieves G. Note that *cont3* $>$ *cont2*, as the first counter refers to a smaller graph, which is easier to be dealt with. Global performances of the selection procedure depend, naturally, on the initial covariance matrix (compare Table 5 with Table 4).

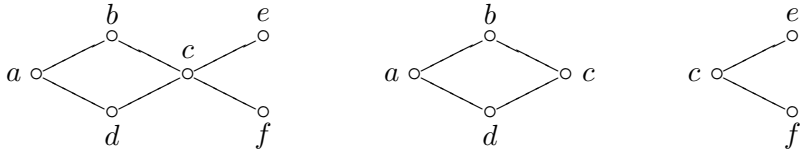

Figure 3: Graph G and induced subgraphs G_A and G_B for the second study with unbalanced samples.

n	n_A	n_B	*cont1*	*cont2*	*cont3*	*cont4*
100	20	80	3990	1200	9510	1120
100	30	70	4174	3519	9423	3296
100	40	60	4400	5790	9290	5350
100	50	50	4324	7001	9310	6508
100	60	40	4326	7848	9059	7114
100	70	30	4254	8254	8414	6944
100	80	20	4274	8475	6675	5671

Table 4: Results of the second study with unbalanced samples. They refer to the graphs in Figure 3 with initial covariance matrix Σ_1.

n	n_A	n_B	*cont1*	*cont2*	*cont3*	*cont4*
100	20	80	2244	1145	9417	1078
100	30	70	2325	1919	9399	1788
100	40	60	2256	2594	9365	2441
100	50	50	2261	3158	9278	2927
100	60	40	2216	3756	8987	3364
100	70	30	2304	3994	8281	3309
100	80	20	2244	4341	6362	2744

Table 5: Results of the second study with unbalanced samples. They refer to the graphs in Figure 3 with initial covariance matrix Σ_2.

3.3 Real data application

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 will 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 concentrate on the six genes and we divide them into two subsets of three and four genes, respectively, with one gene in common, and we go back over the same steps followed in the simulation studies. The graphs are built by using the stepwise procedure and the graphical lasso. The penalization parameter of the graphical lasso for the estimation of G is chosen by looking at the graph in Figure 4, representing the number of edges selected for the penalty parameter varying from 0.01 to 1. The same graphs are obtained for the selection of the penalty parameters for G_A and G_B (figures not displayed). 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 Figure 5), but this is not the case for the stepwise procedure (see Figure 6). 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 it is applied on a set of variables containing both genes. This suggest that the final network selected depends strongly on the chosen subsets of genes.

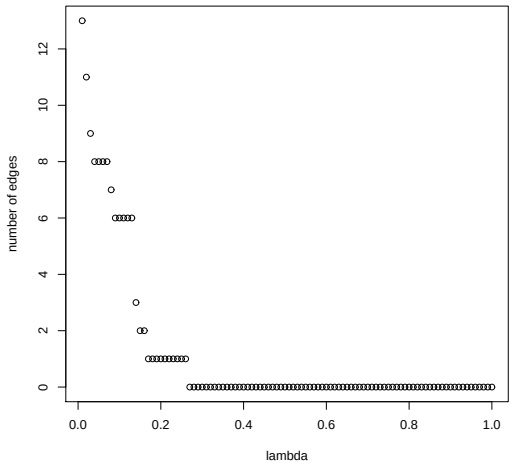

Figure 4: Penalty parameters versus number of edges for the estimation of the graph G with 6 genes.

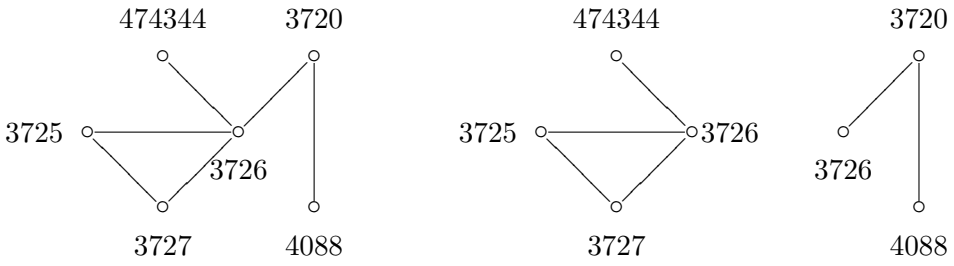

Figure 5: Gaussian graphical models for 6 genes (on the left), for 4 genes (on the center), for 3 genes (on the right) selected using graphical lasso with $\lambda = 0.09, \lambda = 0.1, \lambda = 0.12$, respectively.

