

Multilevel graphical models: a proposal.

Modelli grafici multilivello: una proposta.

Anna Gottard, Carla Rampichini¹

Dipartimento di Statistica “G. Parenti”

Università di Firenze

gottard@ds.unifi.it, carla@ds.unifi.it

Riassunto: Scopo del presente lavoro è introdurre una classe di modelli multilivello: i modelli grafici multilivello, combinando le potenzialità dei modelli grafici nel trattamento di sistemi multivariati con quelle dei modelli multilivello per l’analisi di dati correlati. I modelli proposti sono un utile strumento per l’analisi di strutture di dati complesse. Tali modelli sono qui utilizzati per lo studio delle determinanti della condizione professionale dei laureati dell’Ateneo fiorentino a circa due anni dal conseguimento del titolo. L’approccio seguito consente sia di tenere conto della natura gerarchica dei dati, laureati per corso di laurea, sia di valutare effetti diretti e indiretti delle covariate sulla variabile di risposta.

Keywords: Chain graphs, Graphical models, Multilevel models, University system evaluation.

1. Introduction

Graphical models are able to describe the topology, in terms of conditional independence, of the components of complex probability models: by the graph, the dependency relationships among the variables included in a probabilistic model can be quickly identified. Moreover, it can be understood how the influence of a variable flows through the system to change the distributions over the other variables.

Multilevel models (Snijders and Bosker, 1999) are useful tools to treat clustered correlated data, particularly if one is interested in relations among variables at different levels in a hierarchical system. Data sets with a nesting structure that include unexplained variability at each level of nesting are often properly represented by the hierarchical models used in multilevel analysis: e.g. pupils in classes, children in families, patients in hospitals, etc. The aim of the paper is to combine the topology properties of the graphical models with the ability to treat clustered correlated data of the multilevel models.

2. Multilevel graphical models

In this section, a class of multilevel models, called multilevel graphical models, is introduced. As it is well known, graphical models are a well-made marriage between probability theory and graph theory. A graph is a pair of sets $\mathcal{G} = (V, E)$, where V is the node set and E is the edge set, a subset of all possible pairs from V . Chain graphs are

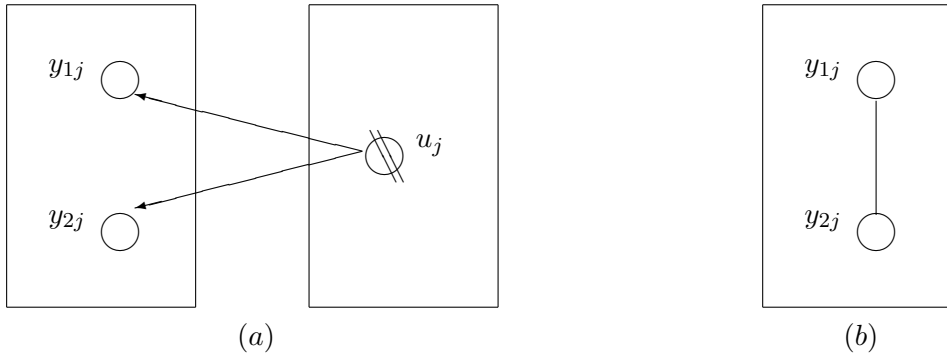
¹Viale Morgagni, 59 - 50134 Firenze.

²Il lavoro è stato svolto nell’ambito del progetto cofinanziato MIUR 2003 “ Apprendimento in modelli grafici e inferenza causale con applicazioni in ambito economico-sociale, genetico e forense”.

acyclic graphs containing both undirected and directed edges. Here, nodes can be partitioned into an ordered sequence of blocks, so that each block is an undirected subgraph and nodes into different blocks can be connected only by arrows. Nodes represent random variables, whereas a missing connection (edges or arrows) between two nodes represents a sort of conditional independence. Undirected edges indicate a symmetric association and arrows indicate an asymmetric one. Theoretical aspects of graphical modelling are comprehensively treated in Whittaker (1990), Lauritzen (1996) and Cox and Wermuth (1996). Typically, graphical models concentrate their attention on the relationships among variables, given that individuals are regarded as independent. This is not true when the responses are correlated. In this case the relationships among individual responses have to be represented in the graph.

In the present work, we propose to represent individual nodes for the response variables in a chain graph. In case of two level hierarchically clustered data, first level observations belonging to the same cluster are correlated, while observations of different clusters are independent. The second level units (clusters) are assumed independent. In multilevel models, this kind of dependence among first level observations is represented by means of a common latent variable, conditionally on which they are independent. In the graph, a latent node of the type $\textcircled{X}$ stands for this latent variable: as made clear in Cox and Wermuth (1996), marginalizing over such a node leads to a correlation between children nodes. To facilitate the reading of the resulting graph, only two individual nodes belonging to a single group are represented. Only one cluster is considered because of the independence of the second level units. This proposed solution on the one hand is easy to be visualized, on the other hand, allows to read from the graph both conditional independencies statements and factorization of the joint probability distribution, according to block-recursive Markov property (see Lauritzen, 1996, p. 54).

Figure 1: A graphical multilevel null model: (a) conditioning on grouping and (b) marginalizing over it.



For example, Figure 1 is a picture of the two-level null model:

$$y_{ij} = \alpha_0 + u_j + \varepsilon_{ij} \quad (1)$$

where y_{ij} is the response variable of the i -th individual of the j -th cluster, $i = 1, 2, \dots, n_j$, $j = 1, 2, \dots, J$, α_0 is the common mean, u_j are i.i.d random variables representing the j -th cluster deviation from the mean, while ε_{ij} are i.i.d. individual residuals. According to the graph of Figure 1, the joint probability distribution can be factorized as:

$$f(\mathbf{Y}, \mathbf{u}) = f(\mathbf{Y} | \mathbf{u})f(\mathbf{u}) = \prod_{j=1}^J f(\mathbf{Y}_j | u_j)f(u_j) = \prod_{j=1}^J \left[\prod_{i=1}^{n_j} f(y_{ij} | u_j) \right] f(u_j) \quad (2)$$

where $\mathbf{Y} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_J\}'$, $\mathbf{Y}_j = \{y_{1j}, \dots, y_{n_{jj}}\}'$, $\mathbf{u} = \{u_1, \dots, u_J\}'$.

3. Application

The proposed models are used to analyze some of the data gathered by a telephone survey conducted, about two years after the degree, on the 2000's graduates of the University of Florence. The main goal is to determine factors influencing graduates job position. Moreover, we want to evaluate the degree programmes on the basis of the probability of finding a stable occupation of graduates. The pure response variable Y , *job position at the date of the interview*, is a polytomous variable (Table 1²), taking on value '0' if the graduate doesn't have a job, '1' if he/she has a temporary job and '2' if he/she has a stable job.

Table 1: *Graduates by job position (2000).*

<i>Job position</i>	<i>Frequency</i>	<i>Percent</i>
0 no job	249	8.02
1 temporary job	1458	46.97
2 stable job	1397	45.01

Table 2: *Variables definition and blocks ordering*

<i>X</i>	<i>Name</i>	<i>Description</i>	<i>Block</i>
11	father educational level	1-5	1 (pure explanatory)
12	mother educational level	1-5	
13	male	1 if male	
21	high school type	4 types	2 (intermediates)
22	high school rank	36-60	
23	regular career	1 if regular	
31	enrollment year	1(<1990)-5(>1995)	3 (intermediates)
32	short degree	1 if short degree	
33	school	11 schools	
41	examinations rank average	18-31	4 (intermediates)
42	duration index	0.91-4.86	
51	training	1 if training done	5 (intermediates)
52	honors	1 if graduate with honors	
53	age at the degree	22-51	
61	degree-interview distance	12-33 months	6 (intermediate)
u_j	course programme	56 course programmes	(latent node)
y_{ij}	job position	0 no, 1 temporary, 2 stable	7 (pure response)

Many observed and unobserved variables may influence the probability of finding a temporary or stable job. Variables can be ordered into a sequence of blocks, according to subject-matter knowledge or time. The unexplained variability at the course programme level is represented by a latent node. The chosen ordering is reported in Table 2.

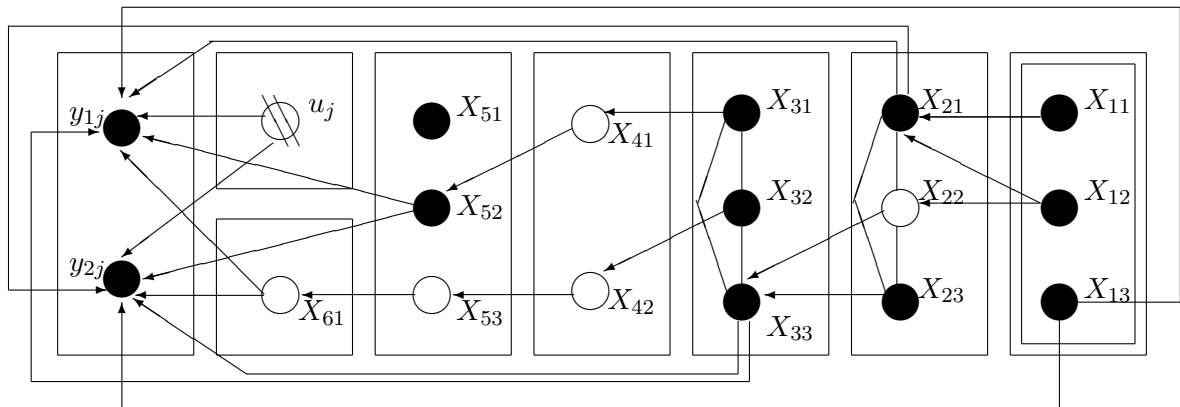
The estimation of a chain graph model is a difficult task, requiring many steps. The estimation strategy adopted in this work is based on fitting univariate appropriate regression models according to the dependent variable scale and to the recursive nature of the chain

²Only graduates having looked for a job have been considered.

graph (Cox and Wermuth, 1996). To take account of the hierarchical structure of the data and of the multinomial nature of the pure response variable, a suitable graphical multi-level model for polytomous response (Skrondal and Rabe-Hesketh, 2003) is developed and fitted by means of maximum likelihood with adaptive Gaussian quadrature.

The dependence structure of the data is quite complex. Only 5 of the 15 explanatory variables have a direct effect on the job position, but only one ($X_{51} = \text{professional training}$) among them has no effect at all on the pure response variable. For example, the parents educational level has only an indirect effect on the pure response through the high school type and the high school rank. The introduction of a latent node, representing the course programme effect, is substantial: the likelihood ratio test comparing the models with and without the latent grouping effect is significant. The proportion of residual variance explained by the grouping is about 8.25%. To give an idea of usefulness of the proposed approach, Figure 2 represents a part of the resulting chain graph, where all undirected edges and only some arrows are drawn to highlight all direct and some indirect effects on the pure response. It is important to stress that graphical models are not a causal models: particularly, in this application, the analysis is conditional on (a) the choice to enroll at the university after the high school and (b) to not drop out during the university career.

Figure 2: A part of the resulting chain graph



References

- Cox D. and Wermuth N. (1996) *Multivariate Dependencies. Models, analysis and interpretation*, London: Chapman and Hall.
- Lauritzen S. (1996) *Graphical Models*, Oxford: Oxford Science Publications.
- Skrondal A. and Rabe-Hesketh S. (2003) Multilevel logistic regression for polytomous data and rankings, *Psychometrika*, 68, 267–287.
- Snijders T. and Bosker R. (1999) *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*, Sage, London.
- Whittaker J. (1990) *Graphical Models in Applied Multivariate Statistics*, Wiley: New York.