

R functions for the computation of joint response graphs induced from a generating DAG

Funzioni R per il calcolo di grafi a catena implicati da un DAG generatore

Giovanni M. Marchetti¹

Dipartimento di Statistica “G. Parenti”

Università di Firenze

giovanni.marchetti@unifi.it

Riassunto: Il *package* `ggm`, scritto in linguaggio R consente di adattare modelli grafici Gaussiani basati su grafi non orientati, su grafi orientati aciclici e su grafi ancestrali. In questo lavoro vengono discusse alcune estensioni che consentono di calcolare a partire da un grafo orientato aciclico generatore, dei grafi di altra natura (per esempio grafi a catena) da esso implicati. Gli algoritmi utilizzati sono basati su semplici trasformazioni dalla matrice di adiacenza del grafo generatore (see Cox and Wermuth (2004)).

Keywords: Graphical Markov models, directed acyclic graphs, induced graphs, chain graphs, gR project.

1. Introduction

The R package `ggm` (see Marchetti and Drton (2003)) is designed for fitting Graphical Markov models to data from Gaussian distributions. The underlying graphs can be directed acyclic graphs (DAGs), concentration and covariance graphs and more generally ancestral graphs. See Lauritzen (1996) for the basic terminology on graphs and Richardson and Spirtes (2002) for the definition of ancestral graph. The package is intended as a contribution to the gR-project described in Lauritzen (2002).

Within `ggm` the graph objects are defined simply by their adjacency matrix, or equivalently by their *edge matrix* (cf. Cox and Wermuth (2004)). The edge matrix $\mathcal{A}$ of a graph is derived simply from the adjacency matrix E as $\mathcal{A} = (I + E)^T$.

The package `ggm` has special functions that allow to define easily some basic types of graphs by using model formulae. For example, in figure 1 the directed acyclic graph to the right is defined with the function `DAG` shown top left and its output is the adjacency matrix of the graph.

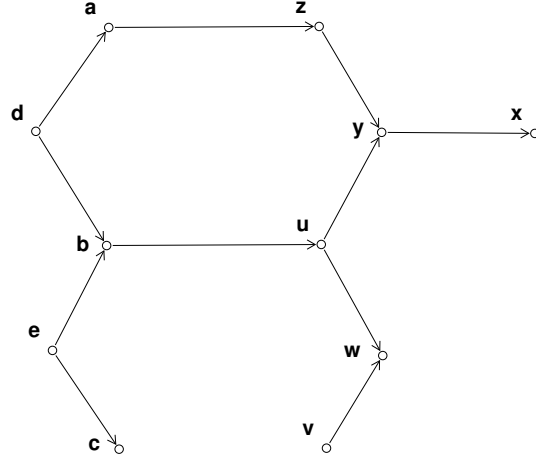
In this paper we describe some useful extensions to the package `ggm` that concern graph computations. In particular we describe procedures to derive from the adjacency matrix of a generating DAG and an ordered partition of its vertices, an implied chain graph after grouping and reordering the variables as specified by the partition. The appropriate theory is discussed in a recent paper by Cox and Wermuth (2004). Here we describe the details of the implementation in R and we show that the adjacency matrix of the chain graph can be computed compactly by a single matrix formula.

¹Indirizzo per corrispondenza: Dipartimento di Statistica “G. Parenti”, viale Morgagni, 59, 50134 Firenze. Il lavoro è stato svolto con finanziamenti MIUR, PRIN 2002.

Figure 1: A generating directed acyclic graph obtained from the function `DAG` using regression model formulae.

```
> dag = DAG(x ~ y, y ~ z + u, w ~ u + v, z ~ a,
+           u ~ b, a ~ d, b ~ d + e, c ~ e)
> dag
```

	x	y	z	u	w	v	a	b	d	e	c
x	0	0	0	0	0	0	0	0	0	0	0
y	1	0	0	0	0	0	0	0	0	0	0
z	0	1	0	0	0	0	0	0	0	0	0
u	0	1	0	0	1	0	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0
v	0	0	0	0	1	0	0	0	0	0	0
a	0	0	1	0	0	0	0	0	0	0	0
b	0	0	0	1	0	0	0	0	0	0	0
d	0	0	0	0	0	0	1	1	0	0	0
e	0	0	0	0	0	0	0	1	0	0	1
c	0	0	0	0	0	0	0	0	0	0	0



2. Chain graphs induced from a generating DAG

Let us denote by V the set of nodes and by $E = [e_{ij}]$ the adjacency matrix of a DAG $D = (V, E)$. Given a set of random variables $Y = (Y_1, \dots, Y_d)$ if they can be associated to the nodes V of the graph such that their density f_Y factorizes as

$$f_Y = f_1 \cdot f_{2|\text{pa}(2)} \cdots f_{d|\text{pa}(d)}$$

where $\text{pa}(i) = \{j \in V : j \rightarrow i\}$ is the set of parents on node $i \in V$, then the random vector Y is said to be Markov with respect to the DAG D . Then, to the set of missing edges of the DAG (or equivalently to the set $M = \{\{i, j\} : e_{ij} = 0, e_{ji} = 0\}$ of structural zeros of the adjacency matrix) corresponds a set of conditional independencies that are satisfied by all distributions that are Markov with respect to the graph (the so called Markov properties, see Lauritzen (1996)).

Given a DAG $D = (V, E)$, a set S of selected nodes, and a disjoint set C of conditioning nodes, we can obtain two induced undirected graphs (see Cox and Wermuth (1996)):

- the induced covariance graph of S given C , which has an undirected edge between i and $j \in S$ iff D does not imply $i \perp\!\!\!\perp j \mid C$;
- the induced concentration graph of S given C , which has an undirected edge between i and $j \in S$ iff the DAG does not imply $i \perp\!\!\!\perp j \mid C \cup S \setminus \{i, j\}$.

The covariance graphs are usually drawn with dashed edges or bidirected edges.

The functions `inducedConGraph` and `inducedCovGraph` are provided in package `ggm` for deriving from a given DAG the induced concentration and covariance graph respectively. The functions use the matrix formulae explained in Cox and Wermuth (2004). The formulae are written more compactly by using the three operators `In`, `anc` and `clos`. Given a square matrix $\mathcal{M}$, $\text{In}\mathcal{M} = \text{abs}(\text{sign}(\mathcal{M}))$ is the indicator matrix of the zeros in the matrix $\mathcal{M}$. Given the edge matrix $\mathcal{A}$ of a DAG, $\text{anc}\mathcal{A} = \text{In}[(2I - \mathcal{A})^{-1}]$ is the edge matrix of the overall ancestor graph. Finally, given the edge matrix $\mathcal{S}$ of an

undirected graph, $\text{clos}S$ is the edge matrix of its transitive closure. We use matrix subscripts to define subsets of matrices. For example, $\mathcal{A}_{aa}$ is the submatrix of $\mathcal{A}$ obtained by selecting the rows and the columns indexed by a . We define analogously the matrix $\mathcal{A}_{ab}$. Note that when a is empty the matrix $\mathcal{A}_{ab}$ has zero rows. In these cases, we will use the convention that the products involving the matrix $\mathcal{A}_{ab}$ return a zero matrix. For example, if $a = \emptyset$ then we have $\mathcal{A}_{ba}\mathcal{A}_{ab} = O_{bb}$.

Then, the induced covariance graph for nodes in S given C has edge matrix

$$(1) \quad \mathcal{S}_{SS|C} = \text{In}[\text{anc}\mathcal{A}_{\bar{C}\bar{C}}\text{clos}(\text{In}(I + \mathcal{T}_{\bar{C}\bar{C}}^T\mathcal{T}_{C\bar{C}}))(\text{anc}\mathcal{A}_{\bar{C}\bar{C}})^T]_{SS}$$

where $\bar{C} = V \setminus C$ and $\mathcal{T}_{ba} = \text{In}[\mathcal{A}_{ba}\text{anc}\mathcal{A}_{aa}]$. Note that this formula is correct even if C is the empty set as the middle factor is the identity matrix and the formula simplifies as $\mathcal{S}_{SS} = \text{In}[\text{anc}\mathcal{A}(\text{anc}\mathcal{A})^T]_{SS}$. An analogous expression is obtained for the edge matrix of the induced concentration graph $\mathcal{S}^{SS|C}$ for S given C .

From a generating DAG and two disjoint sets of nodes S and C we may also derive a chain graph with components C and S . This is defined by two undirected graphs within each component and by a directed graph between them, in such a way that no directed cycles appear. We distinguish two types of directed graphs from C to S :

- the induced *multivariate regression graph* which has an arrow $i \rightarrow j$, with $i \in C$ and $j \in S$, iff D does not imply $i \perp\!\!\!\perp j \mid C$;
- the induced *blocked-concentration graph* which has an arrow $i \rightarrow j$, with $i \in C$ and $j \in S$, iff D does not imply $i \perp\!\!\!\perp j \mid C \cup S \setminus \{i, j\}$;

The edge matrix of the induced multivariate regression graph of S given C is

$$(2) \quad \mathcal{P}_{S|C} = \text{In}[\text{anc}\mathcal{A}_{\bar{C}\bar{C}}\mathcal{T}_{C\bar{C}}^T\text{clos}(\text{In}(I + \mathcal{T}_{C\bar{C}}\mathcal{T}_{C\bar{C}}^T))\mathcal{A}_{C\bar{C}} + \mathcal{F}_{\bar{C}C}]_{SC}$$

where $\mathcal{F}_{ab} = \text{In}[\text{anc}\mathcal{A}_{aa}\mathcal{A}_{ab}]$ and $\mathcal{A}_{bb.a} = \text{In}[\mathcal{A}_{bb} + \mathcal{A}_{ba}\text{anc}\mathcal{A}_{aa}\mathcal{A}_{ab}]$. Finally, the edge matrix of the induced blocked-concentration graph can be obtained from the induced concentration graph for the nodes $C \cup S$, with empty conditioning set, and then extracting the (S, C) matrix block:

$$(3) \quad \mathcal{C}_{S|C} = [\mathcal{S}^{C \cup S \mid C \cup S}]_{SC}.$$

From the above definitions and formulae we may derive induced chain graphs with several components and different interpretations. For example, if the set V is partitioned into three components a , b and c (ordered from a to c), then the induced chain graphs with the Lauritzen-Wermuth-Frydenberg (LWF) and Andersson-Madigan-Perlman (AMP) interpretations (see Andersson and Perlman (2001)) have edge matrices, respectively:

$$(4) \quad \mathcal{S}^{bb|a}, \mathcal{C}_{b|a}, \mathcal{S}^{cc|ab}, \mathcal{C}_{c|ab}$$

$$(5) \quad \mathcal{S}^{bb|a}, \mathcal{P}_{b|a}, \mathcal{S}^{cc|ab}, \mathcal{P}_{c|ab}.$$

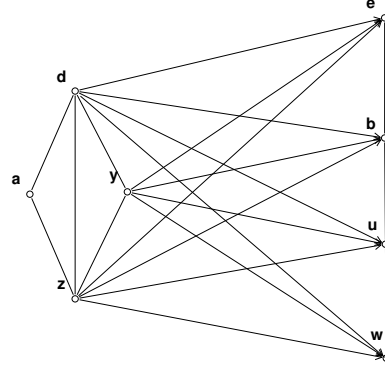
3. R implementation and examples

In R the implementation of the above ideas is straightforward. We add to `ggm` only a function `inducedRegGraph` to compute the adjacency matrix of the multivariate regression graph from equation (2), and a general function `inducedChainGraph` that builds the chain graph, with the required interpretation from (4) and (5).

Figure 2: A chain graph induced by the DAG in fig. 1. The chain components are $A = \{a, d, z, y\}$ and $B = \{b, u, e, w\}$. The interpretation of the graph is the AMP.

```
> cc = list(A = c("a", "d", "z", "y"),
+          B = c("b", "u", "e", "w"))
> cg = inducedChainGraph(dag, cc, type = "AMP")
> cg
```

	a	d	z	y	b	u	e	w
a	0	1	1	0	0	0	0	0
d	1	0	1	1	1	1	1	1
z	1	1	0	1	1	1	1	1
y	0	1	1	0	1	1	1	1
b	0	0	0	0	0	1	1	0
u	0	0	0	0	1	0	0	1
e	0	0	0	0	1	0	0	0
w	0	0	0	0	0	1	0	0



For example, starting from the DAG in figure 1 and choosing as chain components $A = \{a, d, z, y\}$ and $B = \{b, u, e, w\}$ (marginalizing over $M = \{x, c, v\}$) we get the chain graph shown in figure 2. Note that the arguments of `inducedChainGraph` are the adjacency matrix of the DAG, the list of chain components (number of components > 1) and the `type` of the required chain graph (with possible values "AMP" and "LWF"). The function returns the adjacency matrix $E = [e_{ij}]$ of the chain graph. Thus, there is a directed edge $i \rightarrow j$ iff $e_{ij} = 1$ and $e_{ji} = 0$, and an undirected edge $i - j$ iff both $e_{ij} = 1$ and $e_{ji} = 1$.

The zeros in the adjacency matrix of the induced chain graph indicate the independencies that are logical consequences of the generating process underlying the directed acyclic graph. With the induced chain graphs we can easily assess the consequences implied by the DAG when some subsets of variables are considered jointly, with different conditioning set and also with a different ordering.

The function `inducedChainGraph` will be included in the next release of package `ggm`.

References

- Andersson S. Madigan D. and Perlman M. (2001) Alternative Markov properties for chain graphs, *Scandinavian Journal of Statistics*, 28, 33–85.
- Cox D.R. and Wermuth N. (1996) *Multivariate dependencies. Models, analysis and interpretation*, Chapman and Hall, London.
- Cox D.R. and Wermuth N. (2004) Joint response graphs and separation induced by triangular systems, to appear in *Journal of the Royal Statistical Society*, B.
- Lauritzen S.L. (1996) *Graphical Models*, Oxford University Press, Oxford.
- Lauritzen S.L. (2002) gRaphical models in R, *R News*, 3(2)39.
- Marchetti G.M. and Drton M. (2003) `ggm`: an R package for Gaussian graphical models, URL: <http://cran.r-project.org/>.
- Richardson T. and Spirtes P. (2002) Ancestral graph Markov models, *Annals of Statistics*, 30, 962–103.