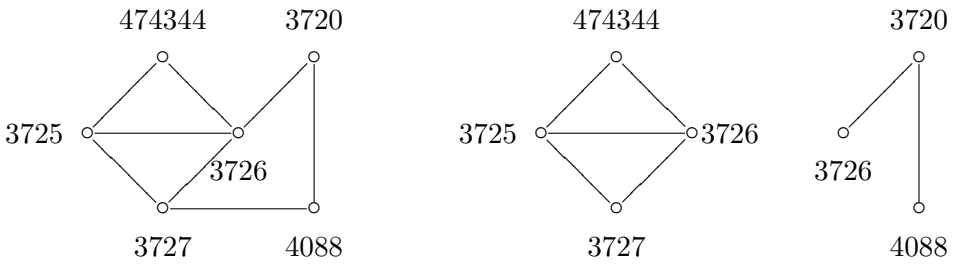

Figure 6: Gaussian graphical models for 6 genes (on the left), for 4 genes (on the center), for 3 genes (on the right) selected using the stepwise procedure.

4 Conclusion

The main interest of this work was to explore the idea of combination of models in the context of model building. In this framework, the intent is to choose a reasonable simple model that is consistent with the available observations taken from heterogeneous studies. In order to achieve this, one may perform separate model selection procedures and then combine their results in some sensible way. Therefore our effort was directed to combine the results of some fixed Gaussian graphical model selections under the assumption of the presence of some overlapping variables.

To this aim, we performed some numerical studies with simulated and real data. The results on simulated data showed that the Gaussian graphical model obtained by combining the results of two Gaussian model selections corresponds more often to the true underlying model from which the observations were generated. In fact, they were compared to the resulting Gaussian graphical model selected on all the variables. Of course, this behaviour was also influenced by the procedure used for model selection and by the initial partial correlation among the variables.
In the brief experiment on real data, we tried to obtain a composition of the results of covariance selections on two sets of observations coming from two sets of variables (genes) with only one variable in common. We checked whether the network searched for all the variables coincided with the network obtained by composition of the two

subnetworks. Again, 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.

This preliminary study of the effectiveness of the proposed model building approach highlights both advantages and disadvantages. On the positive side, such strategy is potentially generic to almost any models in which a common feature can be inferred from a variety of studies, and where the wish is to compose a global picture of the available information based, if possible, on a parsimonious model. On the other side, for the cases in which the substudies to be composed are not naturally defined, we do not yet propose a way of choosing the working subsets. However, particular contexts should suggest appropriate criteria, and 'rules of thumb' may well be developed in the future.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, **9**, 485–516.

Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, **20**, 263–270.

Castelo, R. and Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research*, **7**, 2621–2650.

Dempster, A. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

Drton, M. and Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, **91**, 591–602.

Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Sciences*, **22**, 430–449.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer-Verlag, New York.

Edwards, D. and Havránek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72**, 339–351.

Edwards, D. and Havránek, T. (1987). A fast model selection procedure for large families of models. *Journal of the American Statistical association*, **82**, 205–211.

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

Gabriel, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *Annals of Mathematical Statistics*, **40**, 224–250.

Garret-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., and Gabrielson, E. (2008). Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, **9**, 333–354.

Goh, K.-I., Cusick, M., Valle, D., Childs, B., Vidal, M., and Barabasi, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, **104**, 8685–8690.

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press, Oxford.

Li, H. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, **7**, 302–17.

Massa, M. S. (2008). Combining information from Gaussian graphical models. *PhD Thesis,University of Padova*.

Massa, M. S. and Lauritzen, S. L. (2009). Combining statistical models. *Manuscript*.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, **34**, 1436–1462.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of fitting. *Suri-Kagaku*, **153**, 12–18.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267–288.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. John Wiley & Sons Ltd., Chichester.

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

Acknowledgements

This work was supported by University of Padova grants 070805 and 075919. We wish to thank Lorenzo Maragoni for having carried out some simulations.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy