UNIVERSITÀ DEGLI STUDI DI FIRENZE

DIPARTIMENTO DI STATISTICA "G. PARENTI"

DOTTORATO IN STATISTICA APPLICATA XVI CICLO

Issues on the Estimation of Latent Variable and Latent Class Models with Social Science Applications

Fulvia Pennoni

Supervisor
Prof. Giovanni Marchetti

Our progress
Does Not Consist
in Presuming
we have Arrived
but *in striving*
constantly
Towards the Goal.

St. Bernard

Contents

Acknowledgments

First of all I would like to thank my supervisor, Professor Giovanni M. Marchetti, for his readiness to help me and in sharing ideas with me and solving my problems.

I would like to thank Prof. Elena Stanghellini, who enabled me to get into contact with other researchers working on related topics, essential for developing the basic ideas of this thesis and for her comments on the first drafts.

I am grateful to Professor Brian Francis for giving me the opportunity to work with him and also to Professor Keith Soothill at Lancaster University.

I would like to express my gratitude to all the staff of the Centre of Applied Statistics who provided me with a nice environment when I was working there.

I am indebted to many other persons who inspired me and facilitated my work in many ways. Professor Harry Kiiveri for providing me with his PhD thesis and for many pieces of his advice concerning the differentiation of difficult functions. Professor Nanny Wermuth for her suggestions and examples. She allowed me to attend the workshop on graphical Markov models in Wiesbaden; that event was proved to be a sound source of inspiration.

I have benefitted also from conversation with, and comments by, many other people including Professors Joe Whittaker, Giorgio Calzolari and Marco Barnabani.

I would like to thank also Mr. Alan Airth for checking and correct many English mistakes.

Furthermore I am grateful to Mariella and Paolo for motivating me to undertake this programme of study and to Alessandro and to all my other friends for sustaining me in dark moments and their continuous confidence on my abilities.

Introduction

The following work is made up of separated and distinct research problems but they have in common the presence of latent or hidden variables. These rather different problems illustrate *"how latent variables pervade modern mainstream statistics and are widely used in different disciplines such as medicine, economics, engineering, psychology, geography, marketing and biology"* (Skrondal and Rabe, 2004), they arise in a wide variety of applications.

Although latent variables are part of numerous statistical and data analysis models, we do not have a single general definition of latent variables that would include all of these diverse applications. The techniques involving latent variables have risen and grown almost exclusively within the framework of the social and behavioral sciences but they have different names in the extensive literature which is extended over almost a century and presents a disjointed picture both notationally and conceptually of the treatment of latent variable models. Rather we have definitions of latent variables that are closely tied to specific statistical models and few systematic comparisons of these different definitions and the implications of the differences. Unmeasured variables, factors, unobserved or hidden variables, constructs, true scores, unobserved confounders or missing variables are some of the terms that researchers use to refer to variables in the models that are not present in the data set. What are latent variables? Psychometricians and others have espoused various opinions on the meaning of latent variables and their relevance to empirical inquiry; for a review see Von Eye and Clogg, (1994), Bollen (2002), Borsboom *et al.* (2003).

In the literature at a minimum we can distinguish two broad types of latent variables. The first are notional "true" variables measured with errors, the second are hypothetical concepts which are often related to several observed variables and which are assumed to represent the underlying features of the system under study (hypothetical construct). The first type are variables for which exact measurements are not available. Variables of this type are usually well defined but the available measures are assumed to be imperfect indicators. The model used to take into account the measurement errors is known as the *measurement model*. If the latent variable and the measured variables are both

continuous it is called the *factor model* and it is analyzed in full detail in Section 1.5.1. If the true variable is constructed as categorical and also the observed variables are categorical the models are known as *latent class models*, they are analyzed in Section 2.1.1. Hypothetical constructs such as "self-esteem" and "life-satisfaction" are prominent in psychological research. Sociologists are often concerned with constructs such as "aspiration" and "alienation" whereas political scientists are interested in, for instance, "political efficacy", "expectations", "permanent income". Although such hypothetical concepts and constructs, or latent variables, cannot be directly measured, a number of variables can be used to measure various aspects of these latent variables more or less accurately.

Though this classification into types is useful and often used, it has been criticised. Rubin (1982), for example, has argued that it is at least confusing to say that the observable is some how less "true" than the unobservable. However conceptualization in terms of latent variables is a useful way of looking at the relationships between the observed variables. In such cases the role of the latent variables consists in clarifying the interpretation of relatively complex association structures and the effect of errors of observations in possibly distorting the dependencies.

All latent variable models involve an observable (or manifest) random vector $X = (X_1, ..., X_J)$ and an unobservable (or latent) variable U, which may be either unidimensional or vector valued. In any latent variable model, the manifest variable and the latent variable are assumed to have a joint distribution over a sample space. A basic condition on (X, U) is that of local conditional independence or (local independence) (Holland and Rosenbaum, 1986) that is

$$F(x_1, ..., x_J|u) = \prod_{j=1}^{J} F_j(x_j|u);$$

when the latent variable is held constant the manifest variables should be statistically independent. A variable U for which the latent conditional independence holds is often said to explain completely the association structure between the manifest variables $X_1, ..., X_J$.

Suppes and Zanotti (1981) have shown that if X has a finite number of possible values then there always exists a one-dimensional latent random variable such that (X, U) satisfies this condition. Latent conditional independence taken alone does not allow us to draw scientific conclusions, but *"other conditions such as linearity, monotonicity and functional form are features of latent variables models that give them testable consequences in observed data*" (Holland and Rosenbaum, 1986).

This work is structured as follows. The main core of the work is based on Chapter 3

and Chapter 4. They are written in the article style and they have been submitted for publications to two international reviews as they contain some new methodological and or applicative notes. Chapter 1 and Chapter 2 provide basic background knowledge for the models developed in the last chapters. Every chapter has its own references and a comprehensive bibliography of the thesis is given as last section.

In chapter one we concentrate on the structure of models of dependence and association and on their interpretation using graphical models which have been proved useful to display in graphical form the essential relationships between variables. As a framework for analyzing multivariate data sets, graphical Markov models give a direct and intuitive understanding of the possibly complex underlying dependence structure; second they give a precise representation of qualitative information about conditional independencies in the underlying statistical model, and third the structure of the graph yields direct information about various aspects related to the statistical analysis.

In most cases the statistical meaning of association is some kind of conditional dependence. Thus, a missing edge indicates conditional independence of the corresponding variables, given all the remaining variables. We focus on dependence structures derived from multivariate normal random variables by marginalizing with respect to some components. The results are useful because some components are difficult to measure, their measurement begin either inaccurate or simply incomplete. It is then helpful to know whether an omission induces spurious or misleading association among the remaining fully observed components.

At first we provide the necessary notation and background on graph theory. We describe the Markov properties that associate a set of conditional independence to an undirected and directed graph. Such definitions does not depend of any particular distributional form and hence can be applied to models with both discrete and continuous random variables. In particular we consider models for Gaussian continuous variables where the structure is assumed to be adequately described via a vector of means and by a covariance matrix; the concentration and the covariance graphs models are illustrated where edges represent conditional independence on the first case and marginal independence on the second.

For models with continuous variables the factorization of the joint density provides a simple method for constructing a multivariate distribution with specific conditional independencies: specify univariate regression models such that the explanatory variables in each regression are those which are thought to have a direct influence. Due to the acyclicity this has to be done in a recursive way. The specification of the complex multivariate distribution through univariate regressions induced by a Directed Acyclic Graph (DAG) can be regarded as a simplification, as the single regression models typ-

ically involve considerably fewer variables than the whole multivariate vector. The distributions generated over such graphs are called *triangular systems*. In the present work it is shown that such models are a subclass of the structural equation models developed for linear analysis known as *Structural Equation Models* (SEM).

If the DAG represent systems with latent variablesthe edges corresponding to such nodes are denoted in the graph by a double crossing over the nodes $\{\not\emptyset\}$ to illustrate that they are not observed. The independencies entailed by the DAG which is assumed to describe the data generating process, can be encoded in two types of induced graph: the overall concentration and the overall covariance graph, which are useful to show the overall conditional and marginal independence structure implied by the DAG. Another special class of graphs, called *summary graph* is illustrated, which is an attempt to define a graphical representation of the independence structure of the DAG which results after marginalizing over and conditioning on some variables. It is shown how such graph may represent models known in literature as the instrumental variable model and the seemingly unrelated regression model.

Model identifiability is discussed for models in which the existence of just one variable is hypothesised. Some known results and new graphical criteria based on properties of the induced conditional independence graphs are reported. An overview of the application of the likelihood based inference is also given and some iterative methods are enlightened. The chapter is concluded by some bibliographical notes.

Chapter 2 takes into account model for discrete variables and in particular an introduction to the standard latent class model is provided as a model for measuring one or more latent categorical variables by means of a set of observed categorical variables. We describe some issues on the model identifiability and estimation and then we asses the problem on how the latent class model can be used to model the nature of the latent changes over time when longitudinal studies are used. Recently, latent variables have been used under the name "hidden" variables in Markov modelling. The hidden Markov model is presented which consists of a hidden state variable and a measured state which both varying over time. The conditional independence relations of the model are enlightened which are depicted as in a graphical Markov model. Bibliographical notes conclude the chapter.

In Chapter 3 we consider DAG models in which one of the variables is not observed. Differently from the factor model the latent variable may be not only exogenous but it can be an intermediate variable of the data generating process. Once the condition for global identification has been satisfied, we show how the incomplete log-likelihood of the observed data can be maximize using the EM algorithm. As the EM does not provide the matrix of the second derivatives we show a method for obtaining an explicit

formula of the observed information matrix using the missing information principle. We illustrate the models with several examples taken from the literature. In the first appendix calculations of the derivatives are reported and in the second appendix the R code is reported, to get the estimated standard errors, which is implemented for the R package ggm (Marchetti and Drton, 2003).

In Chapter 4 taking into account the problem of classifying criminal activity the latent class cluster model is extended by proposing a model that also incorporates the longitudinal structure of data using a method similar to a local likelihood approach. The proposed methodology can also be used to classify other types of longitudinal event history where the interest is in the changing nature of activity over time. The chapter starts with the description of the data set which is taken from the Home Office Offenders Index of England and Wales. It contains the complete criminal histories of a sample of those born in 1953 and followed for forty years. The models and the result of the separated analysis for males and females are presented in detail.

Chapter 1

Continuous Graphical Models with latent variables

1.1 Graph Notation and Terminology

In the following the notation and terminology for some types of graphical Markov models will be introduced. It is mainly borrowed from Lauritzen (1996), Anderson *et al.* (1995) and Castelo (2002).

1.1.1 Decomposable Graphical Markov Model

A graph is a pair $G = (V, E)$ where V is the set of vertices and E is a set of edges. In the present context, the set of vertices acts as an index set for some collection of random variables $X_{\mathrm{v}} = (X_1, ..., X_n)$ that form a multivariate distribution of some family P.

The set of edges E is a subset of the set of ordered pairs $V \times V$ that does not contain loops, i.e $(x, y) \in E \Rightarrow x \neq y$, nor multiple edges. Given two vertices a, b we say that they form an *undirected edge* if and only if $(a, b) \in E$ and $(b, a) \in E$. An undirected edge is represented graphically by a solid line joining the two vertices involved, e.g. $a - b$. When all the edges in E are undirected we say that the graph G is an *undirected graph* (UG).

When two vertices are joined by an undirected edge, these two vertices are called *adjacent*. Given a vertex $v \in V$ the *boundary* of v is $bd(v) = \{u \in V | (u, v) \in E\}$. The *closure* of a vertex v is $cl(v) = bd(v) \cup \{v\}$.

A *subgraph* $G_S = (S, E_S)$ is given by a subset $S \subseteq V$ and the induced edge set $E_S = E \cap (S \times S)$. It will be said that G_S is an *induced subgraph* of G. An undirected graph $G = (V, E)$ is said to be complete if and only if every pair of vertices is adjacent. A *clique* is a maximal complete subgraph.

An *undirected path* between two vertices a and b is a sequence $a = v_o, ..., v_n = b$ of distinct vertices such that $n > 0, (v_{i-1}, v_i) \in E$ and $(v_i, v_{i-1}) \in E$ for $i = 1, ..., n$. An *undirected cycle* is an undirected path that begins and ends at the same vertex, i.e $a = b$. An *odd cycle* is a cycle on an odd number of nodes. A subgraph is *connected* if every pair of nodes is connected by a path. By a *connectivity component* we mean a maximal connected subgraph.

Given three subsets of vertices $A, B, S \subset V$, it is said that S *separates* A from B in an undirected graph if and only if every undirected path between vertices in A and B intersects S.

A *chord* in a graph is an edge joining two vertices already connected by a graph. An *undirected chordal graph*, or *chordal graph*, is an undirected graph with no chordless undirected cycles on more than three vertices, chordal graphs are also known as *decomposable, triangulated* or *rigid circuit* graphs, and their properties have been exploited in many areas of research.

We define a *complementary graph* of an undirected graph G as the graph $\bar{G}$ with the same set of nodes, and an undirected edge connecting u and v whenever there is not an edge between u and v in G.

Markov properties and definition

There are three Markov properties for undirected graphs.

Definition 1.1. *Undirected pairwise Markov property (UPMP)*
Let $G(V, E)$ be an undirected graph. A probability distribution P is said to satisfy the undirected pairwise Markov property if, for any pair $u, v \in V$ of non-adjacent vertices, P satisfies

$$u \perp\!\!\!\perp v | V \backslash \{u, v\}.$$

This means that two non-adjacent vertices $u, v \in V$ are conditionally independent given the rest of the vertices.

Definition 1.2. *Undirected local Markov property (ULMP)*
Let $G(V, E)$ be an undirected graph, a probability distribution P is said to satisfy the undirected local Markov property if, for any vertex $v \in V$, P satisfies

$$v \perp\!\!\!\perp V \backslash cl(v) | bd(v).$$

This means that a vertex v is conditionally independent of the rest of the variables without its boundary, given its boundary.

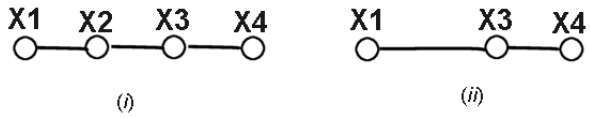

Figure 1.1: *(i) an undirected graph G_1; (ii) an undirected graph G_2 representing the conditional independence structure induced on $\{X_1, X_3, X_4\}$* by G after marginalizing on X_2.

Definition 1.3. *Undirected global Markov property (UGMP)*
Let $G(V, E)$ be an undirected graph, then a probability distribution P is said to satisfy the undirected global Markov property if, for any triple (A, B, S) of disjoint subsets of V, such that S separates A from B in G and A, B are non-empty, P satisfies

$$A \perp\!\!\!\perp B|S.$$

This means that two non-empty subsets of vertices A, B are conditionally independent given a third subset S, if and only if S separates A from B.

In Whittaker (1990) and Lauritzen *et al.* (1990) there is more thorough discussion of these Markov properties. In particular they show the following relations between the properties:

$$UGMP \Rightarrow ULMP \Rightarrow UPMP.$$

This implies that the UGMP is the most comprehensive possible rule for reading off conditional independence restrictions from an undirected graph. If P is strictly positive, then the properties are equivalent (see, e.g. Pearl and Paz, 1987).

We can now introduce the definition of a decomposable graphical Markov model.

Definition 1.4. *Let $G(V, E)$ be an undirected decomposable graph. The set $\mathbf{U}(G)$ of all probability distributions that satisfy the UGMP relative to G is called the decomposable graphical Markov model, determined by G.*

The distinction between decomposable models with chordal graph and non-decomposable models is important. Decomposable models have direct maximum likelihood estimates and allow the factorization of the likelihood function with consequent local computation of important statistical quantities. (Frydenberg and Lauritzen, 1989).

As remarked by Richardson and Spirtes, (2002), Markov models based on UGs are *closed under marginalization* in the following sense: if an undirected graph represents the conditional independencies holding in a distribution then there is an undirected graph that represents the conditional independencies holding in any marginal of the

distribution. For example consider the graph G_1 in Figure 1(i). If we suppose that X_2 is not observed, then it is evident that the conditional independence, $X_1 \perp\!\!\!\perp X_4 | X_3$ which is implied by G_1 is represented by an undirected graph G_2 in Figure 1(ii), which does not include X_2. In addition, G_2 does not imply any additional independence relations that are not implied by G_1.

1.1.2 The Acyclic Directed Graphical Markov Models

An edge is *directed* if and only if $(a, b) \in E \Rightarrow (b, a) \notin E$. A directed edge between two vertices a and b, such that $(a, b) \in E$, will be represented graphically by an arrow pointing from a towards b, i.e $a \rightarrow b$. A graph $G = (V, E)$ is said to be *directed* if all edges in E are *directed edges*.

For a directed edge $a \rightarrow b$ we distinguish between the two joined vertices by specifying that a is the *parent* of b, and that b is the *child* of a. Those parent vertices that have a common child, will be considered as the *parent set* of this child vertex, and it will be noted as $pa(v)$ for any given child vertex v.

A *path* between two vertices a and b is a sequence $a = v_0, ..., v_n$ of distinct vertices such that $n > 0$ and either $(v_{i-1}, v_i) \in E$ or $(v_i, v_{i-1}) \in E$ for $i = 1, ..., n$. A *cycle* is a path where $a = b$. In a directed graph, a *directed path* is formed by directed edges and is a *direction preserving* path. This means that every directed edge in the path points towards the same direction. A given vertex a is called the *ancestor* of b if there is a directed path from a to b. A *directed cycle* is a directed path where the first vertex coincides with the last one. A *acyclic directed graph*, or DAG, is a directed graph without directed cycles. The *skeleton* of a DAG is the undirected graph obtained by transforming the set of directed edges into a set of undirected ones that preserves the same adjacencies.

For a given vertex v, one may consider the set of those vertices that are ancestor of v, which will be called the *ancestor* set of v, and noted $an(v)$. A vertex b is called the *descendant* of a if there is a directed path from a to b, i.e. a is an ancestor of b. The vertices at the end of every directed path that starts at a vertex a will form the descendant set of a, noted $de(a)$. Given a vertex v the *non-descendant* set of v is defined as $nd(v) = V \backslash \{de(v) \cup \{v\}\}$.

An important concept regarding DAGs in this context is the concept of *immorality*. An *immorality* is formed by two non-adjacent vertices with a common child, e.g $a \rightarrow b \leftarrow c$. In the terminology of Cox and Wermuth (1996), an immorality is known as a *sink-oriented V-configuration*, where a *V-configuration* is defined as a triplet of vertices (a, b, c) such that two of them are adjacent to the third one but they are not adjacent

to each other. In the case $a \rightarrow b \leftarrow c$, the vertex b is referred as the *collision* vertex. The terminology of Cox and Wermuth (1996) allows to define further configurations on the three vertices as the *source-oriented V-configuration*, e.g $a \leftarrow b \rightarrow c$, and the *transition-oriented V-configuration*, e.g. $a \rightarrow b \rightarrow c$.

A DAG that has no immoralities is said to be *moral*. A DAG that is not moral can be *moralized* by marring those non-adjacent parents that induce an immorality, i.e. joining them with an undirected edge, and dropping directions on the rest of edges in G. The moralized version of a directed graph G will be noted as G^m.

The statistical meaning of variables corresponding to the different types of nodes as showed in Cox and Wermuth (1996) is the following. The variable at a parent node is *directly explanatory* for the variable at a child node. The variable at an ancestor but not at parent node is *indirectly explanatory* for the variables at descendant node. The variable at a common ancestor node is a common explanatory variable. Further a variable at a transition node is an intermediate variable, at a common sink node is a common response and at a common source node it is a common directly explanatory variable.

Markov properties of DAGs

We are going to illustrate some basic properties connecting graphical structure and probability distributions in a DAG. Three are the Markov properties for DAGs.

Definition 1.5. *Directed pairwise Markov property*
Let $G = (V, E)$ be a DAG, a probability distribution P is said to satisfy the directed pairwise Markov property if, for any pair $u, v \in V$ of non-adjacent vertices such that $v \in nd(u)$. P satisfies

$$u \perp\!\!\!\perp v | nd(u) \backslash \{v\}.$$

The directed pairwise Markov property means that two non-adjacent vertices u and v, such that v is non-descendant of u, are conditionally independent given the non-descendant vertices of u without v.

Definition 1.6. *Directed local Markov property*
Let $G = (V, E)$ be a DAG, a probability distribution P is said to satisfy the directed local Markov property if, for any vertex $v \in V$, P satisfies

$$v \perp\!\!\!\perp \{nd(v) \backslash pa(v)\} | pa(v).$$

The directed local Markov property means that a vertex is conditionally independent of its non-descendants, without its parents, given its parents.

Definition 1.7. *Directed global Markov property*
Let $G = (V, E)$ be a DAG, a probability distribution P is said to satisfy the directed global Markov property if, for any triple (A, B, S) of disjoint subsets of V, where A, B are non-empty, such that S separates A from B in the moralized version of the subgraph induced by the vertices in $An(A \cup B \cup S)$, i.e. in $G^m_{An(A\cup B\cup S)}$, P satisfies

$$A \perp\!\!\!\perp B|S.$$

(Proof: Lauritzen 1996, p.51).
The directed global Markov property means that two non-empty subsets of vertices A, B are conditionally independent given a third subset S if S separates A and B in the moralized subgraph induced by the smallest ancestral set of $A \cup B \cup S$.

We can now state the following definition for directed acyclic Markov model also called *Bayesian Network*.

Definition 1.8. *DAG Markov model*
Let G be a DAG. The set $\boldsymbol{D}(G)$ of all probability distributions that satisfy the directed global Markov property relative to G is called the acyclic directed graphical Markov model, or DAG Markov model, determined by G.

An important property of a distribution P satisfying the local directed Markov property associated with a DAG is that its joint density can be decomposed into conditional probabilities involving only variables and their parents according to the structure of the graph in the following way

$$f(x) = \prod_{v\in V} f(x_v|x_{\mathrm{pa}(v)}).$$

An alternative way of reading conditional independence in a DAG is by using a separation criterion that permits direct reading from the graph whether the defining independence structure of the graph implies a given conditional independent statement. There are two equivalent separation criteria for DAG. One is the *d-separation* criterion of Pearl and Verma (1987). Given two vertices $u, v \in V$ and a subset $S \subseteq V$ where $u, v \notin S$ one says that a path between u and v is *active* with respect to S if

1. every non-collision vertex in the path is not in S or
2. every collision vertex in the path is in S or has a descendant in S.

When a subset S creates an active path between two vertices u and v, then u and v cannot be conditionally independent given S in G. When a path between two vertices u, v is not active with respect to S, one says that the path is *blocked* by S.

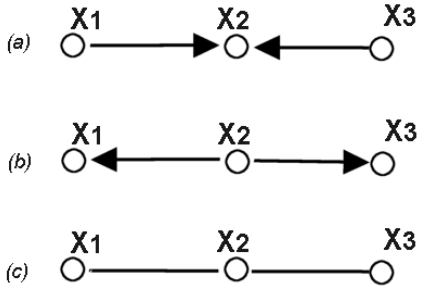

Figure 1.2: *(a) DAG with three observed variables one common child node; (b) DAG with three observed variables and one common parent; (c) moral graph of the former DAG (b).*

Definition 1.9. *d-separation*
Let $G = (V, E)$ be a DAG. For any triple (A, B, S) of disjoint subsets of V, where A, B are non-empty, A and B are d-separated by S if every path between the vertices in A and B is blocked by S.

This *d-separation* rule gives sufficient conditions for two vertices in a DAG to be observationally independent upon conditioning on some other set of vertices.
Lauritzen *et al.* (1990) states another separation criterion based on a moral graph.

Definition 1.10. *Separation*
Let $G = (V, E)$ be a DAG. For any triple (A, B, S) of disjoint subsets of V, where A, B are non-empty, A and B are separated by S if in the moral graph formed from the smallest ancestral set containing $A \in B \in S$ every path from A to B has a node in C.

The graph in Figure 2(a) has vertices $V = \{1, 2, 3\}$ and edges $E = \{(1, 2), (3, 2)\}$. The only independence in this model is $X_1 \perp\!\!\!\perp X_3$ i.e these variables are marginally independent. In addition both influence a third variable X_2, thus conditioning on X_2 corresponds to selecting a subsample of the population and for this subsample X_1 and X_3 are not necessarily independent any more i.e $X_1 \not\perp\!\!\!\perp X_3 | X_2$. Representing the conditional independencies induced by this DAG in an undirected graph where no information on marginal distribution of subvectors can be retained, therefore requires that parents of a common child are linked. Thus the moral graph for Figure 2(a) has edges $E^m = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$. This also implies that no undirected graph can be found to represent the conditional independencies induced by Figure 2(a). In contrast the graph of Figure 2(b) encodes the same conditional independencies of the undirected graph of Figure 2(c) which is its moral graph.

1.2 Background to the graphical Gaussian Markov models

A multivariate Gaussian graphical Markov model is a family of multivariate normal distributions which satisfy a collection of conditional independencies related to an undirected graph. Before showing the graphical representation of such special situation we give an account of the covariance matrices. Linear structures, in fact, which are analysed in the following sections impose some form on the covariance and concentration matrix.

In the following we consider the $p \times 1$ random vector $X_v = (X_1, ..., X_p)$ having a Gaussian distribution with mean $E(X) = \mu$ and covariance matrix $\text{cov}(X) = \Sigma$ assumed positive definite, and its inverse, the concentration matrix, denoted by Σ^{-1}. In what follows we denote vertices by the letters $i, j, .., p$. The diagonal elements of Σ are the *variances* σ_{ii} and those of Σ^{-1} are the *precisions* σ^{ii}. The off-diagonal elements of Σ are the *covariances* σ_{ij} and those of Σ^{-1} are the *concentrations* σ^{ij}. Covariances and concentrations are measures of association; variances and precisions are measures of variability.

If we partition $X = \{X_a, X_b, X_c\}$, the set $V = \{a, b, c\}$ is partitioned accordingly, where a, b, and c are disjoint subsets of V (a and b nonempty) such that $a \cup b \cup c = V$. We indicate with Σ_{ab} the submatrix $[\Sigma]_{a,b}$ of a matrix Σ and with Σ^{ab} the submatrix $[\Sigma^{-1}]_{a,b}$ of its inverse. The covariance Σ and the concentration (or precision) matrix Σ^{-1} of the X_v are then written as:

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} \Sigma^{aa} & \Sigma^{ab} & \Sigma^{ac} \\ \Sigma^{ba} & \Sigma^{bb} & \Sigma^{bc} \\ \Sigma^{ca} & \Sigma^{cb} & \Sigma^{cc} \end{pmatrix}$$

From covariance algebra a marginal correlation ρ_{ij} between X_i and X_j, is expressible via elements of the covariance matrix in a way similar to that in which a partial correlation $\rho_{ij.k}$ between X_i and X_j given all the remaining variables $k = \{1, ..., p\} \setminus \{i, j\}$ is expressible via elements of the concentration matrix or via elements of the conditional covariance matrix of X_i and X_j given $k = \{1, ..., p\} \setminus \{i, j\}$. The following results (see e.g. Wermuth, 1976) hold:

$$\rho_{ij} = \sigma_{ij}(\sigma_{ii}\sigma_{jj})^{-1/2}$$
$$\rho_{ij.k} = -\sigma^{ij}(\sigma^{ii}\sigma^{jj})^{1/2} = \sigma_{ii.k}/(\sigma_{ii.k}\sigma_{jj.k})^{1/2}.$$

As the partial regression coefficients between X_i and X_j given $k = \{1, ..., p\} \setminus \{i, j\}$ is $\beta_{ij.k} = \sigma_{ij.k}/\sigma_{jj.k}$, the partial correlation coefficient relates to it in the following way

$$\rho_{ij.k} = \beta_{ij.k}(\sigma_{jj.k}/\sigma_{ii.k})^{1/2}.$$

For normal distributions independence is equivalent to zero correlation and conditional independence just means zero partial correlation. Since partial correlations are closely connected to the inverse of the covariance matrix the following lemma can be proved [cf. Speed and Kiiveri (1986) and Whittaker (1990) and also Anderson (1958), or Rao (1973)].

Lemma 1.1. *Suppose* $X = \{X_i, i \in V\} \sim N(0, \Sigma)$ *and* a, b *and* c *are pairwise disjoint subsets of* V *(*a *and* b *non-empty), such that* $V = a \cup b \cup c$. *Then:*

(i) the marginal distribution of X_a *is* $N(0, \Sigma_{aa})$;

(ii) the conditional distribution of X_a *given* X_c *is* $N(\Sigma_{ac}\Sigma_{cc}^{-1}X_c, \Sigma_{aa.c})$, *where* $\Sigma_{aa.c} = \Sigma_{aa} - \Sigma_{ac}\Sigma_{cc}^{-1}\Sigma_{ca}$;

(iii) $X_a \perp\!\!\!\perp X_b | X_c \Leftrightarrow \Sigma_{ab.c} = 0 \Leftrightarrow \Sigma^{ab} = 0$.

The statements of the lemma refer to the simultaneous distribution of all variables. The lemma shows that the conditional independence statements concerning the variables if they have a Gaussian distribution can be interpreted in terms of zero partial correlations if it can be assumed that the variables have a finite covariance matrix.

The models with patterns of zero in the correlation matrix are defined as *concentration graphs* where $G_{con}(V, E_{con})$ is the pair set of vertex V associated with X_v and E_{con} is a set of undirected edges such that there is no edge joining two nodes u and s whenever X_u is independent of X_s given all the other variables. Edges in a concentration graph are represented here by full lines, i.e. $u - s$. The *concentration graph* satisfies the undirected pairwise Markov property. In the multivariate normal distribution it follows that a G_{con} can be defined by setting to zero a specified off-diagonal element of the inverse of the variance matrix, such models were proposed by Dempster (1972). The reader is referred to Edwards (2000), Lauritzen (1996) and Whittaker (1990) for statistical properties of these models.

The models with patterns of zeros in marginal correlations had been introduced as linear in covariance models by Dempster (1972) and have more recently been called *covariance graph models* by Cox and Wermuth (1993, 1996). They use an undirected graph with dashed lines to represent marginal association among variables. A *covariance graph* $G_{cov} = (V, E_{cov})$ is the pair of set V of vertices associated with X_v and a set E_{cov} of undirected edges such that there is no edge joining nodes u and s whenever X_u is marginally independent from X_s. Edges in a covariance graph are represented by

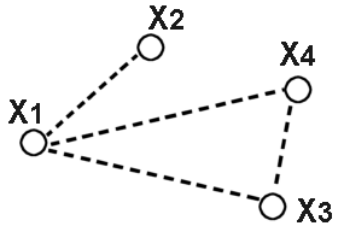

Figure 1.3: *Covariance graph between four observed variables; an edge missing between two nodes means pairwise marginal association.*

dashed lines, i.e. $u - -s$ or by bi-directed edges represented by i.e. ($\longleftrightarrow$) (Drton and Richardson, 2003). For such graphs Kauermann, (1996) studied pairwise, local and global Markov properties (see also Benerjee and Richardson, 2003). Such properties are dual to previous Markov properties associated with undirected graphs.

Figure 1.3 illustrates a covariance graph corresponding to a covariance matrix where there is a structural zero in position $(2, 3)$ and $(2, 4)$, which has therefore the interpretation of linear marginal independence of variable pair X_2, X_3 and X_2, X_4.

1.3 General structural linear equations model

Structural equation models, referred as SEM, are used to represent relations between several response and explanatory variables, some of which may be hidden and some of which may mutually influence one another. Most commonly SEM have been assumed linear but there are important exceptions (e.g. Goldberger and Ducan, 1973). Here we restrict ourselves to the class of linear SEM (LSEM). They have two parts: a system of Gaussian variables (also called systems of linear simultaneous equations) and a path diagram corresponding to the functional composition of variables specified by the structural equation and the correlations among the error terms (Bollen, 1989).

One of the most common specifications of a linear SEM is the so called LISREL (LInear Structural RELations) model (cf. Jöreskog and Sörbom, 1989). A LISREL model has two types of variables: *exogenous*, if they are not influenced by any other variable in the model (they are free variables whose distribution is arbitrary) and *endogenous* if their distributions are determined by those of the exogenous variables. Variables that have been directly observed and measured are called *manifest* variables whereas variables that are hypothesized to have a role in the model but have not been directly observed or measured are called *latent*. The system has the following structure

$$\eta = B\eta + \Gamma\xi + \zeta$$
$$y = \Lambda_y\eta + \epsilon$$
$$x = \Lambda_x\xi + \delta.$$

The first equation is called the *structural equation* because it is the structural part of the model and it describes the relationships between the latent variables. The last two are called the *measurement equations* because they compose the measurement model which describes how each latent variable is measured by the corresponding manifest indicator. Here η and ξ are unobserved endogenous and exogenous random vectors. The unobserved endogenous random variables are only partially explained by the model and ζ is the unexplained component (a random disturbance in the equation), with covariance matrix $\text{cov}(\zeta) = \Psi$. In the last two equations y and x are observed random indicator and ϵ and δ are unobserved random errors or residual vectors. The assumptions of the model can be summarized as follows:

(i) the error variables ζ, ϵ, δ and ξ are mutually uncorrelated,

(ii) $(I - B)$ is non singular.

The fact that the error variables are mutually independent allows the possibility that there are nonzero correlations within the sets of ξ, ϵ, and δ variables or between the ϵ and δ variables. For such reasons LISREL model os a flexible model.

If qualitative variables are present in the model they are usually assumed to be generated by dividing an underlying normal distribution into two or more classes. This implies that interactive effects of two or more variables cannot be represented.

The equations of the LISREL model can be written in block matrix form

$$\begin{bmatrix} \eta \\ y \\ x \end{bmatrix} = \begin{bmatrix} B & 0 & 0 \\ \lambda_y & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \eta \\ y \\ x \end{bmatrix} + \begin{bmatrix} \Gamma \\ 0 \\ \lambda_x \end{bmatrix} \xi + \begin{bmatrix} \zeta \\ \epsilon \\ \delta \end{bmatrix}.$$

If the difference between observed and unobserved variables is neglected a general recursive or non recursive simultaneous equations system of p endogenous and q exogenous variables is the following:

$$Y = BY + CX + E,$$

where X is a vector of q observed or unobserved *exogenous* variables, Y is a vector of p observed or unobserved *endogenous* variables, E is a vector of p unobserved *errors* or

residual variables, B is a $p \times p$ matrix such that $(I - B)$ is invertible, and C is also a $p \times p$ matrix.

It is assumed that:

(*i*) $(E, X) \sim N_{p+q}(0, \Omega)$, i.e, (E, X) has a $(p + q)$-dimensional normal distribution with expectation 0 and (positive definite) covariance matrix Ω,

(*ii*) $\mathrm{Cov}(E_i, X_j)$=0 for $i = 1, ..., p$ and $j = 1, ..., q$. Hence E and X are uncorrelated and have covariance matrix Ψ and Φ, respectively.

In matrix notation the last assumption is

$$\Omega = \begin{bmatrix} \mathrm{cov}(E, E) & \mathrm{cov}(E, X) \\ \mathrm{cov}(X, E) & \mathrm{cov}(X, X) \end{bmatrix} = \begin{bmatrix} \Psi & 0 \\ 0 & \Phi \end{bmatrix}.$$

From this specification it follows that $(Y, X) \sim N_{p+q}(0, \Sigma)$. The next lemma from Kiiveri, Speed and Carlin (1984) gives expressions for Σ and Σ^{-1}.

Lemma 1.2. *Let $Y = BY + CX + E$ be a simultaneous equations system which satisfies (i) and (ii). Then the over all covariance and concentration matrix of X and Y is*

$$\Sigma = \begin{bmatrix} (I - B)^{-1}(\Psi + C\Phi C')(I - B')^{-1} & (I - B)^{-1}C\Phi \\ \Phi C'(I - B')^{-1} & \Phi \end{bmatrix},$$

$$\Sigma^{-1} = \begin{bmatrix} (I - B')\Psi^{-1}(I - B) & (I - B')\Psi^{-1}C \\ -C'\Phi^{-1}(I - B) & C'\Phi^{-1}C + \Phi^{-1} \end{bmatrix}.$$

The matrices B, C, Φ and Ψ contain the coefficients of the linear equations system, and are called the parameter matrices of the model.

We refer to the equation system as *recursive* if the endogenous variables can be ordered in such a way that B becomes a lower (or upper) triangular matrix. For this reason in recent literature they are also called a *triangular system* (Wermuth and Cox, 2004). Some authors only use the term recursive if B is lower triangular and the disturbances for one equation are uncorrelated with the disturbances of the other equation: the covariance matrix of the errors Ψ is diagonal matrix; if Ψ is not diagonal, but B is lower triangular, they call the system *partially recursive* (Bollen, 1990 pp. 81). In such a system feedback relations are absent, i.e. there are no cycles. Otherwise the LSEM is called *non-recursive*.

It is a common practice to draw a *path diagram* (Wright, 1934) of the simultaneous system. As Jöreskog and Sörbom (1989) remark, this has at least two advantages.

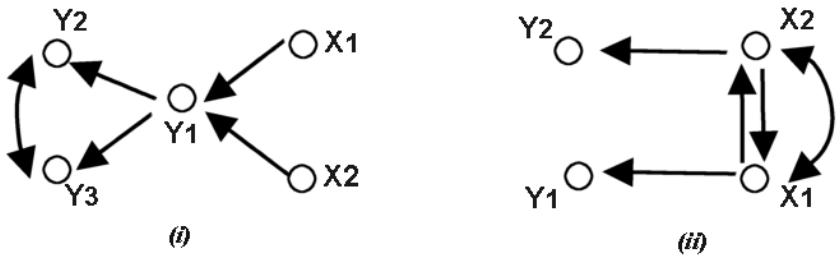

Figure 1.4: *(a) Structural model with correlated residuals; (b) Structural model with feedback loops and correlated residuals.*

First, "the path diagram effectively communicates the basic conceptual ideas of the model." Second, "if the path diagram includes sufficient detail, it can represent exactly the corresponding algebraic equations and the assumptions about the errors terms in these equations" and it portrays also the correlational assumptions about the exogenous variables.

In the path diagram points correspond to variables: sometimes the observed variables are enclosed in boxes and the latent variables are circles, whereas arrows correspond to free parameters in the LSEM. The arrows occurs in the following way: each free parameter in B corresponds to an arrow between two endogenous variables e.g $Y_j \rightarrow Y_i$ if $b_{ij} \neq 0$. Each free parameter in C corresponds to an arrow from an exogenous variable to an endogenous variable, e.g. $X_j \rightarrow Y_i$ if $c_{ij} \neq 0$. Each free parameter in Φ corresponds to bi-directional (two-headed) arrow between two exogenous variables, e.g. $X_i \leftrightarrow X_j$, if $\phi_{ij} = \phi_{ji} \neq 0$. Each free parameter in Ψ correspond to a bi-directional arrow between two error variables, e.g. $E_i \leftrightarrow E_j$ if $\psi_{ij} = \psi_{ji} \neq 0$. These conventions are illustrated by the path-diagram in Figure 1.4 where four observed variables represent: (a) a recursive system with correlated residuals between the endogenous variables Y_1 and Y_2 (b) a non-recursive system with feed-back loops and correlated residuals between X_1 and X_2.

In general if there is a missing edge in the path diagram of a linear structural equation no correlation coefficient of the associated variable pair is implied to be zero, neither the marginal correlation nor any partial correlation. In some structural equation models, however, an independency interpretation is possible: univariate recursive regressions with independent residuals, multivariate regressions and seemingly unrelated regressions can all be regarded as special cases of LSEM for which a graphical representation is possible.

Non-recursive structural equation models are not DAG models and the joint dis-

tribution is no longer specified in terms of the product of conditional distribution of the children given the parent. They are instead represented by directed *cyclic* graphs (Spirtes, 1995). In extending the framework of graphical models also to such SEM with feedback relations among variables, Koster(1996) and Spirtes (1995) proved independently that the set of conditional independence relations and zero partial correlations entailed by a SEM can be read off from the d-separation relations in the associated graph, even in the case of cyclic graphs. In other words d-separation in a cyclic graph still implies independencies in the joint generating distribution, provided that the relations are linear.

The graph in Figure 1.4 (a) is not strictly a directed graph because of the curved line between Y_2 and Y_3 which indicates that error variables ϵ_2 and ϵ_3 are correlated. The error variable for a variable Y represents the sum of all causes of operating on Y other than the substantive variables explicitly included in the model. As illustrated by Richardson and Spirtes (1999) it is possible to convert SEMs with correlated errors into SEMs without correlated errors by adding a latent common cause of the appropriate substantive variables and replacing the previously correlated error variables with uncorrelated ones.

1.4 Linear recursive structural equations

We consider a linear structure equation system associated with a DAG G in N nodes $N = (1, ..., d_N)$ having node i corresponding to a random variable X_i. Where there exists qualitative prior information that specifies constraints on the ordering of random variables, the joint distribution of the observable is not the primitive notion but rather the end result of the specification of a collection of local conditional distributions. More precisely we consider it as the *data generating process* for a DAG model a system of linear recursive regression equations of the form

$$X_v = f_v(X_{\mathrm{pa}(v)}, \epsilon_v) \qquad v \in N$$

where the assignments have to be carried out sequentially, in a well-ordering of the directed acyclic graph G so that at all times, when X_v is about to be assigned a value, all variables in $\mathrm{pa}(v)$ have already been assigned a value. These are variables assumed to be of substantive importance for predicting X_v. The DAG prescribes a stepwise process for generating the distribution where a proper dependence of X_i is to be only on its potentially explanatory variables. The variables ϵ_v or "disturbances" are assumed to be independent; the system is called *recursive* or a *univariate recursive regression system* or a *triangular system.* Often assembling evidence of various kinds gives *"substantive*

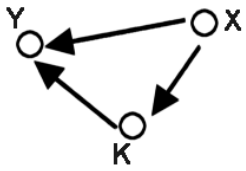

Figure 1.5: *DAG with three variables representing a structural equation system*

knowledge strong enough to relate a response of primary interest via a sequence of intermediate single variables to a purely explanatory variable" (Wermuth, 1999). Such a system is well suited to express a substantive research hypothesis either formulated at the start of an investigation or developed during statistical analysis.

The triangular system is a set of equations of the form

$$\begin{aligned} X_1 + a_{12}X_2 + a_{13}X_3 + ... + a_{1d_N}X_{d_N} &= \epsilon_1 \\ X_2 + a_{23}X_3 + ... + a_{2d_N}X_{d_N} &= \epsilon_2 \\ &\vdots \\ X_{d_N} &= \epsilon_{d_N}, \end{aligned}$$

where X_i is a dependent or endogenous variable for $i = 1, ..., k$ and it is to be determined, i.e. exogenous, for $i = k + 1, ..., d_N$.

As noted in Koster (1999) there are some differences between methods for portraying the relationships between variables by means of a path diagram or graph but in the case of the path diagram of a *recursive* LSEM for Gaussian variables it can be given a consistent interpretation as a graphical model. Kiiveri and Speed (1982) have shown that if the error terms are jointly independent, then any distribution that forms a linear recursive SEM with a directed graph G satisfies the local directed Markov properties for G. "*An edge, missing in the graphical representation, corresponds to a defined conditional independence between two variables and each edge present to a conditional dependence of substantive interest between one variable regarded as explanatory, all taken conditionally on the other explanatory variables.* (Cox and Wermuth, 1996). Assuming uncorrelated exogenous and error variables, Lauritzen *et al.* (1990) have strengthened the results of Kiiveri, Speed and Carlin (1984) such that more Markov properties can be read off the path diagram.

To an LSEM were often given *causal* interpretations, (see e.g Goldberger, 1972)

and the methodology is often called "causal modelling" (see e.g. Blenter 1980). Such types of graph have been called also *causal graphs* (Kiiveri and Speed, 1982) because of the causal interpretation between random variables that can be given to the edges of the graph, due to the fact that each variable is indexed by time. In such a context researchers decompose the effects of one variable on another into direct, indirect and total effects. An edge from X to Y in G means that X is a direct cause of Y relative to the set of vertices V and it expresses the direct effect of X on Y holding constant the other variables in the model. This is a way of interpreting the associations and the dependencies arising in the system. The same causal meaning has been used recently in epidemiological literature (Greenland *et al.* 1999). We take the view of the authors which argue that any causal interpretation should be conducted with extreme caution in the context of observational studies as has been stressed by Guttmann (1977), Cliff (1983), Holland (1988) and Sobel (1995), among others.

The system corresponding to Figure 1.5, with just three variables measured as deviation from their means, is the following

$$\begin{aligned} Y &= \beta_{YK.X}K + \beta_{YX}X + \epsilon_Y \\ K &= \beta_{KX}X + \epsilon_K \\ X &= \epsilon_X. \end{aligned}$$

A measure of the total (overall) effect of X on Y is the marginal regression coefficient between Y and X, marginalizing over K. Just from linearity properties of the system it can be calculated as the sum of effects of "two paths"

$$\beta_{YX} = \beta_{YX.K} + \beta_{YK.X}\beta_{KX},$$

the coefficients for the direct effect of X on Y and terms which can be identified as the indirect effects along each distinct path from X to Y.

Similarly the regression coefficient between Y and K

$$\beta_{YK} = \beta_{YK.X} + \beta_{YX.K}\beta_{KX}.$$

As can be seen, the linear dependence of Y on X is unaltered in slope if and only if $\beta_{YX.K} = 0$ or $\beta_{KX} = 0$, (Cox and Wermuth, 2003). Further the same result holds exactly also for least square estimates of the effects. Spirtes *et al.* (1998) reports a simple rule (Wright, 1934) for calculating the covariances between two variables form a linear recursive system. The covariance between Y and X can be expressed as the sum over all of the collision less paths of the product of the edge labels on the path times the variance of the source node of the path if there is one. For example, in Figure 1.5

$\sigma_{YX} = (\gamma + \xi\alpha)\sigma_{XX}$ where $\gamma = \beta_{YX.K}$, $\xi = \beta_{YK.X}$ and $\alpha = \beta_{KX}$ and hence the partial regression coefficient follows directly $\beta_{YX} = \frac{\sigma_{YX}}{\sigma_{XX}}$.

The arguments we are dealing with apply also to the very much broader family of problems that are called *quasi linear* (Cox and Wermuth, 1996). It means that any dependence present has a linear component and like linear least square regression equations in a multivariate normal framework, any curvature and higher-order interactions present are such that a vanishing linear least-squares regression coefficient implies that no dependence of substantive importance is present.

It is important for interpretation that absence of an edge means an appropriate independence and that presence of an edge implies a dependence strong enough to be of substantive importance. So it must be assumed that there are no *parametric cancellations* (Wermuth and Cox, 1998) or lack of faithfulness in the graph (Spirtes *et al.* 1993). This can occur if the quantitative causal effects of two variables along different directed paths exactly cancel each other out. For example in Figure 1.5 there is no vertex unconditionally d-separated from any other vertex. Assuming that the joint probability distribution over the three vertices is multivariate normal the partial correlation coefficient can be written $\rho_{xy.k} = \rho_{xy} - \rho_{xk}\rho_{yk}/\sqrt{(1-\rho^2_{xk})(1-\rho^2_{yk})}$. It can happen that $\rho_{xy.k} = 0$, because $\rho_{xy} = \rho_{xk}\rho_{yk}$, even though X and Y are not d-separated given K, if the correlations between each pair of variables exactly cancel each other.

Assuming that X is a vector of k mean centered random variables with Gaussian joint distribution with covariance matrix Σ, the recursive system can be written as

$$AX = \epsilon \qquad \text{cov}(\epsilon) = \Delta \tag{1.4.1}$$

where $A = \{-a_{rs}\}$ is upper triangular matrix with ones along the diagonals and with off-diagonal elements corresponding to partial regression coefficients between two variables given the parents, $-a_{rs} = \beta_{rs.pa(r)\setminus s}$ associated with a directed edge between $X_s \leftarrow X_r$; $\Delta = \text{cov}(\epsilon)$ is a nonsingular diagonal covariance matrix of the residuals with elements of partial variances $\delta_{rr} = \sigma_{rr.pa(r)}$ along the diagonal representing the unexplained proportion of the variance of the dependent variable.

Each parameter in the system has a well-understood meaning since it gives for unstandardized variables the amount by which the response is expected to change if the explanatory variable is increased by one unit and all other variables in the equations are kept constant.

From 1.4.1 a triangular decomposition of the covariance matrix Σ and of the concentration matrix Σ^{-1} is given by

$$\text{cov}(X) = \Sigma = (A^{-1})\Delta(A^{-1})', \qquad \Sigma^{-1} = A'\Delta^{-1}A.$$

Whenever there is a zero regression coefficient for variable X_i in equation i this corresponds to the statement that Y_i is conditionally independent of X_j given the directly explanatory variable of X_i.

For joint Gaussian distributions defined by the system this representation implies that every structural zero in A and in A^{-1} is equivalent to a specific independent statement. The linear system that arise can be parameterized in various ways. Different parameterizations typically contain different structural zeros defined as parameters that are implied to be zero in the generating DAG for all members of the family.

Induced graphs

One of the appealing features of studying such a generating process is that it allows different investigation conditioning and or marginalizing over some nodes of the DAG. Induced graphs can be derived starting from such types of DAG and they represent the Markov structure induced by DAG on the observed variables. Two types of such graphs are the induced covariance and concentration graph. The *induced covariance graph* of S given C, $G_{cov}^{S|C}$, is an undirected dashed-line graph of the type described previously in which an edge between i and j is present if and only if $X_i \perp\!\!\!\perp X_j | X_C$ is not implied by the DAG. Similarly, the *induced concentration graph* given C, $G_{con}^{S|C}$, is an undirected full-line graph, in which an edge between i and j is present if and only if $X_i \perp\!\!\!\perp X_j | X_{C \cup S \setminus \{i,j\}}$. Then the first shows the marginal pairwise independencies of variable pairs in S induced after conditioning on C and the latter shows the independencies of variable pairs in S induced after conditioning on C and all remaining variables in S. Such independencies may be derived using separation criteria for DAG and they are not confined to the Gaussian case (Pearl, 1988).

One implication is that when the conditioning set is empty the *overall covariance* graph induced by the DAG has an undirected edge between X_i and X_j if and only if there is a path connecting X_i and X_j which does not contain a collision node. The *overall concentration* graph induced by the DAG has an undirected full line between X_i and X_j if there is an arrow between them in the DAG or X_i and X_j have a common child.

As shown in Section 1.2 an edge missing on such graphs indicates a structural zero in the corresponding induced parameter matrix, i.e. it has a zero in position (i, j) of Σ and of Σ^{-1} respectively. This holds for all linear equations generated over the same graph. Therefore from the first section we know that a missing edge in an induced graph of a Gaussian triangular system indicates both an independence statement and a structural zero correlation in a concentration, covariance or regression coefficient matrix.

In Figure 1.6 an example is presented of the relationship between a DAG and the

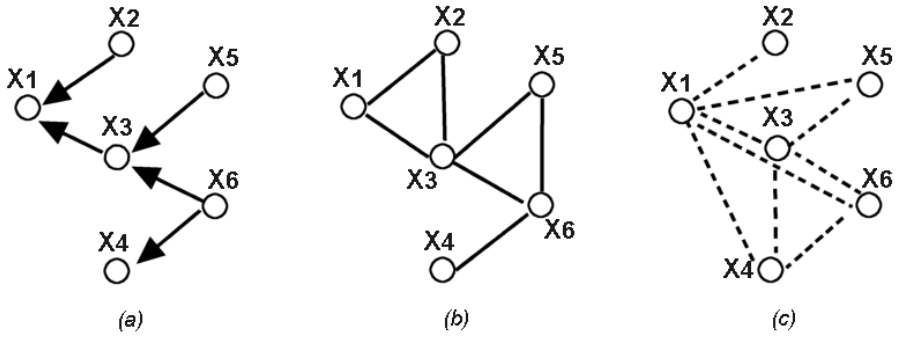

Figure 1.6: *(a) Generating DAG; (b) the overall induced concentration graph G^V_{con}; (c) the overall induced covariance graph G^V_{cov}.*

induced overall concentration and covariance graphs. For instance, a zero element in position $(2, 3)$ in the matrix A, representing the edges of the DAG in Figure 1.6, corresponds to a zero marginal correlation and it leads to a non zero partial correlation corresponding to position $(2, 3)$ of Σ^{-1} as shown by the induced overall covariance and concentration graphs respectively. Wermuth and Cox (2003) derive induced graphs exclusively via transformations of a binary matrix of the starting DAG.

Marginalizing over some nodes of a DAG alters associations among the other variables and typically a graph more complex than the fully directed graph is needed to capture the independence structure because DAG models are a class of graphs that is *not closed under marginalization.* A DAG with hidden variables induces an independence structure over the observed variables that can be represented by a graph called a *summary graph* (Cox and Wermuth, 1996). It gives a summary representation of the system when we marginalize over, or condition on, some variables in the DAG.

Considering a DAG $G = (V, E)$ we derive the independence graph implied for the distribution of X_S, where S is the selected subset of nodes remaining after marginalizing over a subset of nodes M of V, i.e $S = V \backslash M$. The resulting summary graph for the distribution of X_S is denoted by G^S. It may contain three types of edge: directed edge, i.e $(i \prec\!\!\!\!- j)$, dashed line $(i \text{- - - } j)$ or both directed and dashed, i.e $(i \prec\!\!\!\!\text{- - - } j)$. Directed cycles may not occur in a summary graph, but it is possible for there to be a dashed line between i and j and at the same time a directed path from j to i.

By way of illustration we consider the effect of marginalizing over any single node in a DAG. As can be seen from Figure 1.7 the effect of marginalizing over a node is different according as the node is a sink node, or a collision node, or a source node.

Marginalizing over a transition node and over a common source node are all edge-inducing. An induced edge in an independence graph means in general that a specific

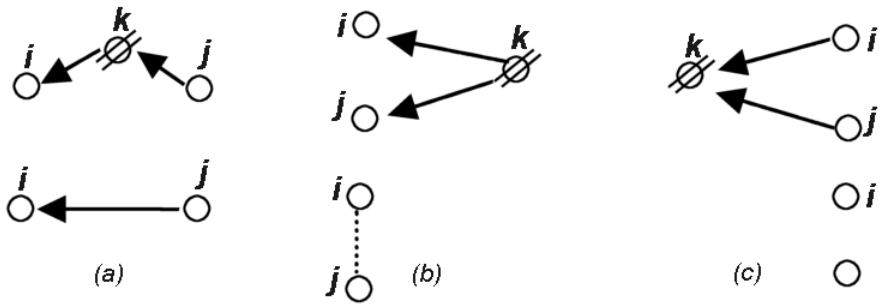

Figure 1.7: *Top: DAG, bottom: summary graphs (a) Effect of marginalizing over transition node, (b) source node, (c) a collision node*

independence statement in the original generating DAG no longer holds in all distributions generated over the new graph. Ignoring the intermediate variable, i.e. marginalizing over k in Figure 1.7(a) leaves i dependent on j. Ignoring the common explanatory variable i.e. marginalizing over k in Figure 1.7(b) leaves the i and j associated. In case (c) instead marginal associations between two explanatory variables cannot happen marginalizing over a child: an effect cannot alter conditions that were determined before the event occurred.

From the illustration of Figure 1.8(a) (see Cox and Wermuth, 1996) it can be seen that when the generating DAG is hypothesized without latent variables the summary graph is the same generating DAG. The overall concentration graph reflects the independencies of $(X_1, X_3) \perp\!\!\!\perp X_4 | X_2$, and the overall covariance graph of $X_3 \perp\!\!\!\perp (X_2, X_4)$.

If in the generating DAG a latent variable is supposed to influence the two primary responses X_1 and X_2, as shown in Figure 1.8(b) the summary graph has a dashed line for X_1 and X_2 and it represents a system of recursive regressions with correlated residuals. The induced covariance graph is independence-equivalent to the summary graph and the induced concentration graph is complete.

In the case of a measurement model as in Figure 1.9, where the influence of two latent variables is hypothesized for three measures, the resulting summary graph is equivalent to a covariance graph and the concentration graph is complete.

As can be seen the summary graph which results from applying the operation of marginalization to the independence model given by the original graph is not always the same independence models. As noted in Wermuth, (1999) different classes of model can be identified which arise by marginalizing over nodes in a directed acyclic graph and which are, in the case of a joint Gaussian distribution, also within the class of linear structural equation models.

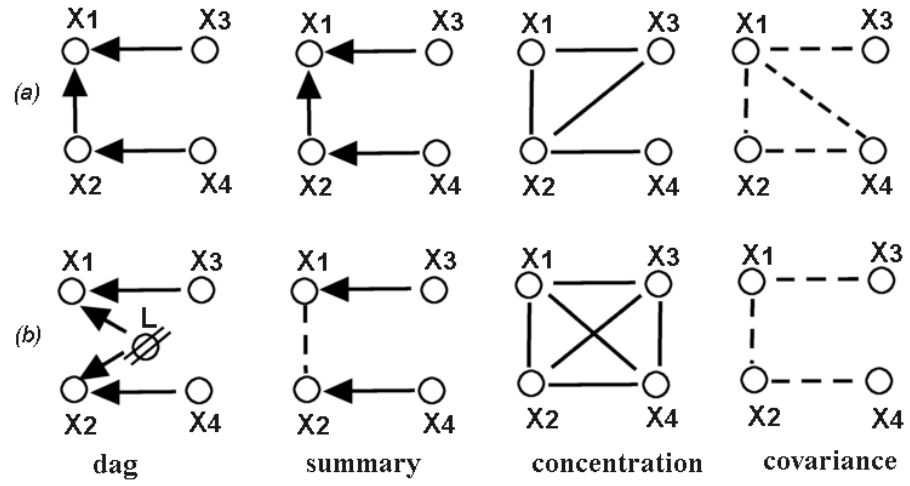

Figure 1.8: *Implications of two DAGs: (a) Derivation of the summary, concentration and covariance graphs; (b) Effect of marginalizing over a source node.*

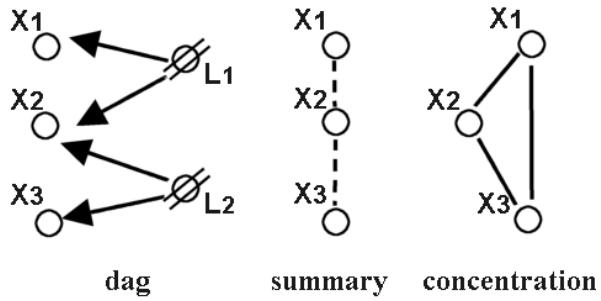

Figure 1.9: *Measurement model; summary graph identical to the induced covariance graph; induced concentration graph.*

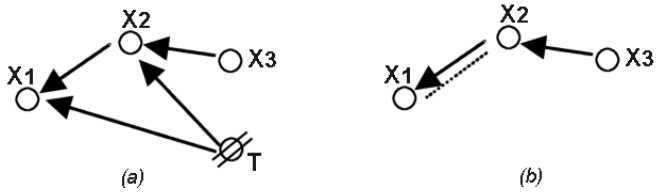

Figure 1.10: *(a) Generating DAG; (b)Corresponding summary graph marginalizing over T.*

For example, considering the DAG in Figure 1.10(a) as a generating process; variable X_3 is called in the econometric literature an *instrumental variable* or *surrogate*. It is an instrument because it allows us to estimate the regression coefficients between two variables X_1 and X_2 in presence of an unobserved confounder T. The instrument is a variable that is marginally correlated with the explanatory variable X_2 and marginally independent of the unobserved variable T. The unobserved confounder is a variable that should have been included in the model but which was not observed, i.e its existence and nature were not appreciated or it was impossible to measure. In the epidemiological literature a confounder is defined to be an unobserved quantity that simultaneously affects the treatment and the response (see also Joffe, 2001). Marginalizing over "T" from the system means that the special independence assumptions made in formulating these equations cannot be empirically tested using only the observed variables and it leads to the summary graph pictured in Figure 6b. A non directional edge in the summary graph is used to indicate that two variables are associated for some reason other that they affect one another. This graph represents a system of linear equations with correlated residuals between X_1 and X_2.

It can be shown that such graphs can also represent the *seemingly unrelated regression equation* (SURE) model by Zellner, (1962). In such models there is interaction between different equations because it is hypothesized that random disturbances associated with some different equations are correlated with each other. In this case the equations are linked statistically, even though not structurally, through the connection of the error distribution inducing the non-diagonality of the associated variance and covariance matrix. For example the summary graph in Figure 1.8(b), linking together the two equations through the dashed line, reflects the link between the error terms of the equation of X_1 and X_2. The generating DAG hypothesises the presence of a hidden common explanatory variable influencing each of the primary response variable and the likelihood that a similar factor may be responsible for the random effects linking the two equations.

1.5 Normal linear factor analysis model

The origins of factor analysis have to be found in Spearman (1904) and an account of his innovative role in the development of the subject can be found in Bartholomew (1995). In the early days of factor analysis, the factor model was used to measure human intelligence. In 1904 Spearman was concerned with the fact that people especially children, who performed well in one test of mental ability also tended to do well in others. This led to the idea that all an individual's scores were manifestations of some

underlying general ability which might be called general intelligence. However, the scores on different items were certainly not perfectly correlated and this was explained by invoking factors specific to each item to account for the variation in performance from one item to another. The aim of the model is to condense a number of variables, often called items, into one or more summary scores. The simplest way in which this might happen would be for the two effects to be independent and additive.

A straightforward generalization of Spearman's *single factor* model is the *common factor model*, which was proposed by Garnett (1919). Suppose that $X_1, ..., X_p$ are continuous variables measured on each individual in a sample from some population. All of $X_1, ..., X_p$ are on an *equal footing*: there is no division into response and explanatory variables. The aim of the study is to "explain" the correlations among the observed variables in terms of a smaller number of unobservable variables also called *common factors*. The general factor model

$$\begin{aligned} X_1 &= u_1 + \lambda_{11}Z_1 + ... + \lambda_{1q}Z_q + E_1 \\ X_2 &= u_2 + \lambda_{22}Z_2 + ... + \lambda_{2q}Z_q + E_2 \\ &\vdots \\ X_p &= U_p + \lambda_{p1}Z_q + ... + \lambda_{pq}Z_q + E_p \end{aligned}$$

or in matrix notation

$$X = \mu + \Lambda Z + E, \tag{1.5.1}$$

where $\mu(p \times 1)$ is a constant term, $\Lambda(p \times p)$ is a matrix of the *factor loadings* which are the same for all individuals, $Z = (Z_1, Z_2, ..., Z_q)'$ are the *common factors* and $E = (E_1, E_2, ..., E_p)'$ are random disturbances specific to each item, they are also called *specific factors*. As can be seen the factor model is a special case of the linear recursive structural equations model where there are no y, η, ϵ, ζ from the system in Section 1.3.

The assumptions are that

i $X_i \sim N_p\ (\mu, \Sigma)$, each X_i follows a multivariate normal distribution;

ii $cov(E, Z) = 0$, the common factors are distributed independently of the specific factors;

iii E(E)= 0, cov(E)= Ψ, with Ψ diagonal, thus they are independent.

The key assumption of the factor model in that $E_1, E_3, ..., E_p$ are uncorrelated thus $X_1, X_2, ..., X_p$ are conditionally uncorrelated given Z. From the assumption

$$X|Z \sim N(\mu + \lambda Z, \Psi),$$

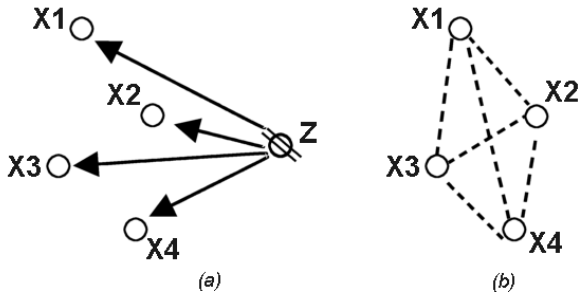

Figure 1.11: *(a)DAG with four observed variables one common factor; (b) resulting summary graph that is a complete covariance graph.*

and if the matrix Ψ is diagonal then it follows

$$X_i \perp\!\!\!\perp X_j|Z \qquad \text{for} \qquad i,j = 1,...,p \qquad i \neq j.$$

As for recursive LSEM the independence structure of the factor model can be captured by a directed acyclic graph. Figure 1.11(a) shows a DAG for four items and one common factor.

The graph says that X_1, X_2, X_3, X_4 are all influenced by Z but they are not directly related to one another. As mentioned before we omit the error terms pointing at the Xs. Figure 1.11(b) shows the summary graph obtained by marginalizing over the common source node: the resulting graph is a full covariance graph where an edge represents the marginal pairwise association of a variable pair (Cox and Wermuth, 1996).

The covariance matrix of X is

$$\Sigma = \text{cov}(X) = \text{cov}(\lambda Z) + \text{cov}(E) = \lambda \text{cov}(Z)\lambda' + \text{cov}(E) = \lambda\lambda' + \Psi. \qquad (1.5.2)$$

The converse also holds: if Σ can be decomposed as in (1.5.2), then the q-factor model (1.5.1) holds. Equation (1.5.2) shows that variances and covariances are linear functions of the parameters. If we allow the latent variables to be correlated, the covariance matrix of the hidden variables will be denoted by $cov(Z) = \Phi$, and the covariance matrix of the observations is induced by the relations between the hidden variables $\Sigma = \lambda\Phi\lambda' + \Psi$.

If Ψ is a diagonal matrix then the factor analysis is scale invariant. In fact, writing $Y = CX$ where $C = \text{diag}(c_1, ..., c_p)$ then $\text{cov}(Y) = (C\lambda)(C\lambda)' + C\Psi C'$ and the q-factor model also holds for Y. So it does not matter whether we work with the covariances or the correlations between $X_1, ..., X_p$ as the same factor model holds. The factor model may therefore be viewed as a particular structure imposed on the covariance matrix Σ, if we consider Σ unstructured there are $p(p+1)/2$ parameters, if we consider Σ structured $(\lambda\lambda' + \Psi)$ there are $pq + p - q(q-1)/2$ parameters, the reduction of

parameters represent the degree of simplification offered by the factor model relative to completely unstructured Σ.

From (1.5.2) the variances of X can be split into two parts

$$\text{var}(X_r) = \sigma_{rr} = \sum_{j=1}^{q} \lambda_{rj}^2 + \psi_r$$

where the first $h_r^2 = \sum_{j=1}^{q} \lambda_{rj}^2$ arises from what is common to all Xs, for which reason it is known as the *communality*, the complementary part, ψ_r, being the variance specific to that particular X_r is known as the *specific* or *unique* variance due to E_r.

Spearman first noted a necessary and sufficient condition for the existence of his general factor model, the so called *tetrad condition* on the correlation matrix. A tetrad condition is an equality among the products of correlations involving a group of four variables. For example, a tetrad condition among W, X, Y, Z is that $\rho_{WX}\rho_{YZ} = \rho_{WY}\rho_{XZ}$. Such a constraint is implied by a model in which there is a single common factor of W, X, Y, Z. In Spearman's case tetrad differences among measures of reading and mathematical aptitude led him to hypothesize that a single common cause, general intelligence, was responsible for performance on all four psychometric instruments.

In 1956 Anderson and Rubin showed that the tetrad condition for a single factor model arises with the column vector λ. If $\lambda\lambda'$ is a positive definite matrix of rank one then the determinant of each of its 2×2 submatrices known as a *tetrad* is zero

$$\begin{vmatrix} \sigma_{il} & \sigma_{im} \\ \sigma_{jl} & \sigma_{jm} \end{vmatrix} = \sigma_{il}\sigma_{jm} - \sigma_{im}\sigma_{jl} = 0$$

this is the *tetrad condition* for unequal i, j, l, m. This generates a block-structure in the covariance matrix.

For example, the model in Figure 1.11 entails the following two tetrad conditions on the covariances (see also Geiger, 1998)

$$\sigma_{12}\sigma_{34} = \sigma_{13}\sigma_{24}$$
$$\sigma_{12}\sigma_{34} = \sigma_{14}\sigma_{23}.$$

In 1931 H.B. Heywood showed that the tetrad condition is a necessary condition for the factor model to hold only because Spearman excluded the case when $\Psi \leq 0$ and this was not an impossible case. There is no inconsistency in the occurrence of a zero residual variance and, taken at its face value, it would simply mean that the variation of the manifest variable in question is wholly explained by the latent variables. In practice this rarely seems plausible (see Bartholomew and Knott, 1999). If one element of Ψ is negative the factor model cannot hold since the variances of the residual vector

would be negative. If a Heywood case arises when the data conform to a linear factor model it would probably be the result of sampling error, a key factor is therefore to have a big sample size. For a given sample size the risk decreases as the number of variables increases. Another cause of the Heywood case is the attempt to extract more factors than are present. This difficulty does not arise if the partial correlations between variables under study are substantial and all positive and if the marginal correlations are close to fulfilling the tetrad conditions (see Cox and Wermuth, 1996).

1.6 Identification and estimation

1.6.1 Identification

A parametric statistical model is identifiable if there is a unique set of model parameters θ that can generate a given distribution. The following definition is useful.

Writing $f(x, \theta)$ for the expression of the likelihood function of the family model considered we define two parameter vectors θ_1 and θ_2 to be *observationally equivalent* if they imply the same $f(x, \theta_1) = f(x, \theta_2)$. The equivalence class at point θ_1 is $\{\theta : \alpha(\theta) = \alpha(\theta_1)\}$.

Definition 1.11. *Local and global identifiability*

(i) A parameter vector $\theta_1 \in \mathcal{A}$ is locally identified if there is an open neighborhood of θ_1 which contains no other θ which is observationally equivalent to θ^1.

(ii) A parameter vector $\theta \in \mathcal{A}$ is globally identified if for any parameter point $\theta^1 \in \mathcal{A}$ there is no other observationally equivalent point $\theta^2 \in \mathcal{A}$.

It should be noted that local identification everywhere in $\mathcal{A}$ is a necessary but not sufficient condition for global identification (see, e.g. Skrondal and Rabe, 2004).

In linear SEM when the latent variables are standardized, the identification problem reduces to assessing whether or not θ is uniquely determined by the covariance matrix Σ of the observed variables. Generally it can be assumed that the distribution of the observed variables is sufficiently well described by the moments of first and second order. In particular, if the distribution is multivariate normal and the observed variables are centered, the second moment order is enough to check model identifiability, and the covariance matrix can be written as a function of a parameter vector, $\Sigma = \Sigma(\theta)$. The parameters in θ are *globally* identified if no vectors θ_1 and θ_2 exist such that $\Sigma(\theta_1) = \Sigma(\theta_2)$ unless $\theta_1 = \theta_2$.

One way to establish identification is algebraically: each element of θ must be solved for in terms of one or more elements of Σ known to be identified. For a review of some algebraic rules of necessary or sufficient conditions for identification see Bollen, (1989). However for complex models this algebraic approach is *"extremely tedious and prone to mistakes"* Bollen, (1989).

Other methods are based on empirical identification which is based on the properties of the estimated parameters. A frequently used rule for identification is based on the estimated information matrix of the second derivatives of the log-likelihood function of the model. Kiiveri (1982) shows that it is sufficient to check the rank of the first derivatives of the covariance matrix of the observed variables with respect to the estimated parameter vector. If this matrix has full rank when evaluated at the solution point then the information matrix is positive definite and the parameters are locally identified at the solution point (e.g. Rothenberg, 1971).

In practice this approach is not feasible because the information matrix is usually analytically intractable in complex models. Another possible solution is to determine the rank by using numerical methods as well as by adding additional constraints needed to obtain local identifiability (see e.g McDonald., 1982).

A necessary condition for identification is that the number of elements of independent parameters θ to be estimated must be less than or equal to the number of unique elements in the sample variance-covariance matrix, as in every linear system. If the non-redundant elements of the variance-covariance matrix are more than the parameters to be estimated, the model is said to be *overidentified*. This gives rise to overidentified conditions on Σ which should hold if the model is true. The overidentified coefficients provide the degrees of freedom for the chi-square test discussed under model evaluation.

The model is *underidentified* when there are fewer non-redundant elements in the variance-covariance matrix than parameters to be estimated and the parameters cannot be consistently estimated; this implies that the degrees of freedom of the model are negative and the model statistics are meaningless. Finally, the model is *just identified* if the number of the parameters to be estimated is the same as the number of non-redundant parameter of the sample covariance matrix. In such case the degrees of freedom of the model are zero.

The *identification problem* of the factor model was discussed in a systematic way by Anderson and Rubin (1956). If the system (1.5.1) or the system (1.5.2) admits a unique solution or a finite number of solutions in $\lambda\lambda'$ and Ψ, given the matrix of the observed variables then the factor model is *globally* identified.

Even if the factor model is identified the solutions cannot be unique for the *indeterminacy* of the factor scores. The subject of factor indeterminacy has a vast history in

factor analysis (Wilson, 1928; Lederman, 1938; Guttman, 1955). It has lead to strong differences in opinion (Steiger, 1979). It is due not simply to measurement errors in the factors, the indeterminacy lies in the fact that it is not known which variables determine the scores. In geometrical terms, factor indeterminacy can be described as follows: if variables are seen as vectors in a vector space, the common factor model postulates a $p + q$ dimensional vector space in which the p observable variables are embedded. The indeterminacy of the common and specific factors results from the fact that it is impossible to identify the $p + q$ basis vectors given p observable variables. Specifically, Guttmann (1955) demonstrated the following lemma:

Theorem 1.1. *Assume that a random vector X satisfies equation (1). Then there exists a random vector W such that*

$$Z = \lambda'\Sigma^{-1}X + W$$

with $E(XW') = 0$ and $E(WW') = I - \lambda'\Sigma^{-1}\lambda$.

A discussion of this issue can be found in Haagen (1992).

One restriction generally imposed for identification is assessing the scale of the latent variables, which is also a way to make them interpretable. The latent variable variance is than fixed to one. Another way is fixing to one a path from the latent variable to one measured variable. For a more detailed discussion see e.g Bollen (1989).

If we restrict ourselves to the single factor model, general results have been established. Anderson and Rubin (1956) gave a simple necessary and sufficient condition to uniquely identify the parameter vector λ.

Theorem 1.2. *A necessary and sufficient condition for identification of a single factor model when Ψ is diagonal is that at least three factor loadings be nonzero.*

Sometimes if the residuals are correlated Ψ is no longer diagonal. This may result from a factor model with more than one latent variable in the case of lack of identification. In this case it is easier to work with the simple factor model with correlated residuals. In fact the concentration matrix of such model has a structure similar to the covariance matrix

$$\Sigma^{-1} = -\delta\delta' + \Psi^{-1} \tag{1.6.1}$$

where $\delta = \Sigma^{xz}\sqrt{\sigma_{xx.z}}$.

A sufficient graphical rule for solving global identifiability of the system has been established by Stanghellini (1997) assessing the structure of zeros in the inverse of the

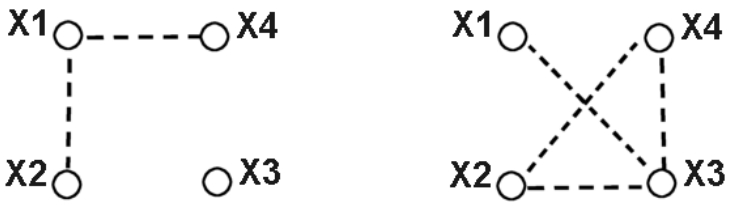

Figure 1.12: *Graph derived after conditioning on Z: (a) possible induced covariance graph of the residual of a single factor model with four observed variables and (b) corresponding complementary covariance graph.*

variance of the residuals Ψ^{-1}. This rule was later proved to be necessary by Vicard (2000). The following lemma (Stanghellini and Wermuth, 2003) based on the properties of the complementary graph of the induced covariance and concentration graphs is a direct consequence of that rule.

Lemma 1.3. *The system (1.5.2) can be solved with respect to* $\lambda\lambda' + \Psi$ *if and only if one of the following conditions holds:*

(i) $\lambda \neq 0$ *and the structure of zeros in* Ψ *is such that every connectivity component of the complementary graph* $\bar{G}$ *of* $G^{X|Z}_{cov}$ *contains an odd cycle;*

(ii) $\delta \neq 0$ *and the structure of zeros in* Ψ *is such that every connectivity component of the complementary graph* $\bar{G}$ *of* $G^{X|Z}_{con}$ *contains an odd cycle.*

In Figure 1.12 is shown the possible induced covariance graph of the residuals obtained conditioning on the common factor, with four observed variables and with correlated residuals between variables X_1, X_2, and X_1, X_4 and the corresponding complementary graph $G^{X|Z}_{cov}$. In such model Ψ has the following structure

$$\Psi = \begin{pmatrix} 1 & \sigma_{12.z} & \sigma_{13.z} & 0 \\ . & 1 & 0 & 0 \\ . & . & 1 & 0 \\ . & . & . & 1 \end{pmatrix}$$

A three-cycle can be seen in $G^{X|Z}_{cov}$ between the variables X_2, X_3, X_4 and this defines the minimum submatrix needed to identify the model.

For the models introduced in Section 1.4 Stanghellini and Wermuth (2004), propose sufficient conditions for global identifiability of DAG models with one unobserved variable. These conditions are formulated in terms of the joint distribution of the variables

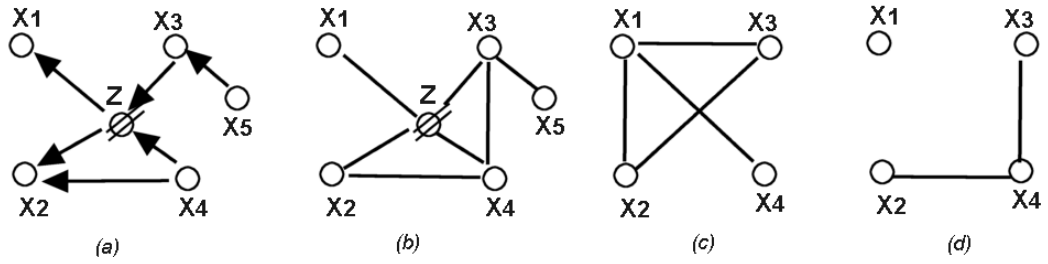

Figure 1.13: *(a) Generating DAG with Z as a node to be marginalized over; (b) the induced overall concentration graph* G^V_{con} *with* $c = bd(Z) = \{X_1, X_2, X_3, X_4\}$*; (c) the concentration graph induced by c* $G^{c|V\setminus c}_{con}$*; (b) the corresponding complementary concentration graph* $\bar{G}$ *of* $G^{c|V\setminus c}_{con}$*.*

and are based based on properties of some conditional independence graph induced by the model. They define a particular class of graphs:

Definition 1.12. *An undirected graph G is G- identifiable if every connectivity component of the complementary graph* $\bar{G}$ *contains an odd cycle.*

Then assuming no parametric cancellations in G_{cov} they derive the following criteria.

Theorem 1.3. *Let* $X = (Y, Z)$ *with marginalization over Z and* $\sigma_{zz} = 1$*. Then the model (1.4.1) is identified if one of the following conditions holds:*

(i) the boundary of the latent variable m, in the covariance graph G^V_{cov} *contains at least three nodes and* $G^{m|Z}$*, the subgraph induced by m in* $G^{Y|Z}_{cov}$ *is G-identifiable;*

(ii) the boundary of the latent variable c, in the concentration graph G^V_{con} *contains at least three nodes and* $G^{m|L}_{con}$*, the subgraph induced by c in* G^V_{con}*, is G-identifiable.*

For example the graph in Figure 1.13 (Stangellini and Wermuth 2004) correspond to an identified model. In this case $c = \{X_1, X_2, X_3, X_4\}$, in *(c)* the subgraph induced by c in G^V_{con} is shown and in *(d)* the complementary graph of this subgraph is presented. The last graph has just one connectivity component which contains the odd cycle formed by $\{X_1, X_2, X_3\}$.

A related concept to the identifiability is that of equivalence which concerns the prospects of distinguishing empirically between different statistical models. If every probability distribution that is generated by one model can be also generated by another, the two models are equivalent and equally well compatible with any data. Equivalence is detrimental when the models are equally compatible with background knowledge, and have the same degrees of freedom. We state the following definition from Pearl (2000, p.145).

Definition 1.13. *Covariance Equivalence*
Two graphical Gaussian linear models are covariance equivalent if and only if they entail the same sets of zero partial correlation. Moreover, two such models are covariance equivalent if and only if their corresponding graphs have the same sets of edges and the same sets of v-configuration.

It provides a test for the equivalence checking the zero partial correlations through the d-separation tests and also through comparison of the corresponding edges and their directionalities. For related topics see also Spirtes *et al.* (1993, 2000) .

1.6.2 Maximum likelihood estimation

Throughout it is assumed that the observed data come from a Gaussian distribution. It is really important that this assumption be checked. Within complete data there are $X = (x_1, ..., x_n)$ independent observations from $N(0, \Sigma)$ and supposing the sample mean vector to be zero, the relevant log-likelihood function can be expressed as

$$l(\Sigma) = \frac{n}{2} \log |\Sigma| + \frac{n}{2} \mathrm{tr}(\Sigma^{-1} S),$$

where S is the sample covariance matrix e.g. Edwards (2000, §3.1). The likelihood equations are the estimating equations obtained by setting the derivatives of the log-likelihood $l(\Sigma)$ with respect to σ_{ij}, $i \longleftarrow j$ to zero. We states the following definition (see e.g. Capitanio *et al.*, 2003, Cox and Wermuth, 1999).

Definition 1.14. *For a family of models specified by a parameter θ taking values in a parameter space Θ, the likelihood of an observed vector x admits a parameter based factorization if $L(\theta; x) = L_1(\theta_1; x) L_2(\theta_1; x)$ where $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ and θ_1 and θ_2 are variation independent, that is $\Theta = \Theta_1 \times \Theta_2$.*

When the variables of the models are all observed the solution of the likelihood equations involves regression of each variable on its direct explanatory variables. This is a well known result from structural equation models (see e.g. Land 1973).

When there are latent variables the maximum likelihood estimators of $(\hat{\Sigma})$ in explicit algebraic form are typically not available. For many models a variety of iterative computational methods for handling unobserved variables are available and solving the likelihood equations, which are the estimating equations obtained by setting the derivatives of the log-likelihood $l(\theta)$ with respect to θ_i and θ_j to zero.

To solve the problem of finding $\hat{\theta} \in \Theta$ such that $l(\hat{\theta}) = \max l(\theta)$ a wide class of algorithms would proceed as follows from a suitable initial value θ_0 of the vector parameter. Given θ_n the $n - th$ approximation to the maximum likelihood estimate θ_n,

let $g_n = \partial l/\partial\theta_n$ the first derivative, or the gradient of the fitting function, evaluated at θ_n and let $A_n = A(\theta_n)$ be a negative definite matrix. Define θ_{n+1} through

$$\theta_{n+1} = \theta_n - c_n A_n^{-1} g_n,$$

where c_n is chosen to maximize $l(\theta)$ in the direction $A_n g_n$ from θ_n, for $c_n \in [-a, a]$ and $a > 0$ fixed throughout.

Many maximization methods are distinguished by their choice of A_n. If the choice of A_n is an identity matrix I this leads to the method of the *steepest ascent*, and the step length can be adjusted on multiplying A_n by a constant c_n to alter the resulting change in θ_{n+1}. The main disadvantage of the steepest ascent method is that it is very slow and it is not very sensitive to the different shapes that the log likelihood may take.

Another choice for A_n is the inverse of the second partial derivatives, or the inverse Hessian matrix, of $l(\theta)$ with respect to or $A_n = [\partial^2 l/\partial\theta\partial\theta']^{-1}$. The choice of A_n is based on a Taylor series expansion of $l(\hat{\theta})$ around θ_n. The Newton-Raphson (NR) algorithm requires the analytic first moreover the second partial derivatives of $l(\theta)$ with respect to θ which may be difficult to calculate and the second derivatives need to be calculated at each step and this can be very time consuming.

Another possible choice of A_n is to use the expected value of the inverse Hessian matrix $A_n = E[\partial^2 l/\partial\theta\partial\theta']^{-1}$ and this is called the Fisher Scoring algorithm. A benefit of the Scoring shared with the Newton's methods is that the inverse of the expected information matrix $E[-\partial^2 l/\partial\theta\partial\theta']^{-1}$ immediately supplies the asymptotic variances and covariances of the maximum likelihood estimates $\hat{\theta}$ (Rao, 1973), since the observed information is under natural assumptions asymptotically equivalent to the expected information.

A modified version of the NR algorithm is often used. The Quasi-Newton methods (Fletcher and Powell, 1963) updates the current approximation A_n to the Hessian matrix by an updating formula $A_{n+1} = A_n - c_n p_n p_n'$ with vector c_n and p_n specified by $c_n = 1/(q_n + A_n s_n)$ where $q_n = (g_{n+1} - g_n)$, $s_n = (\theta_n - \theta_{n+1})$ and $p_n = (q_n + A_n s_n)$, where p_n, q_n and c_n are chosen to increase the function sufficiently in the direction $A_n s_n$. The initial second derivative approximation can be freely specified. In successful applications of quasi-Newton methods choice of the initial matrix A_1 is critical. Setting $A_1 = -I$ is convenient, but often poorly scaled for a particular problem. A better choice is $A_1 = E[\partial^2 l/\partial\theta\partial\theta'|X]$ where X is the observed data. The last choice is generally simple to compute and it does not require the calculation of the second derivative at each step. For example in LISREL (Jöreskog and Sorbom, 1989) program the standard method for fitting models with latent variables is based on such algorithm.

An easier method to implement is the EM algorithm (Dempster, Laird and Rubin,

1977); it is fully described in Chapter 3. If the EM algorithm converges, it converges to a solution of the likelihood equations derived from the log-likelihood function of the observed data. As shown by Kiiveri (1982), the EM algorithm may be slow to converge when there are more latent variables than one. It typically performs well far from the maximum likelihood point and it can be accelerated by using hybrids algorithms that begin as pure EM and gradually make the transition to quasi-Newton methods to perform best near the maximum likelihood point (Lange, 1995a, 1995b).

1.7 Bibliographical note

In his initial exposition of "path analysis" the geneticist Sewall Wright in the early 1920's introduced into statistics the basic idea of directed acyclic graphs whose vertices represent continuous random variables and edges some notions of correlations and causation. One of the appealing features of Wright's method was its ability to produce estimates of path coefficients when some of the variables in the system were not directly observed and some of Wright's examples could be viewed as early attempts at estimating parameters in a model on the basis of missing data approach. The approach was first confined to those problems with essentially linear structures and uncorrelated errors in which the interrelationships are adequately captured by the covariance matrix of the variables. Early descriptions of univariate recursive regressions have been given also by Tinbergen (1937) for the study of the business cycle. Wright's method was early subject to philosophical and methodological criticism; the cause of much misunderstanding seems to be his failure to describe his method in the context of a well defined statistical model. In the 1930s Harold Hotelling invented principal components analysis and Louis Thurstone developed factor analysis. In 1943 Trygve Haavelmo noted an important limitation of univariate linear recursive equations and he developed joint response models with cyclic dependencies. This has led to the overcoming of some of Wright's deficiencies with important developments. The emphasis shifted from being an a posteriori description of an assumed causal process, as Wright viewed his method, to being a tentative test for an assumed causal process. This gave rise to structural equations models (SEM) in econometrics (Goldberger, 1964), linear structural equations models (LSEM); the analysis of covariance structures (Dempster, 1972) in psychometrics (Jöreskog, 1973,1977, 1981) and in sociology (Ducan, 1969; Land, 1973). Covariance selection model were first introduced by Dempster in 1972 and have subsequently been studied by Wermuth (1976a, 1976b), Knuiman (1978) and Speed (1978).

The idea that the probabilistic meaning and implications of structural models are best revealed through an understanding of the independence constraint they impose

seem to have first appeared in Moran (1961) and an approach to causal models in sociology can be found in Blalock (1961, 1971). The conceptual synthesis of models containing structurally related latent variables was developed extensively in sociology during the 1960s and early 1970s. For instance Blalock argued that sociologists should use causal models containing both indicators and underlying variables to make inferences about the latent variables based on the covariances of the observed indicators.

In 1912 Andrej Markov used the notion of conditional independence explicitly to simplify multivariate structures. After the work of Phil Dawid (1979) moving away from the use of SEM Speed Kiiveri and Carlin (1984) preferred to use an appealing factorization of the joint density of the variables under consideration, and the consequent conditional independent constraints on the variables. In fact analysis and interpretation of multivariate data can often be simplified when knowledge about independencies is available or can be derived.

At first the graphs defining graphical Markov models were undirected but soon directed acyclic graphs (DAGs) and chain graphs were used. Chain graphs were introduced by Lauritzen and Wermuth (1989) in which a pair of variables can be linked either by an arrow or by an undirected edge. Undirected and directed acyclic graphs are special types of chain graphs (Lauritzen, 1996, Edwards 2000). Even though the types of graphs encountered in graphical Markov models look quite dissimilar, the way in which for any such graph the associated graphical Markov model is defined, is essentially the same. The vertices of the graph are associated with the random variables that are being studied. For each type of graph a purely graph-theoretical concept of "separation" is introduced, allowing one to state whether or not in a certain graph two given subsets of vertices are separated by a given third subset of vertices. Finally, the graphical Markov model is defined by stipulating that each valid separation statement pertaining to vertices of the graph be mirrored by a valid conditional independent statement pertaining to the associated random variables. Currently there are in the literature two main lines of approach in defining "separation in graphs". One approach, defined by Frydenberg (1990) for the class of chain graphs is based on the operation of "moralization" of subgraphs induced by certain subsets of vertices. The other approach introduced by Pearl (1988) is based on the notion of d-separation and applies to DAGs. In Lauritzen it is shown that both ways of defining separation in graphs are equivalent if the graph is DAG.

Further steps were recently taken by Spirtes (1995) and Koster (1996), who showed how DAGs (in the case of Spirtes) and chain graphs (in the case of Koster) can be generalized so as to permit directed cycles. The work of Koster (1999) provides a probabilistic interpretation for any graph that consists of a finite number of variables

with arrows and undirected edges among them.

In recent times the combination of ideas from the area of graphical Markov models, in which the the Markov properties (i.e., the conditional independence structure of a set of random variables) are accurately portrayed by a certain graph (Frydenberg 1990, Lauritzen and Wermuth 1989, Whittaker 1990) with those from SEM has formed stimulating ground for reinterpreting and enlarging existing results. In fact a graphical approach can be used to solve a number of important problems in SEM; *"the conceptual basis for SEM achieves a new level of precision through graphs. What makes a set of equations "Structural", what assumptions are expressed by the authors of such equations, what the testable implications of those assumption are, and what policy claims a given set of structural equations advertises are some of the questions that receive simple and mathematically precise answers via graphical models"* (Pearl, 2000).

The work of Pearl has suggested that the DAG can also be used to calculate the effects of intervening on an existing causal system by manipulating the values of variables and also it a useful tool to compare the outcomes that would arise under different interventions. As noted by Dawid (2002), this view coincides with enhancing a probabilistic model described by a DAG with further causal interpretation that is not implicit in its nature. Dawid calls the graph with this modified semantics as an *intervention* DAG. A review of graphical causal modelling for epidemiological research can be found in Greenland (2000) and Greenland *et al.* (1999).

When we deal with a latent variable in graphical models we introduce a graph in which some of the nodes are denoted specially to show that they are not observed, in this work following Cox and Wermuth (1996) they are denoted in the graph by a double crossing over the nodes. Latent variables are variables over which we marginalize and the possible impact on the form of the conditional density has to be considered.

In the recent literature on graphical models two other approaches to the problem of marginalizing and conditioning in graphical models have been proposed besides the *summary graph* (Cox and Wermuth 1996; Wermuth and Cox 1998). One considers *ancestral graphs* (Richardson and Spirtes, 2002); and the the other is based on *MC graphs* (Koster, 2002). As noted in Koster (2002) the operation of marginalization and conditioning are defined differently for each type of graph thus after marginalizing on a given dag over a subset of vertices $m \subseteq V$ and or conditioning on a subset $c \subseteq (V \backslash m)$ each of the three approaches may render a distinct (thought separation-equivalent) graphical object.

1.8 References

Anderson T. W. (1958). *An introduction to multivariate statistical analysis.* Wiley, New York.

Anderson T.W., Rubin H. (1956). Statistical Inference in Factor Analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5, 111-150. Univ. California Press, Berkeley.

Andersson S., Madigan D., Perlman M., Triggs C. (1995) On the relations between conditional independence models determined by finite distributive lattices and by directed acyclic graphs. *Journal of Statistical Planning and Inference*, 48, 25-46.

Bartholomew D. (1995). Spearman and the origin and develoment of factor analysis. *British Journal of Statistical and Mathematical Psychology*, 48, 211-220.

Bartholomew D. J., Knott M. (1999). *Latent variables models and factor analysis.* Kendall's Library of Statistics, London: Arnold.

Banerjee M., Richardson T. (2003): On a Dualization of Graphical Gaussian Models: A Correction Note. *Scandinavian Journal of Statistics*, 30, 4, 817-821.

Blalock H. (1971). *Causal Models in the Social Science.* Adline-Atherton, Chicago.

Blalock H. (1961). *Causal Inferences in Nonexperimental Research.* Chapel Hill: University of Noth Carolina Press.

Blenter P. M. (1980). Multivariate analysis with latent variables: causal modeling. *Annual Review of Psychology*, 31, 419-456.

Bollen K. A. (1989). *Structural equations with latent variables.* Wiley, New York.

Capitanio A., Azzalini A., Stanghellini E. (2003). Graphical Models for skew-normal variates. *Scandinavian Journal of Statistics*, Vol. 30, 129-144.

Castelo J. R. V. (2002). *The Discrete Acyclic Digraph Markov Model in Data Mining.* PhD Thesis, Utrecht University.

Cliff N. (1983). Some cautions concerning the application of causal modelling methods. *Multivariate Behavioral Research* 18, 115-126.

Cox D. R., Wermuth N. (1993). Linear dependencies Represented by Chain Graphs. *Statistical Science*, 8, No. 3, 204-283.

Cox D. R., Wermuth N. (1996). *Multivariate Dependencies - Models, Analysis and interpretation.* London: Chapman & Hall.

Cox D. R., Wermuth N. (1999). Likelihood factorizations for mixed discrete and continuous variables. *Scandinavian Journal of Statistics*, 26, 209-220.

Cox D. R., Wermuth N. (2003). A general condition for avoiding effect reversal after marginalization. *J. R. Statist. Soc. B* 4, 937-941.

Dawid A. P. (1979). Conditional independence in statistical theory (with discussion). *Jour. Royal Stat. Society Ser. B*, 41, 1-31.

Dawid A. P. (2002). Influence diagrams for Causal Modelling and Inference. *International statistical review* 70, 2, 161-189.

Dempster A.P. (1969). *Elements of Continuous Multivariate Analysis.* Reading: Addison Wesley.

Dempster A. P. (1972). Covariance Selection. *Biometrics*, 28, 157-175.

Dempster A. P. Laird N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statististical Society* B, 39, 1-38.

Drton M., Richardson T. S. (2003).A New Algorithm for Maximum Likelihood Estimation in Gaussian Graphical Models for Marginal Independence. *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 184-91.

Duncan O. D. (1969). Some linear models for two-wave, two variable panel analysis. *Psychological Bulletin*, 72, 177-182.

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

Frydenberg M., Lauritzen S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, 76, 539-555.

Frydenberg M. (1990). The chain graph Markov property. *Scand. J. Statist.*17, 333-353.

Garnett J. C. (1919). General ability, cleverness and purpose. *British Journal of Psychiatry*, 8, 345-366.

Geiger J. C. (1998). Graphical Models and Exponential Families. *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, 156-165.

Goldberger A. S. (1964). *Econometric Theory.* New York: Wiley.

Goldberger A. (1972). Structural equation methods in the social sciences. *Econometrica* 40, 979-1001.

Goldberger A., Ducan O. (1973). Structural equation models in the social sciences (Seminar Press, New York).

Greenland S. (2000). Causal analysis in the health sciences. *Journal of the American Statistical Association*, 95, 286-289.

Greenland S., Pearl J., Robins JM. (1999). Causal diagrams for epidemiologic research. *Epidemiology* 10, 37-48.

Guttmann L. (1955). The determinacy of factor score matrices with implication of five other basic problems of common-factor theory. *Journal of Mathematical and Statistical Psychology* 8, 65-81.

Guttmann L. (1977). What is not what in statistics. *The Statistician* 26, 81-107

Haagen K. (1992). Il problema dell'indeterminatezza nei modelli con variabili latenti. *Statistica* 3, 365-377.

Haavelmo T. (1943). The statistical implications of a system of simultaneous equations. *Econometrika* 11, 1-12.

Heywood H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Statistical Society, London*, 134, 486-501.

Holland P. W. (1988). Causal inference, path analysis, and recursive structural equation models. In C. C. Clogg, editor, *Sociological Methodology*, 449-493. American Sociological Association, Washington.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417-441, 498-520.

Joffe M.M., Byrne C., Colditz G. A. (2001). Postmenopausal hormone use, screening, and breast cancer: characterization and Control of a Bias. *Epidemiology*, 12, 4, 429-438.

Jöreskog K. G. (1973). A general method for estimating a linear structural equation system. In *Structural Equation Model in the Social Science*(A. S. Goldgerger and O. D. Duncan, eds.) 85-112. Seminar Press, New York.

Jöreskog K. G. (1977). Structural equation models in the social sciences: specification, estimation and testing. In P. R. Kirshnaiah (Eds.) *Applications of statistics*, 265-286, Amsterdam, North Holland.

Jöreskog K. G. (1981). Analysis of covariance structures. With discussion. *Scandinavian Journal of Statistics*, 8, 65-92.

Jöreskog K., G. Sorbom D. (1989). *LISREL 7 - A guide to the Program and Applications*. 2nd Edition, SPSS Publications, Chicago.

Kauermann G. (1996). On a dualization of graphical Gaussian models. *Scandinavian Journal of Statistics*, 23, 105-116.

Kiiveri H. T. (1982). A unified approach to causal models. Ph.D thesis. University of Western Australia.

Kiiveri H. T., Speed T. P. (1982). Structural analysis of multivariate data: a review. In *Sociological Methodology* (S. Leinhardt, ed.) Jossey-Bass, San Francisco.

Kiiveri H. T., Speed T. P., Carlin J. B. (1984). Recursive causal models. *J. Austral. Math. Soc. Ser.* A 30-52

Kiiveri H. T. (1987). An incomplete data approach to the analysis of covariance structure. *Psychometrika*, 52, 539-554.

Knuiman M. (1978). *Covariance selection*. In R.L. Tweedie (Ed.) Proceedings of the conference on Spatial Patterns and Processes. Supplement to Advances in Applied Probability, 10, 123-130.

Koster J.T.A (1996). Markov properties of non recursive causal models. *Ann. Statist.* 24, 2148-2177.

Koster J.T.A. (1999). On the validity of Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand. J. Statist.*, 26, 413-431.

Koster J.T.A. (2002). Marginalizing and conditioning on graphical models. *Bernoulli*, 8, 817-840.

Land K. C. (1973). Identification, parameter estimation and hypothesis testing in recursive sociological models. In A. S. Goldberger and O. D. Duncan (Eds.) Structural equation models in the social science. Seminar Press, New York.

Lange K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society*, B, 57, 425-437.

Lange K. (1995b). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5, 1-18.

Lauritzen S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Lauritzen S., Dawid A., Larsen B., Leimer H. (1990). Independence properties of directed Markov fields. *Networks*, 20, 491-505.

Lauritzen S. L., Wermuth N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17, 31-57.

Lederman W. (1938). The orthogonal transformations of a factorial matrix into itself. *Psychometrika*, 3, 181-187.

Moran P.A.P (1961). *Path coefficient s reconsidered*. Australian Journal of Statistics, 3, 87-93.

Pearl J. (1988). *Probabilistic reasoning in Intelligent Systems*. S. Mateo, CA: Morgan Kaufmann.

Pearl J. (2000). *Causality*. New York: Oxford.

Pearl J., Paz A. (1987). Graphoids: A graph-based logic for reasoning about relevancy relations. In Boulay, B. D. editor, *Advanced in Artificial Intelligence-II*. North Holland.

Pearl J., Verma T. (1987). The logic of representing dependencies by directed graphs. In *Proc. of the conf. of the American Association of Artificial Intelligence*, 347-379.

Rao C. R. (1973). *Linear statistical inference and its applications*. Wiley, New York.

Richardson T., Spirtes P. (1999). Automated discovery of linear feedback models. In *Computation, causation, and Discovery,*(ed. C. Glymour and G. F. Cooper), pp. 253-304. MIT Press, Cambridge, MA.

Richardson T., Spirtes P. (2002). Ancestral Markov graphical models. *Annals of Statistics*, 30, 962-1030.

Rothenberg T. (1971). Identification in parametric models. *Econometrika*, 39, 577-591.

Skrondal A., Rabe-Hesketh S. (2004). *Generalized latent variable modelling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman and Hall, CRC Press.

Sobel M. E. (1995). Causal inference in the social and behavioral sciences. In Arminger A., Clogg C. C., Sobel E. M., editors, *Handbook of statistical modelling for the social and behavioral sciences* 1-38. Plenum Press, New York.

Speed T. P. (1978). *Relations between models for spatail data, contingency tables and Markov fields on graphs*. In R. L. Tweedie (Ed.), Prodeedings of the Conference on Spatial Patterns and Processes. Supplument to Advances in Applied Probability 10, 111-122.

Speed T. P., Kiiveri H. T. (1986). Gaussian Markov distribution over finite graphs. *Annals of Statistics* 14, 138-150.

Spearman C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201-293.

Spirtes P. (1995). Directed cyclic graphical representations of feedback models. In *Uncertainty in Artificial Intelligence* (P. Besnard and S. Hanks, eds.) 491-498. Morgan Kaufmann, San Francisco.

Spirtes P., Glymour C., Scheines R. (1993). *Causation, Prediction and Search*, New York: Springer-Verlag.

Spirtes P., Glymour C., Scheines R. (2000). *Causation, Prediction and Search*,2nd ed. Cambridge, MA: MIT Press.

Spirtes P., Richardson T., Meek C., Scheines R., Glymour C. (1998). Using path diagrams as structural equation modelling tool. *Sociological Methods and Research* 27, 182-225.

Stanghellini E. (1997). Identification of single-factor model using graphical Gaussian rules. *Biometrika*, 84, 241–244.

Stanghellini E., Wermuth N. (2004). On the identification of path analysis models with one hidden variable. *Under revision.*

Steiger J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, 44, 157-167.

Tinbergen J. (1937). *An Econometric Approach to Business Cycle Problems.* Hermann, Paris.

Thurstone L.L. (1938). *Primary mental abilities.* Chicago: University Press.

Vicard P. (2000). On the identification of a single-factor model with correlated residuals. *Biometrika*, 87, 199-205.

Wermuth N. (1976a). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32, 95-108.

Wermuth N. (1976b). Model search among multiplicative models. *Biometrics*, 32, 253-263.

Wermuth N. (1980). Linear recursive equations, covariance selection, and path analysis. *Amer. Statist. Assoc.*,75, 963-972.

Wermuth N. (1999). Effects of an unobserved confounder on a system with an intermediate outcome. ZUMA-Arbeitsbericht 99/07.

Wermuth N. (2003). Analysing social science data with graphical Markov models. In *Highly Structured Stochastic Systems*, P. Green, N. Hjort, Richardson S. (eds.), pp. 45-52. Oxford University Press.

Wermuth N., Cox D. R. (1998). On association models defined over independence graphs. *Bernoulli* 4, 477-496.

Wermuth N., Cox D. R. (2003). On modified triangular systems. *http://psystat.sowi.uni-mainz.de/wermuth/pdfs/papmodtri.pdf*

Wermuth N., Cox D. R. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society*, B, to appear.

Wermuth N., Lauritzen S. (1990). On substantive research hypothesis, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society, Series, B*, 52, 21-50.

Whittaker J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, New York.

Wilson E. B. (1928). On hierarchical correlation systems. *Proceedings, National Academy of Science*, 14, 283-291.

Wright S. (1921). Correlation and Causation. *Journal of Agricultural Research*, 20, 557-585.

Wright S. (1923). The theory of path coefficients: A reply to Niels'criticism. *Genetics* 8, 239-255.

Wright S. (1934). The method of path coefficients. *Annals of Statistics*, 5, 161-215.

Zellner A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* 57, 348-368.

Chapter 2

Latent Class Models

2.1 Latent class analysis

Latent structure analysis arises from models in which the latent variables are categorical and it can be seen as a qualitative version of the factor model for the analysis of qualitative data. The object of latent class model analysis is to define a latent variable, specifically a set of classes of the latent variable, within which the manifest variables are conditionally (or locally) independent. In other words, the association between the manifest indicator *"is explained via a representation in terms of a finite set of latent classes within each of which independency holds"* Cox, (2003).

The latent class model has been considered into two closed related parametrization. The first one (Lazarsfeld and Henry, 1968) is in term of probabilities of belonging to a particular latent class and of obtain a particular scoring pattern on the observed variables, given the latent class one belongs to. The second one was introduced by Haberman (1979), (see also Hagenaars, 1990) is a parametrization in terms of a log-linear model with a categorical latent variable. The basic unrestricted latent class model is identical across the two parameterizations, restricted models with one parametrization may not be readily translated into the alternative parametrization.

For simplicity of exposition, we discuss the probabilistic parametrization of the model following the notation in Bartholomew and Knott, (1999). Suppose that there are p binary observed variables $x_1, x_2, ..., x_p$ with $x_i = 0$ or 1 for all i, $i = 1, ..., p$, which can be collected in a column vector $x = (x_1, ..., x_p)'$. We denote a categorical latent variable by z with levels indexed by r, $r = 1, ..., R$, where R is the number of latent classes. If z can be defined so as to explain the correlations among $x_1, x_2, ..., x_p$ its classes are taken to represent the latent types, as they are defined by the measured variables within the sample population. For this parametrization the assumption of conditional

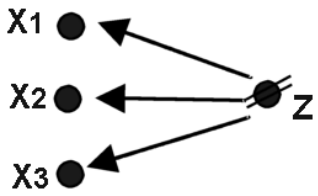

Figure 2.1: *DAG model with one latent variable and three observed binary indicators.*

independence states that the manifest items are independent for individuals with the same position on the latent variable. Such independencies can be encoded in a directed acyclic graphical Markov model, showed in Figure 2.1, of the type equivalent to the factor model, where the variables are depicted by full dots to denote that they are categorical. The example in Figure 2.1 shows that within categories of variable z, the observed variables x_1, x_2 and x_3 are pairwise independent. If x_1, x_2 and x_3 are dichotomous variables there are 2^3 different patterns of responses that can be observed. For example different response vectors are $\{000\}, \{100\}, \{010\}, \{001\}, \{110\}, \{101\}, \{011\}, \{111\}$.

Let $\eta_r = P(z = r)$ be the latent class probability which describes the probability distribution that a random chosen individual is in class r of the latent variable. The number of R classes of the latent variable represents the number of latent types defined by the latent class model for the observed variables. The relative size of each of the R classes provides also significant information for the interpretation of the latent class probabilities. It indicates whether the population is relatively evenly distributed among the R classes, or whether some of the latent classes represent relatively small segments.

A second type of latent class parameter is the conditional probability which is comparable to the factor loading in factor analysis. Let $\pi_{ir} = P(x_i = 1|z = r)$ denote the probability of an individual being in class r of the latent variable with a positive response of the observed variable x_i. For each of the R classes of the latent variable, there is a set of conditional probabilities one for each level of the observed variables. Following the example in Figure 2.1 the R latent classes of z will have three sets of conditional probabilities $(\pi_{1r}, \pi_{2r}, \pi_{3r})$. Within each of the latent classes the conditional probabilities indicate whether observations in class r are likely or unlikely to have the characteristics of each of the observed variables, such probabilities must sum to one. Consequently there is one redundant conditional probability for each observed variable within each latent class. For example if x_1, x_2 and x_3 have I, J and K categories respectively, there are $(I-1)+(J-1)+(K-1)$ conditional probabilities which need to

be identified for each of the R latent classes.

The conditional probability of observing a response pattern x can be expressed as the product of the conditional response probabilities for the separate items. For this model

$$f(x) = \sum_{r=0}^{R-1} \eta_r \prod_{i=1}^{p} \pi_{ir}^{x_i} (1 - \pi_{ir})^{1-x_i}. \tag{2.1.1}$$

A component of latent class analysis is predicting latent class membership for cases showing various observed response vectors. The posterior probability that an individual with response vector x belongs to category r is thus

$$h(r|x) = \frac{\eta_r \prod_{i=1}^{p} \pi_{ir}^{x_i} (1 - \pi_{ir})^{1-x_i}}{f(x)} \qquad r = 0, \ldots, R-1. \tag{2.1.2}$$

This function is used to construct an allocation rule according to which an individual is placed in the class for which the posterior probability is greatest.

2.1.1 Identification, estimation and testing

A difficult problem in latent class models is the identifiability of their parameters. As illustrated in the previous chapter, at paragraph 1.6.1, there are various notions concerning identification in the literature. A latent class model is *globally* identified if there is a *unique* latent structure which generates the manifest probabilities.

A necessary but not sufficient condition for identifiability is that the number of parameters to be estimated independently is not greater than the number of independently observed frequencies. In fact, if there are more model parameters than there are independent cell probabilities then there will be many sets of model parameters leading to the same likelihood. The number of parameters in the model is given by $P = \{R \sum_{j=1}^{t} (c_j - 1) + (R - 1)\}$ where $c_1 \times c_2 \times ... \times c_j$ represents the dimension of the contingency table.

Only for some simple cases an analytic solution has been provided to check identifiability. For instance, the conditions for the parameters to be identifiable, are known for dichotomous observed data and one binary latent variable: the upper limit of the number of classes to be identified, in the absence of any restrictions, is two when there are four dichotomous variables. In fact, the proof that the parameters in a three-class latent class model with four observed variables, are not identifiable was given by Lazarsfeld and Henry (1968, p.65). However, as pointed out by Goodman (2002, p.55), there is an error in the Lazarsfeld and Henry proof. Three latent classes become identifiable when there are five dichotomous variables, in which case the maximum number of identifiable

classes is limited to four (cf. Lazarsfeld and Henry, 1968). Further, it is known that for two-dimensional contingency tables not all parameters for the two class model are identifiable (Clogg, 1981).

As pointed out by Formann (2003), except for some few cases, it not possible to say a priori whether these models may be identifiable. Some criteria for local identifiability "*become available after one has estimated the parameters under a certain model, because identifiability via the parameter estimates also depends on the specific data under investigation.*" (Formann, 2003).

Goodman (1974) made some progress with *local* identifiability, which means that in a small neighborhood the maximum of the likelihood is unique. He stated that the maximum is unique if the transformation of the cell probabilities to the parameters is one to one in a small neighbourhood. If the Jacobian of the transformation, which is the matrix of the partial derivatives of $f(x)$ with respect to the model parameters, has full rank, then all the parameters are locally identifiable, and the model as a whole is said to be locally identifiable. Equivalently, the (expected) matrix of the second order partial derivatives of the log-likelihood possesses eigenvalues smaller than 0 (see Formann, 1985). Goodman proved that the three latent class model with four observed dichotomous variables is not even locally identified.

Maximum likelihood values of the parameters can be found using standard optimization routines such as Newton-Raphson technique (McHugh 1956, 1958) and the EM-algorithm (Dempster *et al.* 1977, Goodman 1978, Wu 1983) . The log-likelihood function for a random sample of size n becomes

$$L = \sum_{k=1}^{n} \ln\left[\sum_{r=0}^{R-1} \eta_r \prod_{i=1}^{p} \pi_{ir}^{x_i}(1-\pi_{ir})^{1-x_i}\right];$$

which as to be maximized subject to $\sum \eta_r = 1$.

The EM algorithm proceeds as follows:

(*i*) it starts with a random choice of an initial set of posterior probabilities $h(r|x)$.

(*ii*) than it uses $\hat{\eta}_r = \sum_{k=1}^{n} h(r|x_k)/n$ for $\{r = 0, ...R-1\}$ and $\hat{\pi}_{ir} = \sum_{k=1}^{n} h(r|x_h)/n\hat{\eta}_r$ for $\{i = 1, 2, ..., p\}$, where $h(r|x_k)$ is the posterior probability to be in class r for an individual with x_k, to obtain a first approximation to $\{\hat{\eta}_r\}$ and $\{\hat{\pi}_{ir}\}$,

(*iii*) the estimates obtained are substituted into the expression 2.1.2 to obtain improved estimates of $h(r|x_h)$,

(*iv*) then the routine returns to compute a second approximation for the parameters $\{\hat{\eta}_r\}$ and $\{\hat{\pi}_{ir}\}$; and it continues the cycle until convergence is attained.

It is useful, as mentioned before, to use different starting values to guard against the risk of taking a local for a global maximum. In multinomial mixture global maxima of the likelihood may exist and this is not only the case for maximum likelihood estimation but also for other methods. If a local maximum has been achieved the problem is to establish it as the unique or global one (see e.g. McCutcheton, 2002). This can be done using different sets of starting values; if they result in different local maxima, then the likelihood is multimodal and one can choose the best maximum as the global one. It seems that, as the number of classes increases, the likelihood threatens to be multimodal (e.g Kollmann and Formann, 1997) and the risk of multiple maxima decreases with increasing sample size. Aitkin *et al.* (1981) provided an illustration of multiple maxima arising with only three latent classes.

The algorithm does not provide the second derivatives needed for the calculation of standard errors of the latent class proportions or conditional probabilities, which can be used to construct confidence intervals that give some sense of the stability of the parameter estimates and to construct various tests of significance. A direct way of obtain standard errors from the EM algorithm has been proposed by Lang (1992) when the latent class model is parameterized in terms of a log-linear model. The estimated standard errors are generally evaluated asymptotically. The second derivatives and cross derivatives of L can be expressed in terms of the posterior distributions and the asymptotic variance-covariance matrix of the estimates is than the inverse of the expectation of the matrix of the negatives of such derivatives. Thus if we have a set of parameters θ then

$$\mathrm{cov}(\hat{\theta})^{-1} = E\left[-\frac{\partial^2 L}{\partial\theta_i \partial\theta_j} \right]_{\theta=\hat{\theta}}.$$

With large n the expectation can be approximated by taking the inverse of the observed second derivative matrix. For the unrestricted latent class analysis, applied to binary data, it has been shown (De Menezes, 1999) that the asymptotic standard errors may poorly approximate the true values when sample sizes are not sufficiently large or there are problems of sparse data. It has been suggested also that one should compare the asymptotic standard errors with standard error estimates obtained using empirical methods such as the *jackknife* and the *parametric bootstrap* (Efron and Gong, 1983). The jackknife is a resampling method that is based on the notion of omitting one observation at a time and then recomputing the statistic of interest (Dayton, 1998). Assume that for a sample of N observations, the computed value of the statistics of interest based on the total sample is W and that if the ith case is omitted, the compute value of the statistic based on the remaining $N-1$ cases is W_i, where $i = 1, ..., N$ now refer to a case and not to a to a manifest variable. Than the jackknife estimates for

the sampling variance of W is $N\sum_{i=1}^{N}(W_i - W)^2/(N-1)$ and the square root of this quantity is the estimated standard error for W.

The parametric bootstrap involves conducting a simulation study in which the sample values of the latent class proportions and conditional probabilities are treated as if they were population values. Random samples of size N are generating using these population values, the statistic of interest is computed from the samples, and the standard deviation of the empirical distribution is used to estimate the standard error of the statistic. For this reason, many common latent class software programs, such as LATENT GOLD (Vermuth and Magidson, 2000), have a facility for estimating standard errors by a parametric bootstrap method.

Several latent class model evaluation procedures have become standard. These techniques rely on a comparison between the expected frequency count given by the estimated latent class model parameters and the cell frequency count found in the sample data. When such frequencies are too far from each other, the model is implausible. Models with many latent classes usually provide expected cell frequencies closer to the observed cell frequencies. The task is to find the most parsimonious model that has an acceptable fit to the observed data. To have such an evaluation four criteria are used, two of which are the Pearson chi-square statistics χ^2 and the log-likelihood ratio test G^2. If n is large compared to 2^p the expected frequency for each response pattern is likely to be large enough to carry out a valid asymptotic chi-squared χ^2 or log-likelihood ratio G^2 test to compare the observed and expected frequencies. The calculation of the degrees of freedom for a model with R classes is $df = 2^p - R(p+1) + 1$.

One problem of the chi-square statistics is that there is some difficulty in rejecting the significance of even quite modest parameters when the sample size is large (see, e.g. McCutcheon 2002). Second latent class models can require the estimation of a rather large number of parameters even for models of modest size.

The other two criteria, are known as information criteria based on concepts derived from information theory, are the AIC (Akaike, 1973) and the BIC (Schwarz 1978, Rafftery, 1995). They penalized the likelihood for the increased number of parameters required to estimate more complex models. There are two equivalent ways to implement the AIC and the BIC for a set of competing latent class models. As originally proposed by Akaike, the AIC is based on the log-likelihood for the data based on, say, h different models begin compared

$$AIC = -2\log L + 2npar,$$

where *npar* is the number of independent parameters that are estimated when fitting

the model to the data. The BIC based on the log-likelihood is the following

$$BIC = -2\log L + (\log N)npar,$$

where N is the sample size. A simple approach is based on likelihood-ratio chi square statistics: the AIC is given by $AIC = G^2 - 2df$ and the BIC is $BIC = G^2 - df[\log(N)]$ where df is the degrees of freedom. Consequently the AIC penalizes the G^2 by the total number of parameters required for the model estimation and the BIC by the total number of parameters required to fit the model and by the total sample size. Consequently, models with lower AICs and BICs are judged to be more appropriate than those with higher values of these criteria. A criticism of the AIC is that it lacks property of asymptotic consistency because it does not directly involves the sample size N. Results on empirical investigations of the AIC and BIC (Lin and Dayton, 1997) suggest that, in the context of latent class models, researchers might prefer the AIC unless the sample contains several thousand cases or the models begin estimated are based on relatively few parameter, in which the BIC is preferable. Often the two measures select the same or very similar models.

Other statistics are known as classification statistics (Vermunt and Magidson, 2000), they contain information on how well we can predict to which latent class cases belong given they observed values, in other words how well latent classes are separated. The estimated proportion of classification error for each latent class may be computed as follows

$$E_r = \sum_{r=1}^{R} \frac{n_r}{N}\big[1 - \max(\pi(r|x_v))\big];$$

where n_r are the frequencies of the response vector x_v, N is the total number of individuals on the sample and $\pi(r|x_v)$ is the posterior probability for the model latent class for response vector x_v. It should be noted that the success of the misclassification for an actual data set tends to be optimistic, because both parameter estimation and classification are based on the same data. Other types of classification errors may be computed (see e.g. Vermunt and Magidson, 2000 p.25).

The Average Weight of Evidence (AWE) criterion adds a third dimension to the information criteria described above. It weights fit, parsimony, and the performance of the classification (Banfield and Raftery, 1993). This measure uses the so-called classification log-likelihood. Which is equivalent to the complete data log-likelihood L^c and it can be defined as

$$AWE = -2L^c + 2\Big(\frac{3}{2} + \log N\Big)npar.$$

The lower AWE, the better a model. Celeux (1997) described various indices that combine information on model fit and information on classification errors.

Some problems for model selection arise for "sparse" data, see for example Collins *et al.*, 1993 and, in a more general context Langeheine *et al.*, 1996. In such cases the parametric bootstrap (Efron, 1982) has become a standard method for model selection in latent class analysis which is used to determine the empirical distribution of the χ^2 test statistic using Monte Carlo sampling (Noreen, 1989). This method requires the fitting of the desired model in the usual way and the generating of a random sample from the population in which the parameter values are set equal to those estimated for the actual sample and then the fitting of the model in each case and the computing of the chosen test of goodness of fit. Then it needs to compare the actual value of the statistic with the bootstrap sampling distribution. The number of bootstrap samples needs to be large enough to give a reasonable estimate of the sampling distribution but the time needed to do the calculations will increase proportionally.

2.2 Dynamic Latent Class Model

Static latent variables as those in the classic latent class models are not expected to change over time, or else the change is not of interest in a particular study. In contrast, dynamic latent variables do change in systematic and important ways over time. The same variable may be thought as static in one context and dynamic in another, depending of the objectives and interests of a particular research. Empirical study of developmental processes, for example, requires the availability of pertinent statistical methods that capture the dynamics underlying age-related changes. Changes in criminal behaviour, in drug use, age-related decline in cognitive ability are examples of a dynamic latent variables.

In longitudinal or panel data, which aries when a sample of units provide responses on multiple occasions, an important feature is that observations at different occasions are clustered in units and influenced by the same shared unit-specific unobserved heterogeneity as well as an unobserved time-varying influence that induces greater dependence between responses occurring closer together in time.

An approach is obtained by means of latent Markov model proposed originally by Wiggins (1955) and then referred as Markov chain models by Lazersfeld and Henry (1968, Chap.9) in the analysis of panel data where few repeated observations (typically three to five) are made on the same people. They have been employed to evaluate theories of development (Langeheine 1994, Langeheine and van de Pol 2002).

An alternative formulation, almost identical, to the latent Markov model has been proposed recently under name the *latent transition analysis* (LTA; Hansen 1991, Collins and Wugalter, 1992). It is a latent class theory approach to measuring stage sequential

dynamic latent variables. As pointed out by Collins and Flaherty (2002), this model is limited to small number of variables and to few stages.

In the time series analysis the application such models are referred as hidden Markov model (MacDonald and Zucchini, 1997). It is increasingly begin adopted in applications since it provides a convenient way of formulating an extension of a mixture model to allow for dependent data.

The term *Markov chain* refers to the discrete variables measured repeatedly over time with the same sample of subjects and that the dynamics across time are modelled by assuming a discrete time process in order to make statements about change or stability or both.

In the following the notation and terminology of the Markov chain is introduced which is mainly borrowed from Bilmes (2002). First we describe a discrete time Markov chain model with observed variables; we do so in some detail because of the close links to the hidden Markov Model. We do not discuss the statistical questions relating the model selection, the properties of maximum likelihood estimators and the question of testing for goodness of fit which are difficult problems in such context and need further study.

A discrete time stochastic process is a collection X_t for $t = 1, ..., T$ of random variables ordered by the discrete time index t.

Definition 2.1. *Stationary Stochastic Process*
The stochastic process $\{X_t : t \geq 1\}$ *is said to be stationary if the two collection of random variables*

$$\{X_{t1}, X_{t2}, ..., X_{tn}\}$$

and

$$\{X_{t1+h}, X_{t2+h}, ..., X_{tn+h}\}$$

have the same joint probability distributions for all n *and* h.

Stationarity is equivalent to the condition that

$$\begin{aligned} &P(X_{t1} = x_1, X_{t2} = x_2, ..., X_{tn} = x_n) \\ &= P(X_{t1+h} = x_1, X_{t2+h} = x_2, ..., X_{tn+h} = x_n), \end{aligned}$$

for all $t_1, t_2, ..., t_n$ for all $n > 0$ for all $h > 0$, for all x_i.
A collection of discrete-valued random variables $\{Q_t\}$ forms an n^{th}-order Markov chain if

$$\begin{aligned} &P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, ..., Q_1 = q_1) \\ &= P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, ..., Q_{t-n} = q_{t-n}) \end{aligned}$$

for all $t \geq 1$, and all $q_1, q_2, ..., q_t$. In other words, given the previous n random variables, the current variables is conditionally independent of every variables earlier than the previous n. A first order Markov chain can be seen as a directed acyclic graphical Markov model as described in Section 1.1.2.

One often views the event $\{Q_t = i\}$ as if the chain is "in state i at time t" and the event $\{Q_t = i\, Q_{t+1} = j\}$ as a transition form state i to state j starting at time t.

A *Markov chain* or *process* is a sequence of events, usually called *states*. It is a stochastic process, in which it is assumed that future states are not influenced by all the past states: given the present the future is independent of the past.

The statistical evolution of a Markov chain is determined by the state transition probabilities $f_{ij}(t) = P(Q_t = j|Q_{t-1} = i)$ of an occurrence at time point t given the occurrence at time point $t-1$. In some cases it can be assumed that the transition probabilities are time invariant. Such a time independent chain is called time-homogeneous because $f_{ij}(t) = f_{ij}$ for all t. The transition probabilities in an homogeneous Markov chain are determined by a time independent stochastic transition matrix $A = \{a_{i,j}\}$. The rows of A form potentially different probability mass functions over the states of the chain; the special case of time $t = 1$ is described by the initial state distribution $\pi = p(Q_1 = i)$. For this reason A is also called a stochastic transition matrix. The stationarity condition of a Markov chain depends on if the Markov chain transition matrix admits a stationary distribution or not, and if it does the current marginal distribution over the states is one of those stationary distributions.

If Q_t is a time-homogeneous stationary Markov chain then

$$P(Q_{t_1} = q_1, Q_{t_2} = q_2, ..., Q_{t_n} = q_n) = P(Q_{t_{1+h}} = q_1, Q_{t_{2+h}} = q_2, ..., Q_{t_{n+h}} = q_n)$$

for all t_i, h, n and q_i.

Using the first order Markov chain, the above can be written as

$$P(Q_{tn} = q_n|Q_{tn-1} = q_{n-1})P(Q_{tn-1} = q_{n-1}|Q_{tn-2} = q_{n-2})... \\ P(Q_{t2} = q_2|Q_{t1} = q_1)P(Q_{t1} = q_1).$$

As it can be seen the time-homogeneous property of a Markov chain is distinct from the stationarity property. Stationarity, however does implies time-homogeneity. An homogeneous Markov chain is stationary only when $P(Q_{t_1} = q) = P(Q_{t_1+h} = q) = P(Q_t = q)$ for all $q \in Q$. Futher details can be found in Guttorp (1995, Chapter 2) and Kao (1997). The sequence of states $Q = \{q_1, ..., q_k\}$ can be observed only through the stochastic processes defined into each state.

Following Bilmes (2002) we provide a formal definition of the Hidden Markov Models

Definition 2.2. *A hidden Markov model is a collection of random variables consisting of a set of T discrete scalar variables $Q_{1:T}$ and a set of other variables $X_{1:T}$ which may be either discrete or continuous (and either scalar or vector valued). These variables, collectively, possess the following conditional independence properties*

$$\{Q_{1:T}, X_{1:T}\} \perp\!\!\!\perp \{Q_{1:t-2}, X_{1:t-1}\}|Q_{t-1} \tag{2.2.1}$$

and

$$X_t \perp\!\!\!\perp \{Q_{\neg t}, X_{\neg t}\}|Q_t \tag{2.2.2}$$

where $X_{\neg t} = (X_{1:T}|X_t) = \{X_1, X_2, ..., X_{t-1}, X_{t+1}, X_{t+2}, ..., X_T\}$ *for each* $t \in 1:T$.

Equations (2.2.1) and (2.2.2) imply a large assortment of conditional independence statements. Equation (2.2.1) states that the future is conditionally independent of the past given the present. One implication is that $Q_t \perp\!\!\!\perp Q_{1:t-2}|Q_{t-1}$ which means that the variables form a discrete time, discrete-valued, first-order Markov chain. Another implication of Equation (2.2.2) is $Q_t \perp\!\!\!\perp \{Q_{1:t-2}, X_{1:t-1}\}|Q_{t-1}$ which means that X_τ is unable, given Q_{t-1}, to affect Q_t for $\tau < t$. This does not imply, given Q_{t-1}, that Q_t is unaffected by future variables. In fact the distribution of Q_t could change, even given Q_{t-1}, when the variables X_τ or $Q_{\tau+1}$ change, for $\tau > t$.

The Equation (2.2.2) states that given an assignment to Q_t, the distribution of X_t is independent of every other variable, both in the future and in the past in the hidden Markov model. One implication is that $X_t \perp\!\!\!\perp X_{t-1}|\{Q_t, Q_{t-1}\}$ which follows since $X_t \perp\!\!\!\perp \{X_{t+1}, Q_{t+1}\}|Q_t$ and $X_t \perp\!\!\!\perp X_{t+1}|Q_{t+1}$.

The two above conditional independence properties imply that, for a given T, the join distribution over all the variables may be expanded as follows

$$\begin{aligned} p(x,q) &= p(x_T, q_T|x_{1:T-1}, q_{1:T-1})p(x_{1:T-1}, q_{1:T-1}) \\ &= p(x_T|q_T, x_{1:T}, q_{1:T-1})p(q_T|x_{1:T-1}, q_{1:T-1})p(x_{1:T-1}, q_{1:T-1}) \\ &= p(x_T|q_T)p(q|q_{T-1})p(x_{1:T-1}, q_{1:T-1}) \\ &= ... \\ &= p(q_1)\prod_{t=2}^{T} p(q_t|q_{t-1})\prod_{t=1}^{T} p(x_t|q_t). \end{aligned}$$

The first and the second row of the equation follow from the chain rule of probability. The third row is obtained since $X_T \perp\!\!\!\perp \{X_{1:T-1}, Q_{1:T-1}|Q_T\}$ and $Q_T \perp\!\!\!\perp \{X_{1:T-1}, Q_{1:T-2}\}$ which follow from Definition 2.2.

A parametrization of an hidden Markov model requires: the distribution over the initial chain variable $p(q1)$, the conditional transition distributions for the first-order

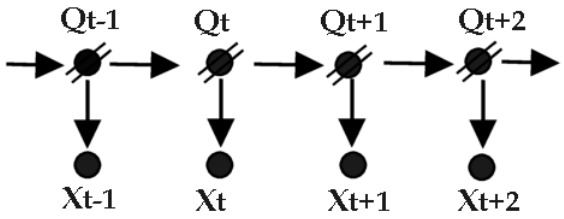

Figure 2.2: *A Hidden Markov Model*

Markov chain $p(q_t|q_{t-1})$ and the conditional distribution for the other variables $p(x_t|q_t)$. In the classic definition of hidden Markov model the initial distribution is labelled π which is a vector of length of the cardinality of the state space; $p(Q_1 = i) = \pi_i$ where π_i is the i^{th} element of π. The observation probability distributions are notated as $b_j(x) = p(X_t|Q_t = j)$ and the associated parameters depend on $b_j(x)$'s family of distributions, when the observations are discrete the distributions are mass functions. Also the Markov chain is typically assumed to be time-homogeneous, with stochastic matrix A where $A_{ij} = p(Q_t = j|Q_{t-1} = i)$ for all t. In the Markov chain $Q_{1:T}$ is typically hidden, which naturally results in the name *hidden* Markov model. The variables $X_{1:T}$ are typically observed.

The collection of observed values $x_{1:t}$ have presumably been produced according to some specific but unknown assignment to the hidden variables. To compute the probability $p(x_{1:T})$ one must therefore marginalize away all possible assignments to $Q_{1:T}$ as follows

$$p(x_{1:T}) = \sum_{q_1:T} p(x_{1:T}, q_{1:T}) = \sum_{q_1:T} p(q_1) \prod_{t=2}^{T} p(q_t|q_{t-1}) \prod_{t=1}^{T} p(x_t|q_t).$$

An hidden Markov model may be graphically depicted as one instance of a DAG model as displayed in Figure 2.2. Using any of the equivalent schemas such as the directed local Markov properties, the conditional independence properties implied by Figure 2.2 are identical to those expresses in Definition 2.2. For example the variable X_t does not depend of any of X_t's non descendants $\{Q_{\neg}t, X_{\neg}t\}$ given X_t's parents Q_t.

2.2.1 Estimation

For the maximum likelihood estimation of the model the EM-type algorithm is usually applied (Dempster *et al.*, 1977). One is interested in computing $p(x_{1:T})$ for a given set of independent and identically distributed observations according to a multinomial

distribution. The joint distribution can be expressed as

$$p(x_{1:T}) = \sum_{q_{t-1}} p(x_t|q_t)p(q_t|q_{t-1})p(x_{1:t}, q_{t-1}).$$

The EM algorithm is commonly used for finding the maximum-likelihood estimate of the parameters of a HMM given the set of observed vectors. This algorithm is also known in the HMM literature as a re-estimation procedure (Baum and Petrie, 1966; Baum *et al.* 1970). The likelihood is constructed from tree ingredients: the initial distribution at the first state of the chain; the time dependent transition probabilities and the conditional densities $\phi(x_t|q_t)$.

The inference calculation involves calculating the probabilities of the hidden states Q_t given the time series of X_t. Baums's forward algorithm is based on recursively evaluating the joint probabilities using the alpha and beta-recursions.

The alpha or forward recursion is based on recursively evaluating the joint probabilities

$$\alpha_q(t) = p(x_{1:T}, Q_t = q)$$

which is the probability of seeing the partial sequence $x_{1:T}$ and ending up in state q at time t. To derive this recursion it was necessary to use only the fact that X_t was independent of its past given Q_t.

The backward procedure (or beta recursion) recursively evaluates the conditional probabilities

$$\beta_q(t) = p(x_{t+1:T}|Q_t = q)$$

which is the probability of ending partial sequence $x_{t+1:T}$ given that we started at state $Q_t = q$ at time t. The E step requires the calculations of the conditional expectation complete-data log-likelihood, where the posterior probabilities can be expressed in terms of the alpha and beta recursions (see e.g. MacLachlan and Peel, 2000). The M-step consists in finding the updated estimates of the parameters from the function at the E step. They are a combination of the maximum likelihood estimates for the multinomial parameters and Markov chain transition probabilities.

2.3 Bibliographical Note

The use of latent class models as a tool to help researchers gain a deeper understanding of the observed relationships among the observed dichotomous (or polytomous) variables has the same short history as the path analysis models in the twentieth century, but it might be worthwhile to note that some mathematical models that were used

earlier in some nineteenth-century models as the work of Peirce (1884) can be viewed as special case of latent class models. The main development of latent class models has taken place during the last half of the twentieth-century: it was developed by Lazarsfeld (1950) and by Lazarsfeld and Henry (1968). Goodman (1974a,1974b) and Haberman (1974, 1976, 1979) were the first to formulate maximum likelihood estimation procedures that could be used in the field of latent class analysis. During the 1950s and 1960s, there were essentially five different methods that were proposed for estimating the parameters in the latent class model (for a review see Goodman, 2002). During the past two decades great progress has been made regarding the estimation and testing of latent class models.

Heberman (1974) showed the connections between latent class models and log-linear models for frequency table with missing cell counts. Bartholomew (1987) classifies both the latent and the manifest variables as metrical or categorical and he uses the phrase: *latent structure analysis* for all models that use categorical latent variables, regardless of the nature of the manifest variables. Test theory models, such as psychological measurement models, known as item response theory (IRT), become popular also under the heading *latent trait models* where it is assumed a continuous latent variable (e.g. ability) and functional relationships are used to model the dependence of discrete random variables (e.g. responses to achievement test items). Latent class and latent trait models treat the manifest indicators as discrete variables, although they can be measured at every measurement level, whereas latent class treat the latent variables as discrete. For a detailed discussion see Heinen (1996).

As previously noted the latent class model assumes that the population consists of T mutually exclusive and exhaustive homogeneous subgroups or latent classes. Each individual belongs to only one latent class. Who belongs to the same latent class has equal probabilities for responding to the item in certain categories. In this sense, latent class model is a *finite-mixture* model (Titterington *et al.*, 1985; McLachlan and Basford, 1988) because the total population is a mixture of finite number of latent classes that differ not only with respect to the conditional response probabilities but also with respect to their relative sizes. During the past decade it has also become clear that particular developments in econometrics, biometrics, and mathematical statistics concerning finite mixture models are identical or at least have very close ties to latent class modelling, thus it enhances the potentialities of latent class analysis.

When the categorical latent variables are regarded as nominal level variables, whose categories are not ordered, there is a close connection between the concepts of "clusters" and "latent classes". Latent class cluster analysis is essentially a variant of what Gibson and Lazarsfeld in the 1950s called *latent profile analysis*, in which the underly-

ing variable is treated as nominal level latent variable, but the observed variables (the indicators) are treated as continuous. Wolfe (1970) was the first who made an explicit connection between latent class and cluster analysis. An important difference between standard cluster analysis techniques and latent class clustering is that the latter is a model-based clustering approach (see e.g. Vermuth and Magidson, 2000). In such cases the goal is to identify distinct different pattern and classifying respondent into groups. This means that a statistical model is postulated for the population from which the sample under study is coming. Thus it is assumed that the data is generated by a mixture of underlying probability distributions (see e.g MacLachlan and Basford, 1988). One advantage of model based clustering is that it provides a precise framework for assessing the resulting partitions of the data and especially for choosing the relevant number of clusters (see e.g. Biernacki *et al.* 1998).

As Clogg (1988) noted, many applications in latent class analysis also aim at more explanatory data reduction. In many examples of social research, in fact, so many different variables are measured that it becomes necessary to compress these data into a smaller set of variables (e.g.; see Aitkin *et al.* 1981, Formann, 1985).

A proposed model called latent transition analysis (LTA; Graham *et al.*, Hansen, 1991; Collins and Wugalter, 1992) consists of appropriate dependence sequences of discrete random variables. It is a latent class theory approach to measuring stage sequential dynamic latent variable and estimating and testing models for stage sequential development.

Alternative formulation of the latent class model for longitudinal categorical data can be found in e.g. Langeheine and van de Pol (2002).

For general hidden Markov models, Lindgren (1978) constructed consistent and asymptotically normally estimator of the component distributions, but he did not consider estimation of the transition probabilities. Leroux establish the consistency of the maximum likelihood estimation for general hidden Markov models under mild conditions, while local asymptotic normality was proven by Bickel ad Ritov (1996). Recently, Bickel *et al.* (1998) showed that under mild conditions the MLE is asymptotically normal and that the observed information matrix is a consistent estimator of the expected information. The relation between hidden Markov models and graphical models has been reviewed also by Smyth *et al.* (1997).

The hidden Markov model is finding widespread in engineering applications, for example as the acoustic model in speech processing see Levinson, Rabiner and Sondhi (1983), in econometrics (Hamilton, 1989; Chib, 1996), biology (Albert, 1991), finance (Ryden *et al.* 1998).

2.4 References

Aitkin M., Anderson D., Hinde J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society*, A, 144, 419-448.

Akaike H. (1973). Information theory and an extention of the maximum likelihood principle. In B. N. Petrov and F. Csàki, (eds.), 2nd International Symposium on Information Theory, Budapest: Akademiai Kiádo, pp. 267-281. Reprinted in Kotz, S. Johnson N. L. (eds.) *Breakthroughs in Statistics, Volume I: Foundations and Basic Theory.* New York: Springer-Verlag.

Banfield J. D., Raftery A. E. (1993). Model-based Gaussian and non-Gaussian Clustering. *Bionetrics*, 49, 803-821.

Bartholomew D. J. (1987). *Latent variables models and factor analysis.* London: Charles Griffin.

Bartholomew D. J., Knott M. (1999). *Latent Variable Models and Factor Analysis.* London: Oxford University Press.

Baum L.E., Petrie T. (1966). Statistical inference for probabilistic functions finite state Markov chains. *Annals of Mathematical Statistics*, 37, 1554-1563.

Baum L. E., Petrie T., Saules G., Weiss N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41 (1), 164-171.

Biernacki C, Celeux G., Govaert G. (1998). Assessing a mixture model for clustering with the intergrated classification likelihood. *Technical Report No. 3521.* Rhône-Alpes INRIA.

Blimes J. (2002). What HMMs Can Do. *Technical Report*, Department of Electrical Engineering, University of Washington.

Cleux B, Biernacki C., Govaert G. (1997). *Choosing Models in model based clustering and discriminant analysis.* Technical Report. Rhone-Alpes: INRIA.

Clogg C. C. (1981). Latent structure models of mobility. *American Journal of sociology*, 86, 836-868.

Clogg C. C. (1988). Latent class models for measuring. In *Latent Trait and Latent Class Models.* R. Langeheine and J. Rost (eds), 173-206. New-York: Plenum Press.

Collins M. L., Wugalter S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioural Research*, 27, 131-157.

Collins M. L., Flaherty B. P. (2002). Latent class models for longitudinal data. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 287-303. Cambridge University Press.

Cox D. R. (2003). Conditional and marginal association for binary random variables. *Biometrika* 90, 4, 982-984.

Dayton C. M. (1999). *Latent class scaling analysis*. Sage Publications.

de Menezes L. M. (1999). On fitting latent class models for binary data. *British Journal of Mathematical and Statistical Psychology* 52.

Dempster A. P. Laird N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statististical Society* B, 39, 1-38.

Efron B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron B., Gong G. (1983). A leisure look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37, 36-48.

Formann A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology* 38, 87-111.

Formann A. K. (2003). Latent Class Diagnosis from a Frequentist Point of view. *Biometrics*, 59, 189-196.

Gibson W. A. (1955). An extension of Anderson's solution for the latent structure equations. *Psychometrika* 20, 69-73.

Goodmann L. A. (1974a). Explanatory Latent Structure Models Using both identifiable and unidentifiable Models. *Biometrika* 61, 315-331.

Goodman L. A. (1974b). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: a modified latent structure approach. *American journal of sociology* 79, 1179-1259.

Goodman L. A. (1978). *Analyzing qualitative/Categorical Data* (ed. J. Magidson), Abt Books, Cambridge, MA.

Goodman L. A. (2002). Latent Class Analysis. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 3-55. Cambridge University Press.

Guttorp P. (1995). *Stochastic Modeling of Scientific Data* . Chapman and Hall, 384 pp.

Haberman S. J. (1974). Loglinear models for frequency tables derived by undirected observation. Maximum-likelihood equations. *Annals of Statistics* 2, 911-924.

Haberman S. J. (1976). Iterative scaling procedures for loglinear models for frequency data derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association, 1975*, 45-50.

Haberman S. J. (1979). *Analysis of Qualitative Data*. Vol.2. New York: Academic Press.

Hagenaars J. A. (1990). *Categorical Longitudinal Data; Loglinear Panel, Trend and Cohort Analysis*. Newbury Park, CA: Sage.

Hansen W. B., Graham J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: peer pressure resistance training versus establishing conservative norms. *Preventive Medicine*, 20, 414-430.

Heinen T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Advanced quantitative Techniques in the Social Sciences, Sage Publications, Thousand Oaks, CA.

Kao E. P. C. (1997). *An introduction to stochastic process*. Duxbury Press, 438 pp.

Kollmann T., Formann A. K. (1997). Using latent class models to analyse response patters in epidemiologic mail survey. On *Applications of Latent Trait and Latent Class Models in the Social Sciences* J. Rost, R. Langeheine (eds.), 345-351. Münster, Germany: Waxmann.

Lang J. B. (1992). Obtaining the observed information matrix for the poisson loglinear model when the EM-algorithm is used. *Biometrika*, 79, 405-407.

Langeheine R. (1994). Latent variable Markov models. In A. von Eye and C.C. Clogg (eds.) *Latent Variable Analysis. Application for Developmental Research*. Thousand Oaks, CA: Sage, pp 373-395.

Langeheine R., Pannekoek J., van de Pol F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociologiacal Methods and Research* 24(4), 492-516.

Langeheine R., van de Pol F. (2002). Latent Markov Chains. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 304-341. Cambridge University Press.

Lazarsfeld P. F. (1950). The logical and mathematical foundation of latent structure analysis. In *Studies in Social Psychology in World War II* Volume IV: *Measurement and Prediction*, S. A. Stouffer, L. Guttman, E. A. Suchman (eds) 362-412. New York: Princeton University Press.

Lazarsfeld P. F., Henry N. W (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lin T. H., and Dayton C. M. (1997). Model selection information criteria for non-nested latent class models. *Journalof Educational and Behavioral Statistics*, 22, 249-264.

Lindgren G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics: Theory and Applications*, 5, 81-91.

MacDonald I., Zucchini W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.

McCutheon A. L. (2002). Basic concepts and procedures in single and multiple group latent class analysis. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 56-85. Cambridge University Press.

McHugh R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* 21, 331-347.

McHugh R. B. (1958). Note on "Efficient estimation and local identification in latent class analysis". *Psychometrika* 23, 273-274.

McLachlan G. J., Basford K. E. (1988). *Mixture Models: Inference and Application to Clustering* . New York: Marcel Dekker.

McLachlan G. J., Peel D. (2000). *Finite Mixture Models*. Wiley Series: New York.

Noreen E. W. (1989). *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York: Wiley.

Peirce C. S. (1884). The numerical measure of the success of predictions. *Science* 4, 453-454.

Raftery A. E. (1995). Bayesian Model selection in social research. In P. V. Marsden (ed.), *Sociological Methodology, 1995*. Cambridge, MA: Blackwell.

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

Titterington D. M, Smith A. F. M., Markov U. E. (1985). *Statistical analysis of Finite Mixture Distributions*. New York: John Wiley.

Vermunt J. K., Magidson J. (2000). *Latent Gold User's Guide*. Belmont: Statistical Innovations Inc.

Wiggins L. M (1973). *Panel Analysis: Latent Probability Models for Attitudes and Behavior Processes*. Amsterdam: Elsevier.

Wolfe J. H. (1970). Pattern clustering by multivariate cluster analysis. *Multivariate Behavioural research*, 5, 329-350.

Wu C. F. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11, 95-103.

Chapter 3

Fitting DAG models with one hidden variable

Abstract: We discuss directed acyclic graph (DAG) models to represent the independence structure of linear Gaussian systems with continuous variables: such models can be interpreted as a set of recursive univariate regressions. Then we consider Gaussian models in which one of the variables is not observed and we show how the incomplete log-likelihood of the observed data can be maximized using the EM. As the EM algorithm does not provide the matrix of the second derivatives we show how to get an explicit formula for the observed information matrix. We illustrate the utility of the models with some examples.

Keywords: Conditional independence graph models; linear systems, directed acyclic graph models, latent variable; identifiability; EM algorithm; standard errors.

3.1 Introduction

The analysis of multivariate data typically deals with complex association structures due to various direct and indirect relations among variables. The idea of graphical Markov models introduced first by Dempster (1972) is to represent the independence structure of a multivariate random vector by a graph where the vertices correspond to variables and the absence of an edge between vertices stands for conditional or marginal independencies. In many applications some dependency structure between

[1]Portions of this work were presented in November 2003 at the gR workshop in Aalborg (DK) by G. Marchetti, F. Pennoni and E. Stanghellini; at the Methodology and Statistics Conference in Ljubljana by F. Pennoni. This research was partially supported by MURST (grant PRIN 2002).

observed variables can be explained by supposing that their distribution arises after marginalizing over, and or conditioning on latent variables.

Such models are particularly of interest in the context in which one variable is not observed and some knowledge about the generating process of the data is available as for example for data collected in the social sciences. In such context appropriate estimation procedures have to be found to estimate the parameters of interest. We focus on maximum likelihood estimation of DAG models with one latent variable which can act as an intermediate, source or collision node. The estimation requires iterative solutions and thus appropriate algorithms.

The outline of the chapter is as follows. In the first section we interpret a DAG for a Gaussian system as a set of recursive univariate regressions and we give some matrix notation. In Section 3.3 we show the observed data log-likelihood and briefly we discuss some identifiability problems. We also illustrate the steps of the EM algorithm for maximum likelihood estimation. Following Kiiveri (1987) we report the explicit form for the second derivatives of observed log-likelihood and in the Appendix A we show how to derive it. In Section 3.4 we give some examples of identifiable DAG's with one hidden variable using real data sets. Computations are carried out in `R` with the package `ggm` (Marchetti and Drton, 2003). In the Appendix B the `R` routines are reported for adding to the `ggm` library the estimated standard errors for the parameter of DAG model with one latent variable.

3.2 Gaussian directed acyclic graph models

Suppose $X = (X_1, X_2, ..., X_k)$ is a finite set of substantive variables of interest ordered in certain way, such that there exist a subset of indices $pa(i) \subseteq \{i+1, ..., k\}$, $i = 1, ..., k$, some independent random variables $\epsilon_1, \epsilon_2, ..., \epsilon_k$ and linear functions $f_1, f_2, ..., f_k$ such that

$$X_i = f_i(X_{pa(i)}, \epsilon_i), \qquad i = 1, ..., k \qquad [X_{pa(i)} \equiv \{X_j : j \in pa(i)\}].$$

The set of equations $X_i = f_i(X_{pa(i)}, \epsilon_i)$ prescribes a stepwise process for generating the distribution where a proper dependence of X_i is to be only on its potentially explanatory variables. The system is called *recursive* or a *univariate recursive regression system* or a *triangular system*.

This system can be represented by a directed acyclic graph (DAG) denoted by $G = (V, E)$ which consists of non empty finite set of vertices $V \equiv \{1, ..., k\}$ representing $X = (X_1, X_2, ..., X_k)$ and a set $E \subseteq V \times V$ of arrows $i \leftarrow j \in E$ iff $j \in pa(i)$ such that there are no direct path that start and end at the same variable. The multivariate

distribution of X_v is called $G - Markovian$ if it fulfils the so called pairwise Markov property

$$X_a \perp\!\!\!\perp X_c | X_c \qquad \text{for all} \quad (a,b) \notin E, a \neq b.$$

For Gaussian distribution this is equivalent to the global Markov property as illustrated in Chapter 1 (Lauritzen and Wermuth, 1989). An important property of a distribution satisfying the global directed Markov property associated with a DAG is that its joint density can be decomposed into conditional probabilities involving only variables and their parents according to the structure of the graph in the following way

$$p(x_1, ..., x_k) = \prod_{i=1}^{k} p(x_i | x_{pa(i)}).$$

Assuming that X is a vector of k mean centered random variables with Gaussian joint distribution with covariance matrix Σ, the recursive system can be written as

$$AX = \epsilon \qquad \text{cov}(\epsilon) = \Delta \tag{3.2.1}$$

where $A = \{-a_{rs}\}$ is upper triangular matrix with ones along the diagonals and with off-diagonal elements corresponding to partial regression coefficients between two variables given the parents, $-a_{rs} = \beta_{rs.pa(r)\setminus s}$ associated with a directed edge between $X_s \leftarrow X_r$; $\Delta = \text{cov}(\epsilon)$ is a nonsingular diagonal covariance matrix of the residuals with elements of partial variances $\delta_{rr} = \sigma_{rr.pa(r)}$ along the diagonal, representing the unexplained proportion of the variance of the dependent variable.

The arguments we are dealing with apply also to the very much broader family of problems that are called *quasi linear* (Cox and Wermuth, 1996). It means that any dependence present has a linear component and like linear least square regression equations in a multivariate normal framework, any curvature and higher-order interactions present are such that a vanishing linear least-squares regression coefficient implies that no dependence of substantive importance is present.

A triangular decomposition of the covariance matrix Σ and of the concentration matrix Σ^{-1} is given by

$$\text{cov}(X) = \Sigma = (A^{-1})\Delta(A^{-1})', \qquad \Sigma^{-1} = A'\Delta^{-1}A.$$

Here we consider the estimation of the unknown parameter Σ or equivalently (A, Δ) of a directed graphical Gaussian model based on an n independent and identically distributed observations $X^{(k)} = (X^1, ..., X^k)$ from X, with zero average constructed from the series of deviates from the mean.

Since our model assumes a zero mean, the empirical covariance matrix is definite to be

$$S = \sum_i X^i X^{i'}/n \qquad i = 1, ..., k.$$

We assume $n \geqslant p$ such that S is positive definite with probability one. Note that the case where the model also includes an unknown mean vector μ can be treated by estimating μ by the empirical mean vector $\bar{X}$.

The density function of X can be expressed as

$$f(x) = (2\pi)^{np/2}|\Sigma|^{-n/2}exp\{-\frac{n}{2}\text{tr}(\Sigma^-1)S\},$$

see e.g. Edwards (2000). Considered as a function of the unknown parameters for fixed data x it gives the likelihood function. The log-likelihood of the model, apart from an additive constant

$$l_X(\Sigma) = \frac{n}{2}[\log|\Sigma^{-1}| - \text{tr}(\Sigma^{-1}S)], \tag{3.2.2}$$

has to be maximized respect to Σ.

It can be shown that

$$\hat{a}_{rs} = -\hat{\beta}_{rs.pa(r)\setminus s} \qquad \hat{\delta} = \hat{\sigma}_{rr.pa(r)}$$

are the maximum likelihood estimates of A and Δ defined by linear regression estimates in the independent equations.

3.3 Unobserved variable: maximum likelihood estimation

Supposing that we observe only a subset $Y^p = (Y^1, ..., Y^p)$ of the variables. The complete data can be seen as $X = (Y, Z)$ where Y denotes the observed components of X and Z denotes the unobserved component. When this is the case the corresponding DAG contains an hidden node.

The relevant log-likelihood function based on the observed components can be written as follows

$$l_Y(\Sigma) = \frac{n}{2}[\log|\Sigma_{yy}^{-1}| - \text{tr}(\Sigma_{yy}^{-1}S_{yy})], \tag{3.3.1}$$

where Σ_{yy} denotes the submatrix referring to Y of the conformably partitioned covariance matrix of X, and S_{yy} is the observed covariance matrix.

The problem of what can be learned from the distribution of the observed variables about the joint distribution specified by the DAG involves identifiability conditions. If A and Δ can be uniquely reconstructed from the covariance matrix of the observed variables the system is said to be globally identified. As illustrated in Section 1.6 Stanghellini and Wermuth (2003) give two sufficient conditions based on properties of the graph for identifiability of DAG Gaussian models with one hidden node. If the sample covariance matrix is positive definite and the DAG considered satisfy one of the given conditions the likelihood surface is unimodal and when fitting the corresponding model a unique global maximum can be achieved.

Maximum likelihood analysis can be conceptualized as maximum likelihood estimation in a multivariate normal model with missing data (Dempster *et al.*, 1977). Following Kiiveri (1987) who first suggested the procedure in a discussion on Jöreskog paper (1981), we describe the maximum likelihood method for fitting such DAGs using the EM algorithm (Dempster *et al.*,1977). This is an iterative algorithm and each cycle, which consists of an E step followed by and M step, increases the likelihood of the parameters. The E step calculates the expected sufficient statistics given the observed data and the current estimate of the parameters and the M step determines the conditional expectations of the sufficient statistics as if they were the observed. For an application of the EM algorithm to estimate the factor analysis model see Rubin (1982).

In the following we explicitly define the E and the M step of the algorithm and present a simple matrix expression for carrying out the computations.

The computations required are particularly straightforward: in the E-step we must compute $Q(\Sigma|\Sigma_r)$ the conditional expected value of the complete data log-likelihood to the observed data Y and a guessed initial value of complete data covariance matrix Σ_r

$$Q(\Sigma|\Sigma_r) = E[l(\Sigma, |Y_1, ..., Y_p, \Sigma_r)].$$

It can be shown that

$$Q(\Sigma|\Sigma_r) = \frac{n}{2}[\ln|\Sigma^{-1}| - \mathrm{tr}[(\Sigma^{-1})E(S|Y, \Sigma_r)]] \tag{3.3.2}$$

where the expected complete data covariance matrix given Y can be written as

$$E(S|Y, \Sigma_r) = \begin{pmatrix} S_{yy} & S_{yy}B' \\ . & BS_{yy}B' + (\sigma^{zz})^{-1} \end{pmatrix} = C(S_{yy}|\Sigma_r) \tag{3.3.3}$$

where the element of the concentration matrix Σ^{-1} corresponding to the missing data Z are noted σ^{zz} and where $B = -(\Sigma^{zz})^{-1}\Sigma^{zy} = \Sigma_{zy}\Sigma_{yy}^{-1}$ are the regression coefficients between Z and Y.

Therefore in the M-step we maximize $Q(\Sigma|\Sigma_r)$ as a function of Σ to produce an improved estimate Σ_{r+1}. This maximization is carried out by fitting the linear recursive regression equations specified by the DAG.

The generalized likelihood ratio test for directed graphical Gaussian models against the saturated model, the deviance at convergence is

$$D = n[\mathrm{tr}(S_{yy}\hat{\Sigma}^{-1}) - \ln|S_{yy}\hat{\Sigma}^{-1}| - m],$$

which has an asymptotic χ^2 distribution with $df = [m(m+1)/2 - m - k]$ degrees of freedom, where m is the number observed variables and k is the number of edges in the DAG.

The EM has the advantage of numerical stability which leads to a steady increase in the likelihood of the data. A negative feature is that it may require many iterations to converge, it is characterized by a slow convergence rate in a neighborhood of the optimal point. It is also sensible to the starting values and it is convenient to choose multiple random starting values.

One major shortcoming is that the observed information matrix is not obtained as a by-product of the algorithm, which is useful to get an estimate of the precision of the estimated parameters to construct confidence intervals and to construct various tests of significance.

As illustrated above the EM finds the value of θ, where $\theta = (\theta_1, ..., \theta_h)$ is the vector of the unknown parameters, $\hat{\theta}$ that maximizes $l_Y(\theta)$, that is the maximum likelihood for θ based on the observed data Y. Following Dempster *et al.* (1977) the observed log-likelihood $l_Y(\theta)$ can be decomposed as

$$l_Y(\theta) = Q(\theta|\theta') - H(\theta|\theta')$$

which leads to a simple expression for the second derivative matrix of the observed log-likelihood derived in terms of the criterion function invoked by the EM algorithm. Minus the second derivative of the log-likelihood is made of two parts

$$-\frac{\partial^2 l_Y}{\partial\theta\partial\theta} = -\frac{\partial^2 Q(\theta|\theta')}{\partial\theta\partial\theta} - \Big(-\frac{\partial^2 H(\theta|\theta')}{\partial\theta\partial\theta}\Big)$$

where Q is as in (3.3.2) and H is the expected value of the conditional density of the complete data X given the observed data Y (Tanner, 1996). Referring to $-Q$ as the *complete information* and to $-H$ as the *missing information* it has the following appealing interpretation: the *observed information* is equal to the complete information minus the missing information due to the unobserved components which has been called the "missing information principle" by Orchard and Woodbury, (1972). A basic result

due to Louis (1982) is that if the distribution of the complete data is in a regular exponential family $-\partial^2 H/\partial\theta\partial\theta = Var_{X|Y}(\partial l_X/\partial\theta)$ the second derivative of the log-likelihood of the observed data can be expressed entirely in terms of the complete data log-likelihood

$$-\frac{\partial^2 l_Y}{\partial\theta\partial\theta} = -E_{X|Y}\Big[\frac{\partial^2 l_X}{\partial\theta\partial\theta}\Big] - Var_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}\Big) \tag{3.3.4}$$

the amount of information lost by observing only Y is determined by the conditional variance of the complete data log-likelihood given Y.

It is important to emphasize that the variance-covariance obtained is based on the first and second derivatives of the observed data log-likelihood and thus is guarantee to be inferentially valid only asymptotically.

Kiiveri (1982) calculated an explicit form for the above expression

$$\frac{\partial^2 l_Y}{\partial\theta_i\partial\theta_j} = \frac{1}{2}tr\Big(\Sigma^{ij}(\Sigma - C) - \Sigma^i\Sigma\Sigma^j\Sigma) + \frac{1}{2}tr(\Sigma^i C\Sigma^j C - \Sigma^i\tilde{C}\Sigma^j\tilde{C})\Big) \tag{3.3.5}$$

where $C = C(S_{yy}|\Sigma)$; $\tilde{C} = [C(S_{yy}|\Sigma) - H]$, where $H = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{\sigma^{zz}} \end{pmatrix}$; and $\Sigma^{ij} = \partial\Sigma^{-1}/\partial\theta_i\partial\theta_j$ and $\Sigma^i = \partial\Sigma^{-1}/\partial\theta_i$.

In Appendix A it is shown how to get such a result and also the explicit formulas for the second derivatives of the observed data log-likelihood for the adopted decomposition to Σ^{-1}.

The EM can suffer from extremely slow convergence specially in problems with more than one latent variable. This defect has prompted a number of suggestions for accelerating the algorithm. A Quasi-Newton method could be used which gradually construct an approximate Hessian from the gradient of the objective function evaluating at successive points encountered by the algorithm avoiding in that way to evaluate the Hessian as the Newton algorithm.

3.4 Implementation and examples

We illustrate the fitting of the models described above on same examples. The `R` package `ggm` (Marchetti and Drton, 2003) is designed for fitting graphical Markov models to data from Gaussian distributions. We implemented the `R` code to add to the existing routines of such package to compute the standard errors for the estimated parameters. Such code is reported in Appendix B. The package is intended as a contribution to the `gR`-project described in Lauritzen (2002).

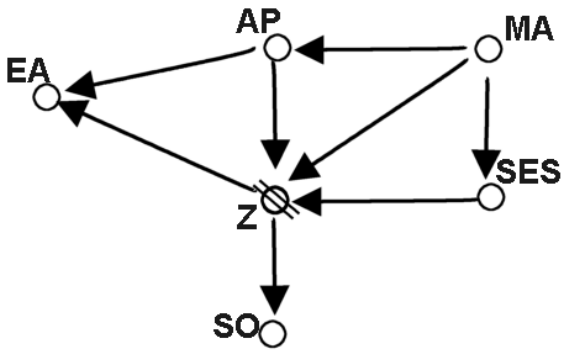

Figure 3.1: *Graphical model for Educational example*

In the following we show an example session in R using ggm and than we reports several examples from the literature.

Example 3.1 (Educational Attainment). We use the the data reported by Sewel, Haller & Ohlendorf (1970) and Wiley (1973). Most measurements used in behavioral and social sciences contains sizeable measurements errors which if not taken into account can cause several bias in results. The following example illustrates the problems with measurement errors in observed variables. The purpose was to describe how well the observed indicators serve as measurement instruments for the latent variable. A sample of 3500 was recorded on the following items, we shorten the variable names:
-MA Mental ability,
-SES Socioeconomic status,
-AP Academic performance,
-SO Significant others'influence,
-EA Educational aspiration.
The postulated model includes a measurement error in SO. We can fit the Graphical gaussian model from Figure 3.1 where the Z variable is not observed. The following is the observed correlation matrix

```
> edu
       ma   ses    ap    so    ea
ma  1.000 0.288 0.589 0.438 0.418
ses 0.288 1.000 0.194 0.359 0.380
ap  0.589 0.194 1.000 0.473 0.459
so  0.438 0.359 0.473 1.000 0.611
```

```
ea  0.418 0.380 0.459 0.611 1.000
```

We define the graphical acyclic directed graph from Figure 3.1, which is done by defining the parents as the regression structure of each node. The resulting graph is represented by an adjacency matrix or equivalently by their *edge matrix* (cf. Cox and Wermuth, 2004) $\mathcal{A} = (a_{ij})$ where $a_{ij} = 1$ if $i \longrightarrow j$.

```
>  dag <- DAG(ap ~ ma, so ~ z, ea ~ ap + z, z ~ ma + ses +ap, ses ~ ma)
>  dag
      ap ma so z ea ses
  ap   0  0  0 1  1   0
  ma   1  0  0 1  0   1
  so   0  0  0 0  0   0
  z    0  0  1 0  1   0
  ea   0  0  0 0  0   0
  ses  0  0  0 1  0   0
```

Routines for testing the state of identification according to the criteria presented in Section 1.6.1 are implemented: the function "checkIdent" which requires to specify the dag and the label of the latent variable, returns which of the conditions are satisfied. For the example considered as the second condition is satisfied the dag is globally identified up to the sign of the regression coefficients of the latent variable.

```
>  checkIdent(dag, "z")
  T1.i T1.ii
 FALSE  TRUE
```

Now we are able to fit the Gaussian DAG model with Z unobserved to the data. The function "fitDagLatent" returns $\hat{\Sigma}$ as `Shat`, the diagonal of $\hat{\Delta}$ as `Dhat`, $\hat{A}$ as `Ahat`. This output also includes the deviance `dev`, the degrees of freedom `df` and the number of iterations `it`. The label of the latent variable needs to be specified and it is possible to set the residual variance or the variance of the latent variable to be 1 specifying `norm=2` or `norm=1` respectively. The starting value for the EM algorithm can be choose with `seed` option.

```
> r<- fitDagLatent(dag, edu,  n,  latent="z", norm = 2, seed=9866)
> r
$Shat
       ma   ses    ap    so    ea     z
ma   1.00  0.29  0.59  0.43  0.43 -0.80
```

```
ses  0.29  1.00  0.17  0.36  0.37 -0.67
ap   0.59  0.17  1.00  0.47  0.45 -0.88
so   0.43  0.36  0.47  1.00  0.61 -1.12
ea   0.43  0.37  0.45  0.61  1.00 -1.14
z   -0.80 -0.67 -0.88 -1.12 -1.14  2.11

$Ahat
      ma  ses   ap so ea    z
ma   1.00 0.00 0.00  0  0 0.00
ses -0.29 1.00 0.00  0  0 0.00
ap  -0.59 0.00 1.00  0  0 0.00
so   0.00 0.00 0.00  1  0 0.53
ea   0.00 0.00 0.03  0  1 0.56
z    0.30 0.48 0.62  0  0 1.00

$Dhat
  ma  ses   ap   so   ea    z
1.00 0.92 0.65 0.40 0.38 1.00

$dev
[1] 7.15

$df
[1] 2

$it
[1] 151
```

Using the following command a summary of the estimated parameters of $\hat{A}$ and of the diagonal of $\hat{\Delta}$ can be obtained with their standard errors, the $z - values$ and the $p - values$ as `tab` and `tab2` respectively.

```
> der<-der2(dag, "z", r, edu, n)
$tab
        value  s.e.       z     p
ma->ses -0.288 0.016 -17.792 0.000
ma->ap  -0.589 0.014 -43.100 0.000
```

```
ma->z    0.303 0.021  14.155 0.000
ses->z   0.479 0.018  27.100 0.000
ap->ea   0.035 0.013   2.660 0.008
ap->z    0.617 0.021  29.468 0.000
z->so    0.532 0.007  72.602 0.000
z->ea    0.556 0.009  61.682 0.000

$tab2
    value  s.e.      z p
ma  1.000 0.027 36.749 0
ses 0.917 0.026 35.889 0
ap  0.653 0.018 37.403 0
so  0.399 0.010 41.091 0
ea  0.378 0.009 41.127 0
z   1.000 0.036 27.486 0
```

Comparing the deviance and the degrees of freedom using the asymptotic distribution of the deviance as χ^2_{df} a satisfying model fit is suggested. It can be seen that the residual variance for SO is 0.399 which means that the reliability of SO is only of the sixty per cent.

Example 3.2 (Criminological research). Let us consider an example from criminological research described by Smith and Patterson (1984). Random samples of persons in sixty residential neighborhoods were interviewed regarding victimization experiences, neighborhood safety and evaluation of police performance. The sample was 1500 people living alone. The seven variables observed were as follows:
- Y_4 number of self reported prior victimizations in the last twelve months,
- Y_5 respondent's age,
- Y_6 respondent sex,
- Y_7 the rate of personal and property victimization per 100 households in the respondent's neighborhood.

The following variables were responses to three questions asking respondents how likely they thought it was that they would be victims of
- Y_3 vandalism during the next year
- Y_2 burglary
- Y_1 robbery.

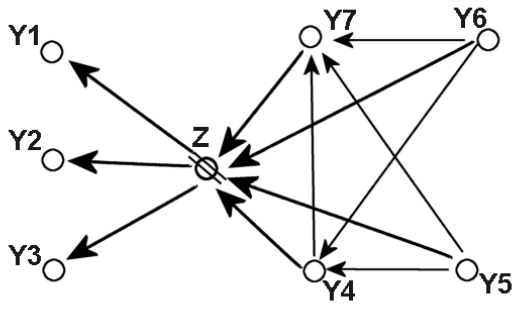

Figure 3.2: *DAG for Criminological example*

Response categories on these items ranged from "not at all likely' to "very likely'. The proposed model was that variables Y_1, Y_2, Y_3 acted as indicators of a latent variable named *perceived risk of victimization.*

Considering a linear structural equations systems represented with the DAG in Figure 3.2, where Z is the latent variable.

We want to estimate the relevant regression coefficients involving Z; the arrows between Y_4, Y_5, Y_6 and Y_7 are considered as dependencies of some interest and they do not imply causal relationship. The dag satisfies the first and the second condition of the Theorem 1.3.1 then the model is globally identified. Fitting this model with the residual variance of the latent variable constrained to be one we get a deviance of 43.63 on 8 degrees of freedom. It can be seen from Table 3.2 that a significant fit can be achieved by adding a direct edge from age (Y_5) to the perceived risk of robbery (Y_1). The results from the new model are similar to those of the previous model with the addition of a positive effect of the respondent's age on the robbery indicator of perceived risk as displayed in Table 3.2. It can be seen the non significant z-statistic for the regression coefficient of sex at the 5% level; it appears that prior victimization (Y_4) and victimization rate (Y_7) have the greatest effect on the latent variable. The estimated residual variances for Y_1,

Table 3.1: *Estimated partial residual variances for the model in example 1*

	δ	s.e	z
$\delta_{1.Z}$	0.4691	(0.0303)	15.4818
$\delta_{2.Z}$	0.3611	(0.0132)	27.3560
$\delta_{3.Z}$	0.4390	(0.0161)	27.2671

Y_2, Y_3 and Z are shown in Table 3.1 with their standard errors and z-values. Table 3.3

Table 3.2: *Estimated regression coefficients for the model in the example 1*

Arrows	Estimates	s.e	z
$Y_1 \leftarrow Z$	0.6805	0.0486	14.0021
$Y_2 \leftarrow Z$	0.7350	0.0503	14.6123
$Y_3 \leftarrow Z$	0.6887	0.0480	14.3479
$Z \leftarrow Y_4$	0.2541	0.0303	8.3861
$Z \leftarrow Y_5$	−0.1343	0.0305	-4.4033
$Z \leftarrow Y_6$	−0.0168	0.0305	-0.5508
$Z \leftarrow Y_7$	0.2474	0.0299	8.2742
$Y_1 \leftarrow Y_5$	0.1165	0.0200	5.8250
Deviance	13.76	*df* 7	$p < 0.06$

gives the observed correlation matrix (upper triangular) and the estimated correlation matrix (lower triangular) for the last models.

Table 3.3: *Observed (upper diagonal) and Estimated (lower diagonal) covariance matrix*

Y_1	−	0.575	0.540	0.169	-0.014	-0.023	0.224	
Y_2	0.575	−	0.598	0.240	−0.144	−0.088	0.215	
Y_3	0.539	0.599	−	0.246	−0.128	−0.092	0.182	
Y_4	0.198	0.237	0.222	−	−0.184	−0.148	0.168	
Y_5	−0.014	−0.141	−0.132	−0.184	−	0.236	−0.027	
Y_6	−0.048	−0.082	−0.077	−0.148	0.236	−	−0.102	
Y_7	0.198	0.217	0.203	0.168	−0.027	−0.102	−	
Z	0.783	0.869	0.814	0.323	−0.191	−0.111	0.295	1.182
	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Z

Example 3.3 (Reader reliability in essay scoring). In an experiment to establish methods of obtaining reader reliability in essay scoring 126 examinees were given a three-part English composition, source: Jöreskog and Sorbom (1989). Each part required the examinee to write an essay and for each examinee, scores were obtained on the following: -Y_1 original part 1 essay

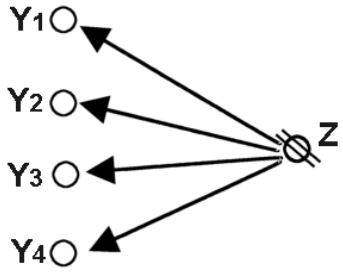

Figure 3.3: *Graphical model for reliability example*

-Y_2 handwritten copy of the original part 1 essay
-Y_3 a carbon copy of the handwritten copy
-Y_4 the original part 2 essay.
The investigator would like to know whether the four scores can be used interchangeably or whether scores on copies Y_2 and X_3 are less reliable than the originals Y_1 and Y_4.

The graphical model is the following

The estimated model is a measurement model, it satisfies the conditions for global identifiability. The results of the estimated parameters are given below on Table 3.4. The deviance of the model is 2.298 with df=$\{2\}$

Table 3.4: *Estimates coefficients β and partial residual variances δ for reliability example*

β	Est	s.e	δ	Est	s.e
$\beta_{y1.z}$	4.573	0.18	$\delta_{y1.z}$	4.160	0.51
$\beta_{y2.z}$	2.268	0.41	$\delta_{y2.z}$	21.04	2.66
$\beta_{y3.z}$	2.651	0.35	$\delta_{y3.z}$	15.71	1.98
$\beta_{y4.z}$	4.535	0.10	$\delta_{y4.z}$	1.3	0.16

In Table 3.4 the observed and the estimated covariances are reported. The reliability can be tested using the measures defined in Bollen (1989, Ch.6).

Example 3.4 (Hodge and Treiman's study (1968)).
Six variables were measured to study the relation between social status and social participation, see also Jöreskog and Goldberger (1975):
-CH Church attendance,
-ME Membership,

Table 3.5: *Observed (upper diagonal) and Estimated (lower diagonal) covariance matrix for reader reliability example*

Y_1	25.070	12.436	11.726	20.751
Y_2	12.239	28.202	9.228	11.973
Y_3	12.121	7.095	22.739	12.069
Y_4	20.739	12.139	12.022	21.871
	Y_1	Y_2	Y_3	Y_4

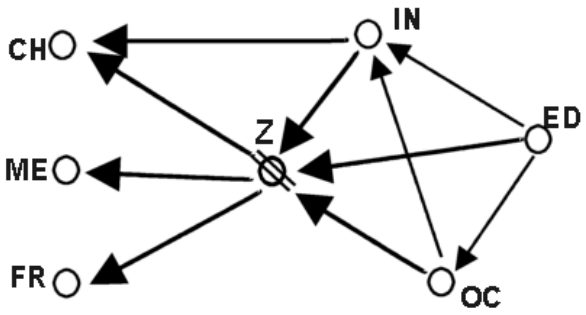

Figure 3.4: *Graphical model for the social status example*

-FR Friends seen,
-Z Social Participation,
-IN Income,
-OC Occupation,
-ED Education.

The postulated structural equation model corresponds to the graphical model in Figure 3.4. The model is globally identified.

The estimated regression coefficients using as constrain $\delta_{z.par(z)} = 1$ for the model without the edge between Church Attendance CH and Income IN has a deviance of 12.52 with 6 degrees of freedom.

Adding a direct edge from Income IN to Church attendance CH we get the estimates shown in Table 3.7 with a deviance of 6.79 with 5 degrees of freedom. The observed and the estimated correlation matrix are given in Table 3.6.

Table 3.6: *Observed (upper diagonal) and Estimated (lower diagonal) correlation matrix*

CH	$-$	0.360	0.210	0.100	0.156	0.158	
ME	0.348	$-$	0.265	0.284	0.192	0.324	
FR	0.208	0.286	$-$	0.176	0.136	0.226	
IN	0.100	0.287	0.169	$-$	0.304	0.305	
OC	0.129	0.211	0.124	0.304	$-$	0.334	
ED	0.208	0.312	0.184	0.305	0.344	$-$	
Z	0.499	0.697	0.410	0.411	0.303	0.477	$-$
	CH	ME	FR	IN	OC	ED	Z

Table 3.7: *Estimated regression coefficients for the model in the example 3.4*

Arrows	Estimates	s.e	z
$CH \leftarrow IN$	-0.127	0.052	-2.477
$Z \leftarrow IN$	-0.280	0.056	-5.017
$IN \leftarrow OC$	0.226	0.044	5.091
$Z \leftarrow OC$	-0.106	0.058	-1.849
$IN \leftarrow ED$	0.227	0.044	5.104
$OC \leftarrow ED$	0.344	0.041	8.306
$Z \leftarrow ED$	-0.326	0.058	-5.571
$CH \leftarrow Z$	-0.551	0.086	-6.369
$ME \leftarrow Z$	-0.692	0.077	-9.031
$FR \leftarrow Z$	-0.420	0.062	-6.593

3.5 Appendix A: Proof of (3.3.5)

To simplify the notation we write l_X for $l_X(\Sigma)/n$ and l_Y for $l_Y(\Sigma)/n$.

The first and the second derivatives of (3.3.2) when Σ is a function of a vector of parameters θ are

$$\frac{\partial l_X}{\partial \theta} = \frac{1}{2}\mathrm{tr}(\Sigma^i(\Sigma - S))$$

$$\frac{\partial^2 l_X}{\partial \theta_i \partial \theta_j} = -\frac{1}{2}\mathrm{tr}(\Sigma^{ij}(\Sigma - S) - \Sigma^i \Sigma \Sigma^j \Sigma)$$

where $\Sigma^i = \partial\Sigma^{-1}/\partial\theta_i$ and $\Sigma^{ij} = \partial\Sigma^{-1}/\partial\theta_i\partial\theta_j$. For the parametrization considered

$\Sigma^{-1} = A'\Delta^{-1}A$ the explicit derivatives have the form

$$\frac{\partial l_X}{\partial A_{ji}} = [(\Sigma - S)A\Delta^{-1}]_{ji}; \qquad \frac{\partial l_X}{\partial \Delta_{ii}} = \frac{1}{2}[A(\Sigma - S)A']_{ii};$$

$$\frac{\partial^2 l_X}{\partial A_{ji}\partial A_{ml}} = -(A^{lj}A^{im} + S_{li}\Delta^{-1}_{mj}); \qquad \frac{\partial^2 l_X}{\partial \Delta_{ii}\partial \Delta_{ll}} = -\frac{1}{4}(\Delta^{il}\Delta^{il} + \Delta^{il}\Delta^{il})$$

where A^{ji} denotes the $(j,i)th$ element of A^{-1} and Δ^{il} is the $(i,l)th$ element of Δ^{-1}.

To get the derivatives of the incomplete log-likelihood l_Y as in (3.3.1) Dempster *et al.* (1977) showed that

$$\frac{\partial l_Y}{\partial \theta} = E_{X|Y}\Big(\frac{\partial l_X}{\partial \theta}\Big)$$

the observed score is equal to the expected score of the complete data log-likelihood conditioned on the observed data. This expression becomes

$$\frac{\partial l_Y}{\partial \theta_i} = \frac{1}{2}\mathrm{tr}\Big(\Sigma^i(\Sigma - C)\Big)$$

where $C = C(S_{yy}|\Sigma)$ is defined in (3.3.3). The first part of the right hand side of (3.3.4) is minus the conditional expected value of second derivative of (3.3.2)

$$-E_{X|Y}\Big[\frac{\partial^2 l_X}{\partial\theta_i\partial\theta_j}\Big] = -\frac{1}{2}\mathrm{tr}(\Sigma^{ij}(\Sigma - C) - \Sigma^i\Sigma\Sigma^j\Sigma).$$

The second part of the right hand side of (3.3.4) can be written

$$-Var_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}, \frac{\partial l_X}{\partial\theta'}\Big) = -E_{X|Y}\Big\{\Big[\frac{\partial l_X}{\partial\theta} - E_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}\Big)\Big]\Big[\frac{\partial l_X}{\partial\theta} - E_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}\Big)\Big]'\Big\} =$$

$$= -E_{X|Y}\Big\{\frac{1}{2}\mathrm{tr}\Big[(\Sigma^i(C-S))(\Sigma^i(C-S))'\Big]\Big\} = -\frac{1}{2}\mathrm{tr}\Big(\Sigma^i C\Sigma^j C - \Sigma^i\tilde{C}\Sigma^j\tilde{C}\Big).$$

So (3.3.4) is established. In the parameterizations (A, Δ) we get the second derivatives

$$\frac{\partial^2 l_Y}{\partial A_{ji}\partial A_{ml}} = -(A^{lj}A^{im} + C_{li}\Delta^{-1}_{mj}) + C_{li}[\Delta^{-1}A'CA\Delta^{-1}]_{mj} +$$

$$+[CA\Delta^{-1}]_{mi}[\Delta^{-1}A'C]_{lj} - \tilde{C}_{li}[\Delta^{-1}A'\tilde{C}A\Delta^{-1}]_{mj} - [\tilde{C}A\Delta^{-1}]_{mi}[\Delta^{-1}A'\tilde{C}]_{lj}$$

And the derivatives respect to Δ^{-1}

$$\frac{\partial^2 l_Y}{\partial \Delta_{ii}\partial\Delta_{ll}} = -\frac{1}{4}\Big(\Delta^{il}\Delta^{il} + \Delta^{il}\Delta^{il}\Big) + \frac{1}{4}\Big\{[A'CA]_{il}[A'CA]_{il}$$

$$+[A'CA]_{il}[A'CA]_{il} - [A'\tilde{C}A]_{il}[A'\tilde{C}A]_{il} - [A'\tilde{C}A]_{il}[A'\tilde{C}A]_{il}\Big\}.$$

3.6 Appendix B

R code to computed the inverse variance and covariance matrix of the complete data log-likelihood.

```
der2 <- function(amat, latent,  fit, Syy,  n)
{
   cmqi <- function (Syy, Sigma, z)
     {
     ## Computes the matrix C(M | Q) by Kiiveri (1987), Psychometrika.
     ## It is a slight generalization in which Z is not the last element.
     ## z is a Boolean vector indicating the position
     ## of the latent variable in X.
     y <- ! z
     Q <- solve(Sigma)
     Qzz <- Q[z,z]
     Qzy <- Q[z,y]
     B <- - solve(Qzz) %*% Qzy
     BSyy <- B %*% Syy
     E <- Sigma*0
     E[y,y] <- Syy
     E[y,z] <- t(BSyy)
     E[z,y] <- BSyy
     E[z,z] <- BSyy %*% t(B) + solve(Qzz)
     dimnames(E) <- dimnames(Sigma)
     E
   }
   A<-(fit$Ahat)
    ## the functions take the output returned by fitDagLatent
   Shat  <- (fit$Shat)
   Khat  <- solve(Shat)
   Dhat <- (fit$Dhat)
   Delta <- diag(fit$Dhat)
   d1<-solve(Delta)
   dimnames(d1) <- dimnames(A)
   AA <- solve(A)
   nod <- rownames(amat)
```

```
nam <- rownames(Syy)
sek <- intersect(nam, nod)
sek <- c(sek, latent)
amat <- amat[sek,sek, drop=FALSE]
nod<-rownames(amat)
wherez <- is.element(nod, latent)
QQ <-cmqi(Syy, Shat, wherez)
q <- ncol(A)
H <- matrix(0, q ,q)
H[wherez,wherez] <- 1/(Khat[wherez,wherez])
Qtil <- QQ-H
Stil <- Shat-H
e <- d1%*%t(A)%*%QQ%*%A%*%d1
f <- QQ%*%A%*%d1
g <- t(f)
nn <- d1%*%t(A)%*%Qtil%*%A%*%d1
vv <- Qtil%*%A%*%d1
va <- t(vv)
ij <- matrix(nod[allEdges(amat)], ncol=2)
## Names of vertices of the edges
ij[, 2:1]
## The order of the indices in S is reversed wrt amat
p <- nrow(ij)
k <- c()
#print(ij)
for(u in 1:p)
{
  for(v in 1:p)
  {
    i <- ij[u,1]; j <- ij[u,2]
    l <- ij[v,1]; m <- ij[v,2]
    new <- AA[l,j]*AA[i,m]- QQ[l,i]*d1[m,j] + QQ[l,i]*e[m,j]
           + f[m,i]*g[l,j] - Qtil[l,i]*nn[m,j] - vv[m,i]*va[l,j]
    k = c(k,new)
  }
}
p<- nrow(ij)
```

```
k <- matrix(k, p, p, byrow=TRUE)
ed <- paste(ij[,1],"->", ij[,2], sep="")
dimnames(k) <- list(ed,ed)
p <- ncol(Syy)
npar <- apply(topSort(amat), 2, sum)
df<- (n-npar)
kinv  <- solve(-k)
kinv1 <- kinv/df
seA <- sqrt(diag(kinv1))
uu <- t(A)%*%QQ%*%A
oo <- t(A)%*%Qtil%*%A
nek <- - 1/4*(d1%*%d1 + d1%*%d1)
       + 1/4*(uu%*%uu + uu%*%uu) - 1/4*(oo%*%oo + oo%*%oo)
seD<-diag(sqrt(solve(-nek)/df))
n<-ncol(A)
pp = c()
At<- -t(A)
 for (i in 1:n) {
   for (j in 1:n) {
     if (amat[i,j]==1)
     pp=cbind(pp,At[i,j])
    }
  }
pp <-as.vector(pp)
TT <- pp/seA
TT <- as.vector(TT)
names(TT) <- names(seA)
PP<-2*(1-pnorm(abs(TT)))
tab <-round(cbind(value = pp, s.e. = seA, z = TT, p = PP),4)
DD <- Dhat/seD # per Delta
DD <- as.vector(DD)
P <- 2*(1-pnorm(abs(DD)))
tab2<-round(cbind(value = Dhat, s.e. = seD, z = DD,p = P),4)
cat("\n")
list(tab = tab, tab2 = tab2)
 }
```

3.7 References

Bollen K. A. (1989). *Structural equations with latent variables.* Wiley, New York.

Cox, D.R., Wermuth N. (1996). *Multivariate Dependencies - Models, Analysis and interpretation.* London: Chapman & Hall.

Cox D. R., Wermuth N. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society*, B, to appear.

Dempster A. P. (1972). Covariance Selection. *Biometrics*, 28, 157-175.

Dempster A. P. Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, 39, 1-38.

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

Hodge R.W., Treiman, D.J. (1968). Social participation and social status. *American Sociological Review*, 33, 723-740.

Jöreskog K. G. (1981). Analysis of covariance structures. With discussion. *Scandinavian Journal of Statistics*, 8, 65-92.

Jöreskog, K. G., Goldberger (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. . *Journal of the American Statistical Association*, 10, 631-639.

Jöreskog K., G. Sorbom D. (1989). *LISREL 7 - A guide to the Program and Applications.* 2nd Edition, SPSS Publications, Chicago.

Kiiveri H. T. (1982). A unified approach to causal models. Ph.D thesis. University of Western Australia.

Kiiveri H. T. (1987). An incomplete data approach to the analysis of covariance structure. *Psychometrika*, 52, 539-554.

Lauritzen S. L. (1996). *Graphical Models.* Oxford: Clarendon Press.

Lauritzen S. L. (2002). gRaphical models in R, *R News*, 3(2)39.

Lauritzen S. L., Wermuth N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17, 31-57.

Louis T. A. (1982). Finding the observed information matrix when using the EM-algorithm. *Journal of the Royal Statistical Society, B, 44, 226-233.*

Marchetti G. M., Drton M. (2003). `ggm`: an `R` pachakage for Gaussian graphical models, URL: *http://cran.r-project.org/.*

Orchard T., Woodbury M. A. (1972). A Missing Information Principle: Theory and Applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, (vol.1), 697-715.

Patterson S. (1984). *Journal of research in crime and deliquency*, Vol. 21 No. 4, 333-352.

Rubin D. B., Thayer D.T.(1982). EM Algorithms for Maximum Likelihood Factor Analysis. *Psychometrika*, 47,1, 69-76.

Sewell W. H., Haller A. O., and Ohlendorf G. W. (1970). The educational and early occupational status attainment process: revisions and replications. *American Sociological Review*, 35, 1014-1027.

Stanghellini E., Wermuth N. (2003). On the identification of path-analysis models with one hidden variable. *Submitted for publication in Biometrika.*

Tanner M.A. (1996). *Tools for statistical inference.* New York: Springer.

Wiley D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger and O. D. Duncan, eds., *Structural Equation Models in the Social Sciences.* New York: Academic Press, pp. 69-83.

Chapter 4

Classifying criminal activity: A latent class approach to longitudinal event data

Abstract: This article addresses the problem of classifying criminal behaviour, that is, of determining which types of offending co-occur in time. Using the Home Office Offenders Index, the complete criminal histories of a sample of those born in 1953 are examined. A local likelihood approach to latent class analysis is proposed, and is used to classify criminal activity, and state transitions over time can be examined. The proposed methodology can be used to classify other types of longitudinal event history where the interest is in the changing nature of activity over time.

Keywords: Criminal Careers; Latent class models; Local likelihood; Criminal pathways

4.1 Introduction

In criminological research, a common task is that of classifying criminal behaviour - of determining which types of criminal activity co-occur in time. In this paper, we are specifically concerned with allowing for behavioural change over the life course - thus the concern is focused specifically on identifying specific groups of criminal activity in

[1]This is a joint work with Professors Brian Francis and Keith Soothill. Some results of the work have been presented at the 54th Session (2003) of the International Statistical Institute (Berlin) and at the Academy of Criminal Justice Sciences Annual Meeting 2004 (Las Vegas) by B. Francis, F. Pennoni and K. Soothill. I was supported by Marie Curie Fellowship scheme. The authors are grateful to David Firth for helpful comments.

time and allowing individuals to move between these groups as an individual proceeds in their criminal history.

We start this paper by considering how such a task has been addressed by criminologists, first by attempting to classify individuals, and then through more recent dynamic approaches. Classification of criminals was an early concern of criminal research; for example Lombroso (1876) - often designated "the father of criminology" attempted to classify offenders according to facial appearance and head shape, and developed typologies of offenders. This essentialist stance of classifying the criminal rather than the behaviour continued for much of the century since Lombroso's pioneering text. The introduction of labelling theory and the work of Becker (1963) and Matya (1964), amongst others, stressed the importance of the social process and the societal reaction to the criminal act. Labels and labelling became pivotal in explanations, and Clinard and Quinney (1973) began to capture this shift, labelling individuals with criminal behaviour as "professional fringe violator" or with psychological labels such as "self-centred over indulged person". For much of this work, typologies were often developed without statistical methodology, with the result that there were poor allocation rules of new cases to groups (Gibbons, 1972) and by the mid 1970s there was, indeed, a despondency about the future of research in typologies (Gibbons, 1975).

More recent work on typologies (Nagin and Land, 1993; Nagin, *et al.*, 1995) has started to examine changes in activity over time, concentrating on the frequency or quantity of offending rather than the type or quality of offending. This approach is based on a definition of a "criminal career" as the " characterization of the longitudinal sequence of crimes committed by an individual offender" (Blumstein *et al.*, 1986). The appropriate way to understand criminal career is to take a longitudinal look at individuals over the life-course. (Farrington and West, 1990, Lauritsen, 1998, McCord 1982). "The criminal career approach emphases the need to investigate such questions as why people start offending (onset), why they continue offending (persistence), why offending becomes more frequent or more serious (escalation), and why people stop offending (desistence). The factors influencing onset may differ from those influencing other criminal career features such as persistence, escalation, and desistance, because the different processes occur at different ages" (Farrington, 1997). Addressing these and related issues requires knowledge about individual criminal careers, they initiation, their termination and the dynamic changes between these end points (Blumstein *et al.*, 1986).

The approach of Nagin *et al.* (1993), using a semi-parametric Poisson model, searches for subgroups of offenders with similar patterns in offending frequency over the criminal career and concentrates on the varying frequency of total offending over

time. In this way different trajectories of offending frequency over time can be identified - such groups include "adolescence limited" (who offend then stop within adolescence), and the "high level chronics" (who continue to offend at a high rate).

A final strand of work has examined the issue of transitions and specialisation. The mix of different offence types among active offenders is another important criminal career dimension. Wolfgang (1972) suggested the formation of transition matrices to assess offence specialisation over time, with transition probabilities from offence i to offence j being estimated. Stander *et al.* (1989) using data on 698 male adult prisoners, examined transitions between Home Office offence groups rather than offences and found that offending transition behaviour over the first ten occasions was stationary. However, problems with this approach can be seen. Offenders would tend to commit many different types of offence during their careers, and it is quite likely that even those who specialise in fraud, or violence will also be involved in petty theft if the opportunity arises. Moreover, the reliance of this approach on transition between conviction occasions, means that age is not controlled for, for example, some individuals the move between the 2nd and 3rd conviction will occur before 16, and for others after age 25.

Thus, on the one hand, methods assessing transition in criminal behaviour are not satisfactory, and on the other hand, work on typologies is mainly concentrating on the frequency of crime rather than its varying nature. The problem is to identify which offences co-occur at the same period or age in the criminal history of an offender and to identify the number of types and the nature of these types of activity. In addition, criminal careers develop over time and techniques to describe them should therefore also be able to capture transitions from one offence *type* to another. The technique should be able to allow transition matrices to be examined for any age, although actual behaviour transitions may well occur at different ages for different individuals.

This paper adopts a latent class approach, which provides advantages over other methods of typology construction: it is based on a well formulated statistical model, and does not assign units absolutely to classes but estimates posterior probabilities of class membership for each unit. Latent class analysis is an increasingly popular technique in criminological applications (see, e.g. Fergusson *et al.*, 1991; Van der Heijden *et al.*, 1997; Uebersax, 1997); moreover the substantial body of work on criminal trajectories has been recognised as a form of latent class analysis (d'Unger *et al.*, 1998).

This paper proceeds as follows. Section 4.2 describes the dataset of criminal conviction histories which provides the motivation for this work. Section 4.3 describes existing work in examining patterns of criminal behaviour, and proposes a new more flexible approach based on local likelihood, and Section 4.4 provides the results of the analysis. The paper concludes with a brief discussion.

4.2 The data

The data set used for the present analysis was derived from the Offenders Index, a Home Office research data set. The Offenders Index is a court based record of the criminal conviction histories of all offenders in England and Wales from 1963 to the current day which contains a one in thirteen sample of all offenders born in 1953. We have restricted our analysis to the Offenders Index cohort data for 1953, which consists of all offenders born in four specified weeks in 1953, with their complete criminal histories from the age of criminal responsibility, 10 years, until the end of 1993. This provided us with a set of offenders which are a one in thirteen sample of all offenders born in 1953. The index stores the dates of convictions, the detailed offence code of the conviction and the disposal or sentence.

Some features of the data set need to be mentioned. As the data set is based on conviction records, it has the limitation that neither arrests nor cautions are included so not all criminal activity is registered. In addition, the Offenders Index records standard list offences only - thus some of the more minor offences tried in magistrates courts are omitted - these are mainly less serious motoring offences (Francis *et al.*, 2004). The conviction histories are formed by joining together court conviction records through a record linkage process, matching convictions to individuals by name, gender and data of birth. This process may be subject to error, particularly for female offenders (Francis and Crosland, 2002). Finally, the Offenders Index contains no record of deaths, immigration or emigration, and so there is no information on when individuals are at risk of conviction and when they are not. This includes "immigration" and "emigration" to Scotland and Northern Ireland. However, these limitations are balanced by the strength of the data set of providing complete criminal conviction histories for all offenders in England and Wales from 1963. To date, the complete offenders index contains data on over 6 million individuals; the publicly available 1953 cohort data on 11402 individuals appearing in an England and Wales Court between 1963 and 1993.

Following the work of Francis *et al.* (2004) we simplify the data, reducing the offence codes to 73 major offences, after combining categories and eliminating offences with less than ten occurrences in the whole cohort (thus removing 0.005% of sample numbers). For example, murder was combined with manslaughter and attempted murder.

The data consists of $9,232$ male offenders and $2,170$ females; so that 81% of sample members begin males and 19% female.

Prime *et al.* (2001) analysed the criminal histories of offenders born in 1953, 1958, 1963, 1968, 1973 and 1978. It can be seen from Figure 4.1 that there are differences within the six birth cohorts examined and in the percentage of the population with a

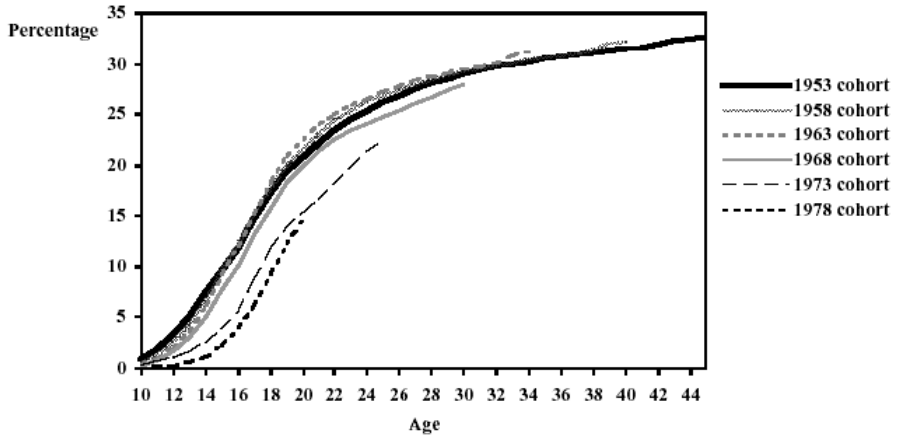

Figure 4.1: *Estimated Cumulative percentage of the male population with a conviction for various birth cohorts. Source: Prime et al. 2001.*

conviction. Before the age of fifteen, males born in a later cohort years are less likely to be convicted for an offence than males born in 1953. After the age of fifteen, males born in 1978 and 1963 are more likely to have been convicted of an offence, with 20 and 21 per cent respectively convicted for an offence before the age of twenty compared to 19 per cent of the population born in 1953.

They report also that thirty three per cent of men born in 1953 had at least one conviction for a "standard list" offence before the age of forty six. Figure 4.1 shows that for the 1953 birth cohort, the percentage of the population with at least one conviction increases with age, rising from 7 per cent before the age of fifteen, to 19 per cent before the age of twenty; 28 per cent before the age of thirty, and 31 per cent before the age of 40. Most offenders are first convicted of an offence between the ages of about thirteen and twenty. The number of new offenders tails off with increasing age and only 2 per cent of the population are first convicted of an offence in their late thirties to mid forties.

The cumulative percentage of the female population with a conviction is showed in Figure 4.2. They also show that nine per cent of women born in 1953 had been convicted of a standard list of offence before the age of forty six. At the younger ages the percentage of the female population with a conviction is only between a eighth and a twentieth of the percentage of the male population with a conviction. The differences between six birth cohorts, shown in Figure 4.2, are very similar to those observed for males in Figure 4.1.

To carry out our analysis we simplify the time axis, using age in completed years

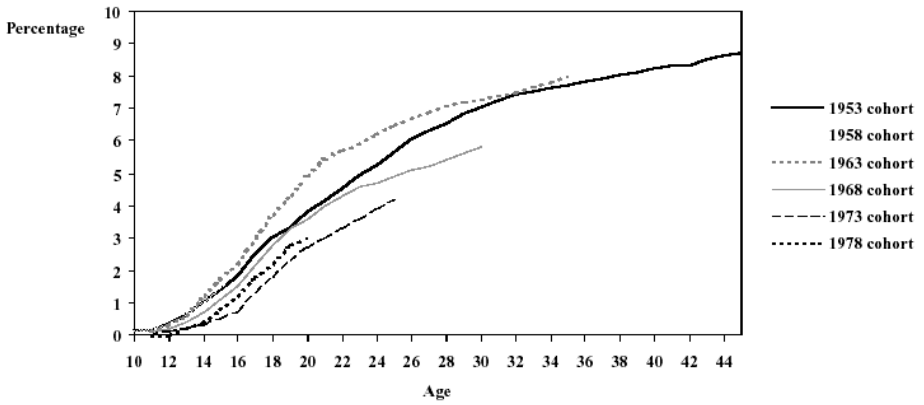

Figure 4.2: *Estimated cumulative percentage of the female population with a conviction for various birth cohorts. Source: Prime et al. 2001.*

rather than working with continuous time - age thus takes 30 distinct values, from 10 to 40. We choose to analyze males and females separately, as the expectation is that they will have rather different patterns of criminal activity.

Court data provide information on all offences at each record appearance. The data set indicates that 47.6 per cent of offenders who appear in court are charged with one offence only. Younger man predominated, 45.7 per cent being between 10 and 20 at the time of the sentence, 21 per cent between 21 and 25, and 32 per cent over 26 years.

4.3 Models for local patterns in event data

The problem of determining classes of criminal activity which co-occur in time can be rephrased more generally as one which identifies common patterns of event types which co-occur in a small time interval in a collection of sequences of longitudinal event data, where typically there will be a large number of different event types. A related problem has been addressed by Abbott and Barman (1997), who, in examining the structure of academic papers in the American Journal of Sociology, determined different types of sentence, and then attempted to find common subsequences of sentence types which occur in these articles. However, this work is not strictly relevant for criminal conviction data as Abbott and Barman were primarily concerned with the order of events within the subsequence, whereas the particular ordering of criminal convictions at a point in a local time neighbourhood is not important. More recently, work by Hellerstein *et al.* (2002) has addressed the problem of detecting patterns of event in a computer networks using an approach entitle “multi attribute frequent pattern mining” but their approach

has no underling statistical model.

The approach in this study is to use latent class analysis to determine classes of criminal activity. Two methods of carrying out such local clustering in time are considered - the "segmentation" approach and a "local neighborhood" approach proposed here.

4.3.1 The segmentation approach to local clustering

We provide a short summary of the segmentation approach, which was proposed by Francis *et al.* (2004). This approach segments the age axis into a number of predetermined age group segments, and examines offending behaviour within those age groups. Francis *et al.* chose the age groups 10-15, 16-20, 21-25, 26-30, 31-35, with a five year window apart from the first group, and termed each age segment an *age strip*. The primary consideration was to obtain an interval of sufficient length to gain a picture of the individual's offending career in that portion of the life course.

A record was constructed for each age strip for each individual, which consisted of a set of binary items, one for each offence category, with each item coding 1 if the offender had been convicted at least once for that offence category in the age group under study, and coded 0 if not. Records were omitted if they contained no convictions. Thus for each offender i, a set of $1 \geq r(i) \leq 6$ records was constructed, one for each age strip - the total number of strips was $S = \sum_i r(i)$ and the strips were indexed by s. The complete data for all offenders in each study therefore formed what can be thought of as a *prevalence matrix*.

The latent class model was specified as follows. Let O_{sj} be the observed binary response for strip s and offence category j, with $O_{sj} = 1$ if offence j is committed within the strip s, and 0 otherwise. Over all s and j, the O_{sj} form a prevalence matrix $\boldsymbol{O}$, which has rows O_s. We suppose that here there are K classes or clusters in the date set; let class k $(k = 1, ..., K)$ have probability $\pi(k)$. Define p_{jk} to be the probability that there is at least one conviction for offence j in a strip, given that the strip belongs to class k. Than we can write the likelihood L of observing the prevalence matrix $\boldsymbol{O}$ as a weighted sum over the K classes of individual class likelihoods:

$$L = f(\boldsymbol{O}) = \prod_i \prod_r^{r(i)} \sum_k \pi(k) p(O_s|k) = \prod_s^S \sum_k \pi(k) p(O_s|k),$$

where

$$p(O_s|k) = \prod_j p_{jk}^{O_{sj}} (1 - p_{jk})^{(1-O_{sj})}.$$

We assume that any correlation that might exist within age strips belonging to the same offender is modelled by the latent class structure. The latent class problem then becomes one of obtaining estimates of the p_{jk} for each strip and of the $\pi(k)$ by maximizing the above likelihood.

Unlike other methods of clustering, there has been much work and discussion on the appropriate measure to be used when determining K, the optimal number of latent classes or clusters. The concept of optimality is that there will be some number of classes where all classes are distinct, but where adding an additional class to the model provides no extra explanatory power. However standard likelihood theory breaks down, because restricting a $k+1$ class solution to k classes would involve setting $\pi(k+1)$ to zero, giving a likelihood ratio test for a parameter that lies on the boundary of the parameter space. This makes formal statistical significance testing using differences in log-likelihoods impossible. To prevent this, as showed in Section 2.1.1 of Chapter 2, most authors suggest to use information criteria such as the AIC (Akaike, 1974) and BIC statistic (Schwarz 1978, Raftery, 1995). As showed in Section 2.1.2 both are based on the likelihood but apply corrections for the number of parameters fitted and additionally (for BIC) for the number of observations. Guidance is provided by the appendix to D'Unger *et al.* (1998), which suggests that the BIC should be used.

We briefly illustrate the main findings of such analysis with Table 4.1 and 4.2 where it is shown a brief "pen-portraits" (Francis *et al.*, 2004) for each cluster for males and females respectively. The best solution for the *male* cohort was provided by nine latent class model and three class model provided the best solution for describing the five-year *female* criminal histories. Table 4.1 and 4.2 reports the estimated cluster proportions for the nine clusters for males and for three cluster for females. The names given to the clusters are based on the conditional posterior probability for an age strip to be in a cluster having one or more conviction for a particular offence category j.

The segmentation approach has a number of problems. First, the size of the strip and the strip cut-points are pre-determined rather than determined from the data. There is no reason why the cut-points of 16, 21, 26 etc. are the correct values to choose, and more carefully chosen cut-points might give a different solution. Additionally, every offender might have their own personal pathway through criminal behaviour, and different offenders might have different cut-points and pathways. For example, one offender may make a transition from one pattern of offending to another at age 18, another may not make a transition until age 22, and a third may not make a transition at all. We therefore generalise the above approach.

Table 4.1: *Male results using the segmented approach: nine cluster solution*

No. of Clusters	Proportion	Offending profiles
1	18.5%	Marginal lifestyle
2	16.6%	Non-violent property
3	12.3%	Fraud and general theft
4	11.8%	General violence
5	9.9%	Petty theft
6	8.6%	Aggressive property offending
7	8.3%	Vehicle theft
8	7.8%	Wounding
9	6.0%	Shoplifting

Table 4.2: *Female results using the segmented approach: three cluster solution*

No. of Clusters	Proportion	Offending profiles
1	59.4%	Versatile offending
2	36.3%	Shoplifting
3	4.3%	Trust violation

4.3.2 A local neighborhood approach

An alternative way of proceeding is to form a set of individual-specific offence prevalence vectors for age strips centered at each possible age in the data set and to carry out a clustering procedure on all active age profiles for all offenders. We need to define a local neighbourhood $N(a_t)$ around each age a_t, i.e. a pre specified time interval around a_t. Let the neighbourhood around a_t extend from a_{t0} to a_{t1}, where $a_{t0} \leq a_t \leq a_{t1}$, than for each local neighbourhood offences falling within the neighbourhood are identified which contribute to the construction of the prevalence vector at age a_t, and offences outside the neighbourhood which do not.

Following the ideas of local regression smoothing, and local likelihood, we can specify the nature and the size of the local neighborhood through a kernel function K. For example, Wu and Tuma (1990) proposed a local hazards survival analysis which was developed by Betensky *et al.* (2002) and placed in the framework of local likelihood. The idea in these papers is to estimate the hazard function locally - within a defined neighbourhood region defined by a kernel function, while still obtaining global estimates of the regression parameters. The kernel function in our application will similarly weight

points in a region around an age a_t defined by a chosen range of the neighbourhood or *bandwidth* h, giving us local estimates of the probability of class membership at any age. When the neighbourhoods overlap it allows for events at a given age to contribute to more than one strip with weights determined by the kernel, and thus introduces a degree of smoothing in the way latent class membership changes from year to year.

As in local regression also here the choice of kernel function does not greatly influence the efficiency of the approximation of the target function (see e.g Hastie and Tibshirani, 1990), more important is the choice of the value of the bandwidth because it controls the degree of smoothing. The choice of h can be viewed as a smoothing parameter on the analysis. Increasing the value of h will decrease the ability of the model to detect changes over time, until, when h approaches the age range, we are analysing the offence profile of an individual's complete criminal career rather than periods within individuals. Small values of h will induce substantial noise in the age to age variation of latent class membership.

After examining various bandwidth in preliminary analyses we chose symmetric range of the neighbourhood. We report results based on a bandwidth of five years.

Various choices of symmetric kernel function can be made - Fahrmeir and Tutz (2001) for example, give a wide range of choices. A straightforward choice is to follow Wu and Tuma (1990) and choose a rectangular or uniform kernel function with overlapping regions centred on a_t:

$$K(t) = K\left(\frac{a - a_t}{h}\right) = \begin{cases} 1/(2h+1) & |a - a_t| \leq h \\ 0 & otherwise \end{cases}$$

Other kernels could be used but are less useful when constructing binary prevalence vectors.

4.4 Model Specification

As the previous analysis we use a latent class cluster analysis, a model-based clustering procedure which assigns windows to classes with estimated probabilities. As mention in Chapter 2 one advantage of model based clustering is that it provides a precise framework for assessing the resulting partitions of the data and especially for choosing the relevant number of clusters.

To estimate a model where the latent classes are estimated globally over all data points we need to treat the data points like local events in the neighbourhood of age a_t

[2]Further work is needed on the consequences for estimator of choosing different bandwidth, i.e. of choosing symmetric versus asymmetric range. Such work is beyond the scope of this study.

(Tibshirani and Hastie, 1997). We extend the definition of the vector O to be vector of binary indicators where $O_{jt} = 1$ if offender i is convicted for offence j in a region defined by the bandwidth centered on age a_t. The binary variables $O_{jt} = 1$ are collected into the prevalence vector O_t. Than $p_{jt|k}$ is the probability that there is at least one offence of type j the region given the membership of a_t on class k; and $\pi(k)$ is the probability of the latent class in the region.

With N individuals the likelihood can be expressed as the product of the conditional pattern probabilities in all the regions defined by the kernel

$$L(O_t) = \prod_i \prod_t \sum_k \pi(k) p(O_{it} = o_{it}|k) K(t)$$

where

$$p(O_{it} = o_{it}|k) = \prod_j p_{jt|k}^{o_{jt}} (1 - p_{jt|k})^{1-o_{jt}}$$

The posterior probability of the class membership overall the points of the regions of the bandwidth is

$$H(k|\boldsymbol{O}_{it}) = \frac{\prod_i \prod_t \pi(k) p(O_{it} = o_{it}|k) K(t)}{L(O_{it})}.$$

The choice of the uniform kernel permitted to carry out the estimation of the model as in standard latent class analysis. The analysis were carried out with the software program Latent Gold Version 3.0.1 (Vermunt and Magidson, 2000) which can deal with large numbers of items and has the ability to deal with individual level data rather than data in tabular form.

To find maximum likelihood for the model parameters Latent Gold uses both the EM (Dempster *et al.*, 1977) as defined in Chapter 2 and the Newton Rahpson algorithm (McHugh, 1956, 1958). The program starts with the EM until either the maximum number of EM iterations or the EM convergence criterion is reached. Than the program switches to NR iterations which stops when the maximum number of NR iterations or the overall convergence criterion is reached. This is a way to combine the advantages of both algorithms, that is, the stability of the EM even when far away from the optimum and the speed of the Newton-Raphson when close to the optimum (cf. Vermunt and Magidson, 2002).

The greater the number of latent classes the more the models can suffer of identifiability problems as such described in Section 2.1.1 of Chapter 4. To avoid obtaining a local rather then a global solution we used a multiple set of starting values in order to check for different solutions of log-posterior values. We also choose to perform 500 EM

iteration and 150 NR iterations. The algorithms were quite slow to converge because of the large size of the data set.

It should be note that the model considered in this article typically implied non linear moment structures. It follows that local identification at one point in the parameter space does not imply local identification elsewhere in the parameter space and the parameter points can often be found which are not locally identified.

4.5 Results

Model selection was based on BIC and on the estimated proportions of non correct classifications. In fact when classification of cases is based on assignment to the class having the highest membership probability, the proportion of cases that are expected to be misclassified is reported by the statistic known as classification error (cf. Section 2.1.2). The closer this value is to 0 the better.

It should be noted that the BIC we use is computed by using the log-likelihood value and the number of parameters rather then by using the G^2 and the number of degrees of freedom. This means that data at the individual level is used to calculate the diagnostic values rather than data from full 2^{71} table of cross-classifications.

The BIC and the classification errors were recorded for both males and females. Table 4.3 reports the BIC and the classification error values for the estimated one to fourteen class models for males.

It can be seen that the BIC is lowest for the fourteen class model and it seems to decrease increasing the number of classes. In this work we have therefore used BIC as a guide, also taking into account other diagnostic values such as the classification error. From table 4.3 the classification error reaches the minimum value for the eleven and the twelve class models. For the principle of parsimony we assume the eleven class model to be the best solution for the *male* cohort.

Table 4.4 reports the cluster proportions together with a label assigned to each of the eleven latent classes and the top nine profile probabilities of cluster membership given that the offence of type j has occurred, which show how the clusters are related to the offences.

The strips are divided reasonably evenly between the clusters, as follows: cluster 1 - 19.3 per cent; cluster 2 - 13.2 per cent; cluster 3 - 12.5 per cent; cluster 4 - 10.3 per cent; cluster 5 - 9.9 per cent; cluster 6 - 8.8 per cent; cluster 7 - 7.8 per cent; cluster 8 - 5.8 per cent; cluster 9 - 5.1 per cent; cluster 10 - 4.7 per cent; cluster 11 - 2.6 per cent.

Hence, from the table it can be noted that the first two clusters in particular, include a vast range of criminal activity, together with cluster 7, cluster 8 and cluster 9. The

Table 4.3: *BIC and classification error values for males latent class analysis*

No. of Clusters	BIC	E_j
1	916583.33	0
2	904911.70	0.0547
3	898304.10	0.1999
4	893198.08	0.2468
5	889213.78	0.2145
6	884929.26	0.1558
7	880789.64	0.1506
8	878430.67	0.1152
9	875665.31	0.1744
10	872397.61	0.1104
11	870079.45	0.0724
12	867624.68	0.0679
13	867555.97	0.1180
14	864466.92	0.0882

Table 4.4: *Top nine profile probabilities for males: 11 cluster solution*

Cluster 1 *Marginal Lifestyle*		Cluster 2 *Mixed Offending*		Cluster 3 *Petty Theft*	
$\pi(1)$	0.19	$\pi(2)$	0.13	$\pi(3)$	0.12
54.Reciving stolen goods	0.20	49.Petty theft	0.56	49.Petty theft	0.99
45.Stealing	0.11	30,27.Commercial burglary	0.50	46.Shoplifting	0.05
104.Assault	0.10	Vehicle offences	0.45	54.Reciving stolen goods	0.04
28.Burglary in a dwelling	0.07	28.Burglary in a dwelling	0.32	33.Going equipped	0.03
165.Posssesion of weapons	0.05	Criminal demage	0.30	45.Stealing from vehicle	0.02
20.Assault on females	0.04	8.Maliciuos wounding	0.25	28.Burglary	0.02
47.Stealing/machine	0.04	54.Reciving stolen goods	0.23	47.Stealing from machine	0.02
195.Minor offences	0.04	Fraud	0.19	29.Aggravated/burglary	0.02
80,83.Absconding from custody	0.03	46.Shoplifting	0.17	Forgery	0.01

Cluster 4 *Demage, Violence*		Cluster 5 *Car Crazies*		Cluster 6 *Violence*	
$\pi(4)$	0.10	$\pi(5)$	0.10	$\pi(6)$	0.09
Criminal damage	0.99	Vehicle Offences	0.99	8.Maliciuos wounding	0.99
8.Maliciuos wounding	0.15	49.Petty theft	0.14	49.Petty theft	0.07
49.Petty theft	0.08	45.Stealing from vehicle	0.11	Criminal damage	0.03
104.Assault	0.05	30,27.Commercial burglary	0.08	104.Assault	0.03
195.Minor offences	0.04	33.Going equipped	0.06	46.Shoplifting	0.03
Vehicle Offences	0.04	28.Burglary in a dwelling	0.04	30,27.Commercial burglary	0.03
46.Shoplifting	0.04	Criminal damage	0.04	Vehicle Offences	0.03
30,27.Commercial burglary	0.03	195.Minor offences	0.03	34.Robbery	0.02
80,83.Abscombing from custody	0.02	54.Reciving stolen goods	0.02	54.Reciving stolen goods	0.03

Cluster 7 *Commer. Burglary*		Cluster 8 *Shoplifting*	
$\pi(7)$	0.08	$\pi(8)$	0.06
30,27.Commercial burglary	0.99	46.Shoplifting	0.99
49.Petty theft	0.19	54.Reciving stolen goods	0.02
28.Burglary in a dwelling	0.08	80,83.Abscombing from custody	0.02
Criminal demage	0.07	104.Assault	0.01
Vehicle offences	0.06	39.Stealing a person	0.04
46.Shoplifting	0.06	195.Minor offences	0.01
29.Aggravated burglary	0.05	-	-
54.Receiving stolen goods	0.04	-	-
32.Robbery	0.03	-	-

Cluster 9 *Fraud*		Cluster 10 *Drugs*		Cluster 11 *Stealing*	
$\pi(9)$	0.05	$\pi(10)$	0.05	$\pi(11)$	0.03
Offraud	0.99	Offdrug	0.99	41.Steeling by an employee	0.99
49.Petty theft	0.27	46.Shoplifting	0.05	52.Falsifying accounts	0.09
54.Receiving stolen goods	0.14	54.Receiving stolen goods	0.03	49.Petty theft	0.08
46.Shoplifting	0.08	49.Petty theft	0.03	54.Stealing	0.05
Forgery	0.08	80,83.Abscombing	0.03	Vehicle offences	0.04
80,83.Abscombing custody	0.07	Criminal damage	0.03	Fraud	0.03
30,27.Commercial burglary	0.05	30,27.Abscombing custody	0.02	Forgery	0.02
40.Stealing in a dwelling	0.05	28.Burglary in a dwelling	0.02	46.Shoplifting	0.02
8.Malicious Wounding	0.05	Vehicle offences	0.01	30,27.Commercial burglary	0.02

other clusters show a more specialized criminal activity. In fact the profile probability is high for one offence and low for the remainders. This result is satisfying, as it shows that offending groups are readily identified using the local likelihood approach. Whilst group in the segmentation approach were often defined by a wide range of offences, this is true only for a few of the clusters found in the analysis.

In terms of criminal activity, people behave differently at different ages. Considering the assignment of each age with the highest posterior probability to a cluster we can examine in details how offender change their offending behavior as they become older.

Considering *male* variations by age, for every age a_t, we identified the number of offenders classified into that offence cluster. A graph of such numbers is than obtained to have a picture of changing activity over time.

The histograms in Figure 4.3-4.4 show the row posterior probabilities - this have been smoothed to show the overall trend line, which has been superimposed. The offending profiles vary dramatically with age. Some interesting patterns emerge and compared to the "segmentation" approach they give a better picture. The problem of naming the clusters seems less difficult as offenders will often participate in a range of criminal activity, but they seem to have a predominant activity.

Combining the results from the tables and the histograms we can analyse the results as follows. Cluster 1 is the largest cluster and it is not a very specialized cluster, in

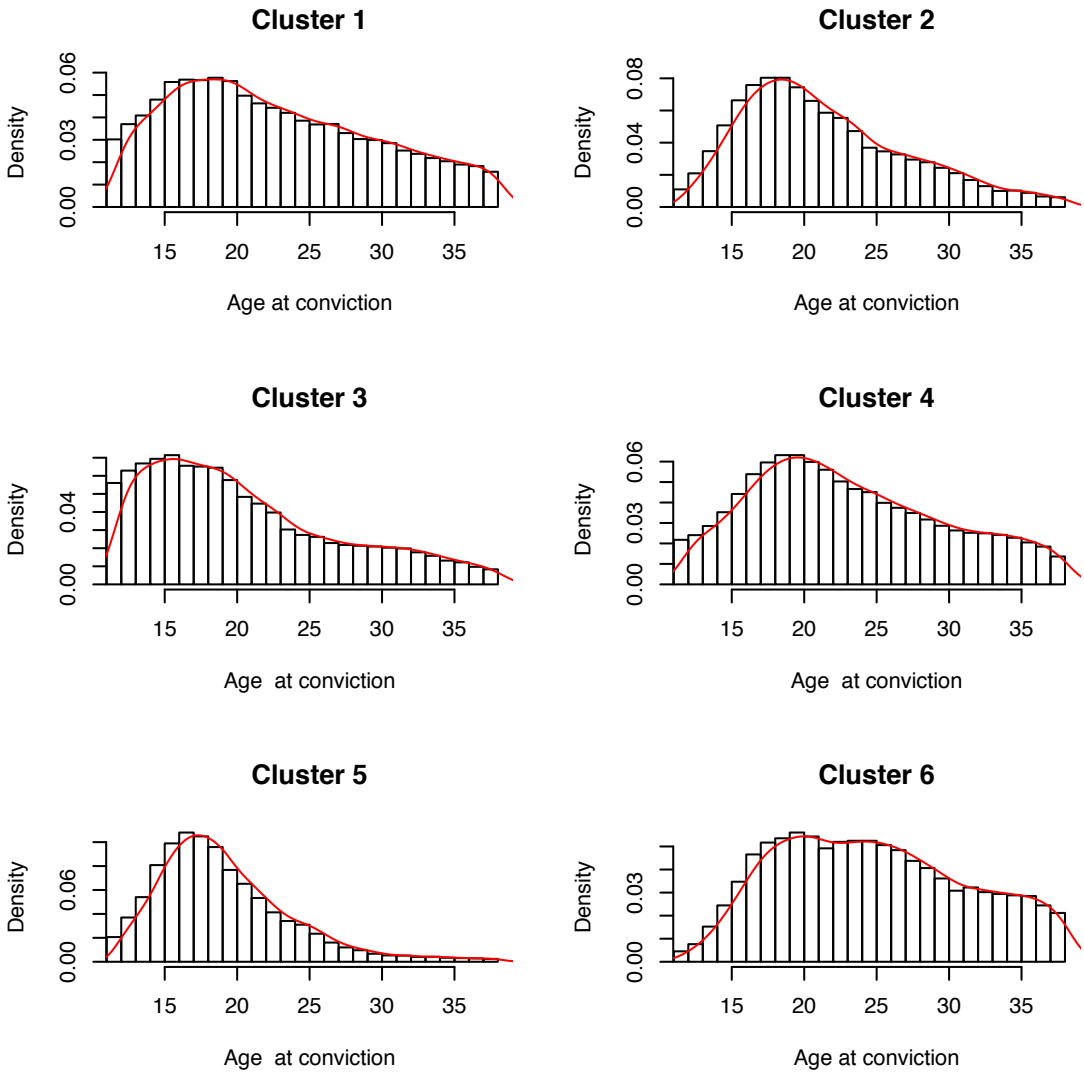

Figure 4.3: *Histogram of posterior probabilities of cluster membership for the first six clusters for males 1953 cohort.*

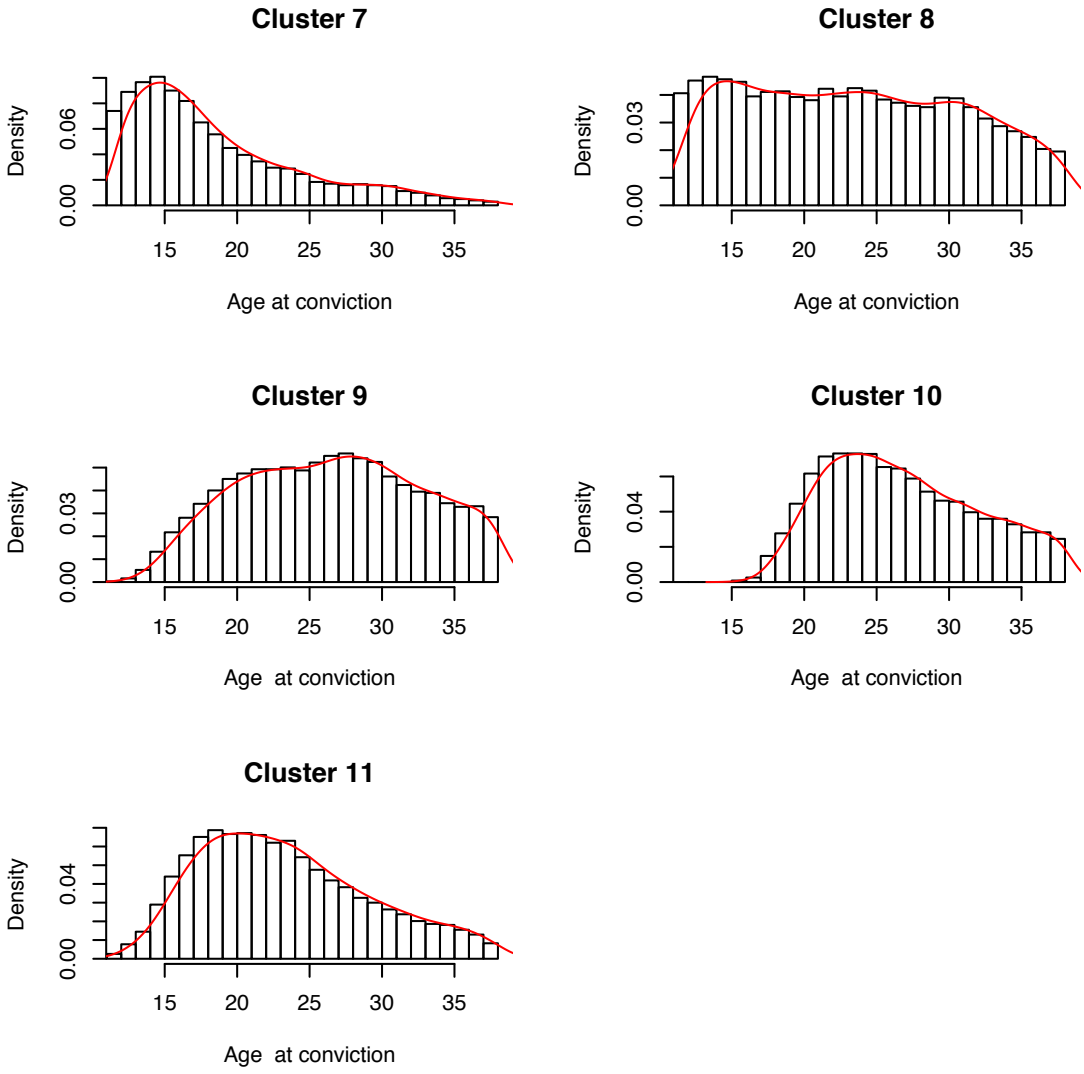

Figure 4.4: *Histogram of posterior probabilities of cluster membership for the last five clusters for males 1953 cohort.*

fact, it is characterized mainly by receiving stolen goods (0.20), stealing from a vehicle (0.11) and assault (0.10). It has high activity between 15 and twenty but it seems to be made of criminal activities which last over the years. Cluster 2 is also a not very specialized cluster: offenders are involved in petty theft (0.56), commercial burglary (0.50), theft from vehicle, theft of vehicles and driving licence offences (0.45); burglary in a dwelling (0.32) and other minor offences; it reaches a peak on age 18 and 19 and soon it seems to decline quite speedily. Cluster 3 predominantly involves petty theft (0.99) and it is a specialized cluster: the other offences has probability less or equal to 0.05; it is an offence which involve all young men until 20, it reaches the peak around age 15. Cluster 4 involves other criminal damage (0.99) and malicious wounding (0.15); it involves all ages over 15. Cluster 5 is characterized by non-violent offenders involved in vehicle theft (0.99) and petty theft (0.14) and other stealing (0.11); mainly 16-20 years old and then it soon declines. Cluster 6 is predominantly involved in malicious wounding, and it has the highest probability of having threats to murder and kidnap (0.99). It does not reach reach the peak until 20 with another peak around 24-25. Class 7 is characterized by a very hight probability of breaking into shops and commercial property (0.99), with smaller probability of pretty theft (0.19) and of burglary (0.08); it as hight activity between 12-20 and then it declines soon. Class 8 involves shoplifting (0.99) and also receiving or handling stolen goods (0.01); it can be seen that it has three peaks: one around 13-14, around 24-25 and 30-31. Class 9 is characterized by a very hight probability of fraud offences (0.997) with pretty theft (0.27) and receiving stolen goods (0.14) also contributing. It does not reach the peak until age 28. Class 10 is involved in drugs (0.99) and shoplifting (0.05); it does not happen before age 16 and it reaches the peak from age 22 to 25. Class 11 is involved predominantly in stealing from an employee (0.99) and false accounting (0.09).

Female

Using the same analysis of the outcomes for females we report in Table 4.5 the BIC and the classification error values for the estimated one to fourteen latent class model of females using the "local neighborhood" approach.

A six-cluster model was chosen which seems to provide the best solution for describing the *female* 1953 cohort criminal histories. Table 4.6 shows the labels assigned to each cluster, the estimated six cluster proportions and the top nine profile probabilities of cluster membership for each offence. Figure 4.5 reports the histograms of the posterior probabilities together with assignment of each age a_t to clusters.

The first four clusters are the largest. Cluster 1 has 14 per cent probability and it seems to involve predominantly the shoplifting offence with (0.99) and pretty theft

Table 4.5: *BIC and classification error values for female latent class analysis*

No. of Clusters	BIC	E_j
1	107631.80	0
2	101890.20	0.0321
3	100895.00	0.0322
4	100342.80	0.0965
5	99770.61	0.1066
6	99400.56	0.0679
7	99211.48	0.061
8	99122.03	0.0881
9	99098.11	0.0717
10	98918.20	0.0647
11	99049.74	0.0532
12	99510.87	0.0617
13	99349.82	0.0484
14	99438.28	0.0569

Table 4.6: *Top nine profile probabilities for female: 6 cluster solution*

Cluster 1 *Shoplifting*		Cluster 2 *Fraud*		Cluster 3 *Petty Theft*	
$\pi(1)$	0.38	$\pi(2)$	0.25	$\pi(3)$	0.14
46.Shoplifting	0.99	Fraud	0.28	49.Petty theft	0.99
49.Petty theft	0.02	40.Stealing in a dwelling	0.07	Fraud	0.14
54.Receving stolen goods	0.02	43.Absracting electricity	0.06	46.Shoplifting	0.05
-	-	47.Stealing from machines	0.06	54.Receving stolen goods	0.03
-	-	46.Shoplifting	0.06	30,27.Commercial burglary	0.02
-	-	28.Burglary in a d	0.03	40.Stealing in a dwelling	0.02
-	-	30,27.Commercial burglary	0.03	41.Steling by an employee.	0.02
-	-	45.Stealing from vehicle	0.02	-	-
-	-	-	-	-	-

Cluster 4 *Violence*		Cluster 5 *General criminality*		Cluster 6 *Stealing*	
$\pi(4)$	0.14	$\pi(5)$	0.04	$\pi(6)$	0.04
8.Maliciuos wounding	0.34	49.Petty Theft	0.86	41.Steeling by an employee	0.99
Criminal damage	0.28	46.Shoplifting	0.41	52.Falsifying accounts	0.18
104.Assault	0.18	Fraud	0.40	46.Shoplifting	0.03
46.Shoplifting	0.06	54..Receving stolen goods	0.34	Fraud	0.03
30,27.Commercial burglary	0.06	8.Maliciuos wounding	0.24	-	-
195.Other offences	0.05	Criminal damage	0.20	-	-
80,83.Abscombing cus.	0.05	Forgery	0.17	-	-
28.Burglary in a dwelling	0.04	Vehicle offences	0.16	-	-
54.Receving stolen goods	0.03	30,27.Commercial burglary	0.13	-	-

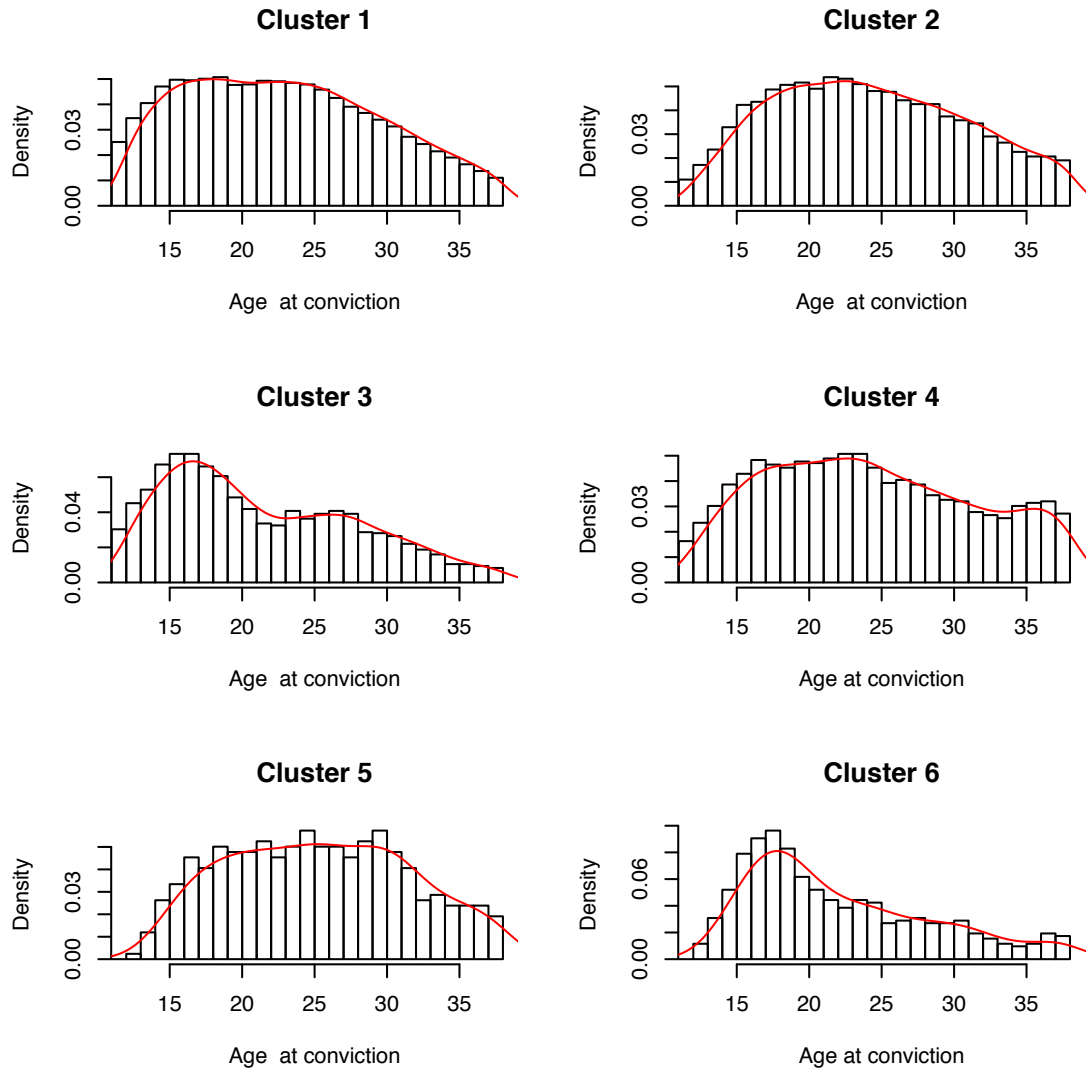

Figure 4.5: *Histogram of posterior probabilities of cluster membership for the six clusters for females 1953 cohort.*

(0.02); it is an all ages offence although it does tend to tail off once over 30. Cluster 2 is characterized by fraud (0.28), stealing in a dwelling (0.07) and abstracting electricity (0.06). It is also an all ages offence. Cluster 3 is petty theft (0.99) and fraud (0.14); it has a peak between 15 and 18, and also another peak around 27. Cluster 4 is another large cluster (0.14) per cent, it is involved in violence, assault (0.34) and criminal damage (0.28) and it is an all age crime with peak in 27 and another one in 37. Cluster 5 is not very large 4 per cent, it is characterized by petty theft (0.85) with shoplifting (0.42) and fraud (0.40). Cluster 6 is a small cluster 4 per cent; it is involved with sealing by

an employee (0.99) and false accounting (0.18); offenders are mainly 15-19 year olds .

For example, for female analysis, we can see that stealing from an employee is very much an activity for young girls, reaching a peak at age 17 and then declining. The probability of fraud, in contrast, hardly changes, from age 15 to age 25. Involvement with car offending appears to be a type of activity which offenders grow out of and move on to other activity, or to stop offending. Violent offending and being convicted for it, however, is more likely to continue until the offender is 30 and has a for less steep.

4.6 Discussion

It can be seen that age is a crucial variable in understanding criminal histories. The proposed local neighborhood approach seems to give a more accurate picture of the changing over time. The use of latent class model with local kernel smoothing to represent time related change provides a general framework to build behavioural typologies that encompasses situations involving for example a psychological study on the core behaviour of children.

Selecting an appropriate latent class model involves comparisons among models. We illustrate the potential usefulness if the BIC procedure and the classification error.

A more appropriate description of the changes over time could be useful, for example comparing different birth cohorts, because typologies may not remain constant over time i.e. typologies from 1983 birth cohort may not be the same as those from 1953 birth cohort.

However the analysis provide a case study in applying relative complex latent class models to a substantive problem in criminology. These results contributes to understanding the crime behaviour and can help to decide policy for juvenile crime.

An unresolved question is to better assess if there are many transitions between different latent classes and when such transitions occur. We think that hidden Markov models can be useful to generate understanding in the area of offence patterning. The basic assumption of this model is that the offending pattern of an offender within a certain age strip depends only on a discrete latent variable representing his/her tendency to commit crimes, which follows a first order homogeneous Markov process. Such model may be useful to test hypothesis of interest, for example by restricting appropriately the transition matrix models can be tested in which only pre-specified transition can occur. A disadvantage of such modelling is that it requires fixed intervals like the segmentation approach under representative offences have to be chosen for interval in which more than one offence occur. The application of hidden Markov modelling of

offence patterning will constitute material for further work.

4.7 References

Abbot A., Barman E. (1997). Sequence comparison via Alignement and Gibbs sampling. *Sociological Methodology*, 27, 47-87.

Akaike H. (1973). Information theory and an extention of the maximum likelihood principle. In B. N. Petrov and F. Csàki, (eds.), 2nd International Symposium on Information Theory, Budapest: Akademiai Kiádo, pp. 267-281. Reprinted in Kotz, S. Johnson N. L. (eds.) *Breakthroughs in Statistics, Volume I: Foundations and Basic Theory.* New York: Springer-Verlag.

Becker H. (1963). *Outsiders.* New York: The Free Press.

Betensky R.A., Lindsey J.C., Rayan L.M., Wand M.P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21, 263-275.

Blumstein A., Cohen J., Roth J. A., Visher C. A (1986). *Criminal Careers and "Career Criminals".* Panel on research on criminal careers, Commission on behavioural and social sciences and education, National Research Council, Washington D.C.: National Academy Press.

Clinard M.B. and Quinney R.(1967). *Criminal behavior systems: a typology.* New York: Holt, Reinhart and Winston.

Clinard M.B. and Quinney R.(1973). Corpoeate criminal behaviour. In *Criminal Behaviour System: a topology reviewed.* M. B. Clinard, R. Quinney, New York: Holt, Reinhart and Winston.

Dempster A. P., Laird N. M., Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* B, 39, 1-38.

D'Unger A. V., Lund K. C. , McCall P. L., Nagin D. S. (1998). How many latent classes of delinquent/criminal careers. Results from mixed poisson regression analysis. *American Journal of Sociology*, 103, 1593-630.

Fahrmeir L., Tutz G. (2001). *Multivariate Statistical Modelling Based on generalized Linear Models* (2nd Edition). New York: Stringer Verlag.

Farrington D. (1997). Human development and criminal careers. In Rod Morgan *et al.* (1997) *The Oxford Handbook of Criminology*(2nd); Oxford, Oxford University Press.

Farrington D., West D. (1990). The Cambridge Study in Delinquent Development: A Long-Term Follow-up Study of 411 London Males. In H. J. Kerner and G. Kaiser eds. *Criminality: Personality, Behavior and Life History.* Berlin: Springer-Verlag, pp. 115-138.

Fergusson D.M., Horwood L. J., Lloyd M. (1991). A latent class model of child offending. *Criminal Behaviour and Mental Health*, 1, 90-106.

Fergusson D. M., Horwood L. J., Nagin D. S. (2000). Offending trajectories in a New Zealand birth cohort. *Criminology* 38, 525-551.

Francis B., Crosland P. (2002). The Police National Computer and the Offenders Index: can they be combined for research purposes? Full report:
`http://www.homeoffice.gov.uk/rds/pdfs2/pncandoir170.pdf`
Home Office:London.

Francis B., Soothill K. and Fligelstone R. (2004). Identifying patterns of offending behaviour: A new approach to typologies of crime. Accepted for first (launch) issue of *European Journal of Criminology* (1) 47-82.

Gibbons, D.C. (1965) *Changing the lawbreaker: the treatment of delinquents and criminals.* Englewood Cliffs: Prentice-Hall Inc.

Gibbons D.C. (1972). *Society, crime and criminal careers.* Englewood Cliffs: Prentice Hall Inc.

Gibbons D.C. (1975). Offender typologies- two decades later. *British Journal of Criminology*, 15, (2), 140-156.

Hastie T. and Tibshirani R. (1990). *Generalised Additive Models.* Wiley: New York.

Hellerstein J. L., Ma S., Perng C. S. () Discovering actionable patterns in event data. *IBM system journal*, 41 (3), 475-493.

Lauritsen J. L. (1998). The age-crime debate: Assessing the limits of longitudinal self-report data *Soc. Forces*, 77 (1): 127-154.

Lombroso (1876). *L'Uomo delinquente.* Torino: Fratelli Bocca.

Matza D. (1964). *Delinquency and Drift*. New York: Wiley.

McHugh R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* 21, 331-347.

McHugh R. B. (1958). Note on "Efficient estimation and local identification in latent class analysis". *Psychometrika* 23, 273-274.

Nagin D.S., Farrington D.P., Moffitt T.E. (1995). Life-course trajectories of different types of offenders. *Criminology*, 33 (1) 111-139.

Nagin D.S., Land K.C. (1993). Age, Criminal Careers and population heterogeneity: Specification and estimation of a nonparametric mixed-Poisson model. *Criminology*, 31 327-362.

Prime J., White S., Liriano S., Patel K. (2001). *Criminal Careers of those born between 1953 and 1978*. Statistical Bulletin 4/01. London: Home Office.

Raftery A. E. (1995). Bayesian Model selection in social research. In P. V. Marsden (ed.), *Sociological Methodology, 1995*. Cambridge, MA: Blackwell.

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

Stander J., Farrington D., Hill G. and Altham P. (1989). Markov Chain Analysis and Specialisation in Criminal Careers. *British Journal of Criminology*, 29, 317-335.

Tibshirani R., Hstie T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.

Uebersax J.S. (1997). Analysis of student problem behaviors with latent trait, latent class, and related probit mixture models. In: Rost J, Langeheine R, eds. *Applications of Latent Trait and Latent Class Models in the Social Sciences*. New York, NY: Waxmann; pp. 188-195.

van der Heijden P., 't Hart H., Dessens J. (1997). A parametric bootstrap procedure to perform statistical tests in a LCA of anti-social behaviour. In: Rost J, Langeheine R, eds. *Applications of Latent Trait and Latent Class Models in the Social Sciences*. New York, NY: Waxmann; p.196-208.

Vermunt J. K., Magidson J. (2000). *Latent Gold User's Guide*. Belmont: Statistical Innovations Inc.

Vermunt J. K., Magidson J. (2002). Latent Class Cluster Analysis. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 89-106. Cambridge University Press.

Wolfgang M., Figlio R. and Sellin T. (1972). *Delinquency in a birth cohort.* Chicago: University of Chicago Press.

Wu L. L., Tuma N. B. (1990). Local Hazard Models. *Sociological Methods*, 20, 141-180.

Bibliography

Abbot A., Barman E. (1997). Sequence comparison via Alignement and Gibbs sampling. *Sociological Methodology*, 27, 47-87.

Aitkin M., Anderson D., Hinde J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society*, A, 144, 419-448.

Akaike H. (1973). Information theory and an extention of the maximum likelihood principle. In B. N. Petrov and F. Csàki, (eds.), 2nd International Symposium on Information Theory, Budapest: Akademiai Kiádo, pp. 267-281. Reprinted in Kotz, S. Johnson N. L. (eds.) *Breakthroughs in Statistics, Volume I: Foundations and Basic Theory.* New York: Springer-Verlag.

Anderson T. W. (1958). *An introduction to multivariate statistical analysis.* Wiley, New York.

Anderson T.W., Rubin H. (1956). Statistical Inference in Factor Analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5, 111-150. Univ. California Press, Berkeley.

Andersson S., Madigan D., Perlman M., Triggs C. (1995) On the relations between conditional independence models determined by finite distributive lattices and by directed acyclic graphs. *Journal of Statistical Planning and Inference*, 48, 25-46.

Banfield J. D., Raftery A. E. (1993). Model-based Gaussian and non-Gaussian Clustering. *Bionetrics*, 49, 803-821.

Bartholomew D. J. (1995). Spearman and the origin and develoment of factor analysis. *British Journal of Statistical and Mathematical Psychology*, 48, 211-220.

Bartholomew D. J. (1987). *Latent variables models and factor analysis.* London: Charles Griffin.

Bartholomew D. J., Knott M. (1999). *Latent variables models and factor analysis.* Kendall's Library of Statistics, London: Arnold.

Banerjee M., Richardson T. (2003): On a Dualization of Graphical Gaussian Models: A Correction Note. *Scandinavian Journal of Statistics*, 30, 4, 817-821.

Baum L.E., Petrie T. (1966). Statistical inference for probabilistic functions finite state Markov chains. *Annals of Mathematical Statistics*, 37, 1554-1563.

Baum L. E., Petrie T., Saules G., Weiss N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41 (1), 164-171.

Becker H. (1963). *Outsiders.* New York: The Free Press.

Betensky R.A., Lindsey J.C., Rayan L.M., Wand M.P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21, 263-275.

Biernacki C, Celeux G., Govaert G. (1998). Assessing a mixture model for clustering with the intergrated classification likelihood. *Technical Report No. 3521.* Rhône-Alpes INRIA.

Blalock H. (1971). *Causal Models in the Social Science.* Adline-Atherton, Chicago.

Blalock H. (1961). *Causal Inferences in Nonexperimental Research.* Chapel Hill: University of Noth Carolina Press.

Blenter P. M. (1980). Multivariate analysis with latent variables: causal modeling. *Annual Review of Psychology*, 31, 419-456.

Blimes J. (2002). What HMMs Can Do. *Technical Report*, Department of Electrical Engineering, University of Washington.

Blumstein A., Cohen J., Roth J. A., Visher C. A (1986). *Criminal Careers and "Career Criminals".* Panel on research on criminal careers, Commission on behavioural and social sciences and education, National Research Council, Washington D.C.: National Academy Press.

Bollen K. A. (1989). *Structural equations with latent variables.* Wiley, New York.

Bollen K. A (2002). Latent Variables in Psychology and the Social Science. *Annual Review of Psychology*, 53, 605-634.

Borsboom D., Mellenbergh G. J., van Heerden J. The theoretical status of latent variables. *Psychological review*, vol. 110, No. 2, 203-219.

Capitanio A., Azzalini A., Stanghellini E. (2003). Graphical Models for skew-normal variates. *Scandinavian Journal of Statistics*, Vol. 30, 129-144.

Castelo J. R. V. (2002). *The Discrete Acyclic Digraph Markov Model in Data Mining*. PhD Thesis, Utrecht University.

Cleux B, Biernacki C., Govaert G. (1997). *Choosing Models in model based clustering and discriminant analysis*. Technical Report. Rhone-Alpes: INRIA.

Cliff N. (1983). Some cautions concerning the application of causal modelling methods. *Multivariate Behavioral Research* 18, 115-126.

Clinard M.B. and Quinney R.(1967). *Criminal behavior systems: a typology*. New York: Holt, Reinhart and Winston.

Clinard M.B. and Quinney R.(1973). Corpoeate criminal behaviour. In *Criminal Behaviour System: a topology reviewed.* M. B. Clinard, R. Quinney, New York: Holt, Reinhart and Winston.

Clogg C. C. (1981). Latent structure models of mobility. *American Journal of sociology*, 86, 836-868.

Clogg C. C. (1988). Latent class models for measuring. In *Latent Trait and Latent Class Models*. R. Langeheine and J. Rost (eds), 173-206. New-York: Plenum Press.

Collins M. L., Wugalter S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioural Research*, 27, 131-157.

Collins M. L., Flaherty B. P. (2002). Latent class models for longitudinal data. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 287-303. Cambridge University Press.

Cox D. R. (2003). Conditional and marginal association for binary random variables. *Biometrika* 90, 4, 982-984.

Cox D. R., Wermuth N. (1993). Linear dependencies Represented by Chain Graphs. *Statistical Science*, 8, No. 3, 204-283.

Cox D. R., Wermuth N. (1996). *Multivariate Dependencies - Models, Analysis and interpretation.* London: Chapman & Hall.

Cox D. R., Wermuth N. (1999). Likelihood factorizations for mixed discrete and continuous variables. *Scandinavian Journal of Statistics*, 26, 209-220.

Cox D. R., Wermuth N. (2003). A general condition for avoiding effect reversal after marginalization. *J. R. Statist. Soc. B* 4, 937-941.

Cox D. R., Wermuth N. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society*, B, to appear.

Dayton C. M. (1999). *Latent class scaling analysis.* Sage Publications.

Dawid A. P. (1979). Conditional independence in statistical theory (with discussion). *Jour. Royal Stat. Society Ser. B*, 41, 1-31.

Dawid A. P. (2002). Influence diagrams for Causal Modelling and Inference. *International statistical review* 70, 2, 161-189.

D'Unger A. V., Lund K. C. , McCall P. L., Nagin D. S. (1998). How many latent classes of delinquent/criminal careers. Results from mixed poisson regression analysis. *American Journal of Sociology*, 103, 1593-630.

de Menezes L. M. (1999). On fitting latent class models for binary data. *British Journal of Mathematical and Statistical Psychology* 52.

Dempster A.P. (1969). *Elements of Continuous Multivariate Analysis.* Reading: Addison Wesley.

Dempster A. P. (1972). Covariance Selection. *Biometrics*, 28, 157-175.

Dempster A. P. Laird N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statististical Society* B, 39, 1-38.

Drton M., Richardson T. S. (2003).A New Algorithm for Maximum Likelihood Estimation in Gaussian Graphical Models for Marginal Independence. *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 184-91.

Duncan O. D. (1969). Some linear models for two-wave, two variable panel analysis. *Psychological Bulletin*, 72, 177-182.

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

Efron B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron B., Gong G. (1983). A leisure look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37, 36-48.

Fahrmeir L., Tutz G. (2001). *Multivariate Statistical Modelling Based on generalized Linear Models* (2nd Edition). New York: Stringer Verlag.

Farrington D. (1997). Human development and criminal careers. In Rod Morgan *et al.* (1997) *The Oxford Handbook of Criminology*(2nd); Oxford, Oxford University Press.

Farrington D., West D. (1990). The Cambridge Study in Delinquent Development: A Long-Term Follow-up Study of 411 London Males. In H. J. Kerner and G. Kaiser eds. *Criminality: Personality, Behavior and Life History.* Berlin: Springer-Verlag, pp. 115-138.

Fergusson D.M., Horwood L. J., Lloyd M. (1991). A latent class model of child offending. *Criminal Behaviour and Mental Health*, 1, 90-106.

Fergusson D. M., Horwood L. J., Nagin D. S. (2000). Offending trajectories in a New Zealand birth cohort. *Criminology* 38, 525-551.

Formann A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology* 38, 87-111.

Formann A. K. (2003). Latent Class Diagnosis from a Frequentist Point of view. *Biometrics*, 59, 189-196.

Francis B., Crosland P. (2002). The Police National Computer and the Offenders Index: can they be combined for research purposes? Full report:
`http://www.homeoffice.gov.uk/rds/pdfs2/pncandoir170.pdf`
Home Office:London.

Francis B., Soothill K. and Fligelstone R. (2004). Identifying patterns of offending behaviour: A new approach to typologies of crime. Accepted for first (launch) issue of *European Journal of Criminology* (1) 47-82.

Frydenberg M., Lauritzen S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, 76, 539-555.

Frydenberg M. (1990). The chain graph Markov property. *Scand. J. Statist.*17, 333-353.

Garnett J. C. (1919). General ability, cleverness and purpose. *British Journal of Psychiatry*, 8, 345-366.

Geiger J. C. (1998). Graphical Models and Exponential Families. *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, 156-165.

Gibbons, D.C. (1965) *Changing the lawbreaker: the treatment of delinquents and criminals.* Englewood Cliffs: Prentice-Hall Inc.

Gibbons D.C. (1972). *Society, crime and criminal careers.* Englewood Cliffs: Prentice Hall Inc.

Gibbons D.C. (1975). Offender typologies- two decades later. *British Journal of Criminology*, 15, (2), 140-156.

Gibson W. A. (1955). An extension of Anderson's solution for the latent structure equations. *Psychometrika* 20, 69-73.

Goldberger A. S. (1964). *Econometric Theory.* New York: Wiley.

Goldberger A. (1972). Structural equation methods in the social sciences. *Econometrica* 40, 979-1001.

Goldberger A., Ducan O. (1973). Structural equation models in the social sciences (Seminar Press, New York).

Goodmann L. A. (1974a). Explanatory Latent Structure Models Using both identifiable and unidentifiable Models. *Biometrika* 61, 315-331.

Goodman L. A. (1974b). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: a modified latent structure approach. *American journal of sociology* 79, 1179-1259.

Goodman L. A. (1978). *Analyzing qualitative/Categorical Data* (ed. J. Magidson), Abt Books, Cambridge, MA.

Goodman L. A. (2002). Latent Class Analysis. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 3-55. Cambridge University Press.

Greenland S. (2000). Causal analysis in the health sciences. *Journal of the American Statistical Association*, 95, 286-289.

Greenland S., Pearl J., Robins JM. (1999). Causal diagrams for epidemiologic research. *Epidemiology* 10, 37-48.

Guttmann L. (1955). The determinacy of factor score matrices with implication of five other basic problems of common-factor theory. *Journal of Mathematical and Statistical Psychology* 8, 65-81.

Guttmann L. (1977). What is not what in statistics. *The Statistician* 26, 81-107

Guttorp P. (1995). *Stochastic Modeling of Scientific Data* . Chapman and Hall, 384 pp.

Haagen K. (1992). Il problema dell'indeterminatezza nei modelli con variabili latenti. *Statistica* 3, 365-377.

Haavelmo T. (1943). The statistical implications of a system of simultaneous equations. *Econometrika* 11, 1-12.

Haberman S. J. (1974). Loglinear models for frequency tables derived by undirected observation. Maximum-likelihood equations. *Annals of Statistics* 2, 911-924.

Haberman S. J. (1976). Iterative scaling procedures for loglinear models for frequency data derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association, 1975*, 45-50.

Haberman S. J. (1979). *Analysis of Qualitative Data*. Vol.2. New York: Academic Press.

Hagenaars J. A. (1990). *Categorical Longitudinal Data; Loglinear Panel, Trend and Cohort Analysis*. Newbury Park, CA: Sage.

Hansen W. B., Graham J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: peer pressure resistance training versus establishing conservative norms. *Preventive Medicine*, 20, 414-430.

Hastie T. and Tibshirani R. (1990). *Generalised Additive Models*. Wiley: New York.

Heinen T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Advanced quantitative Techniques in the Social Sciences, Sage Publications, Thousand Oaks, CA.

Hellerstein J. L., Ma S., Perng C. S. () Discovering actionable patterns in event data. *IBM system journal*, 41 (3), 475-493.

Heywood H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Statistical Society, London*, 134, 486-501.

Hodge R.W., Treiman, D.J. (1968). Social participation and social status. *American Sociological Review*, 33, 723-740.

Holland P. W. (1988). Causal inference, path analysis, and recursive structural equation models. In C. C. Clogg, editor, *Sociological Methodology*, 449-493. American Sociological Association, Washington.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417-441, 498-520.

Joffe M.M., Byrne C., Colditz G. A. (2001). Postmenopausal hormone use, screening, and breast cancer: characterization and Control of a Bias. *Epidemiology*, 12, 4, 429-438.

Jöreskog K. G. (1973). A general method for estimating a linear structural equation system. In *Structural Equation Model in the Social Science*(A. S. Goldgerger and O. D. Duncan, eds.) 85-112. Seminar Press, New York.

Jöreskog K. G. (1977). Structural equation models in the social sciences: specification, estimation and testing. In P. R. Kirshnaiah (Eds.) *Applications of statistics*, 265-286, Amsterdam, North Holland.

Jöreskog K. G. (1981). Analysis of covariance structures. With discussion. *Scandinavian Journal of Statistics*, 8, 65-92.

Jöreskog, K. G., Goldberger (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. . *Journal of the American Statistical Association*, 10, 631-639.

Jöreskog K., G. Sorbom D. (1989). *LISREL 7 - A guide to the Program and Applications*. 2nd Edition, SPSS Publications, Chicago.

Kao E. P. C. (1997). *An introduction to stochastic process*. Duxbury Press, 438 pp.

Kauermann G. (1996). On a dualization of graphical Gaussian models. *Scandinavian Journal of Statistics*, 23, 105-116.

Kiiveri H. T. (1982). A unified approach to causal models. Ph.D thesis. University of Western Australia.

Kiiveri H. T., Speed T. P. (1982). Structural analysis of multivariate data: a review. In *Sociological Methodology* (S. Leinhardt, ed.) Jossey-Bass, San Francisco.

Kiiveri H. T., Speed T. P., Carlin J. B. (1984). Recursive causal models. *J. Austral. Math. Soc. Ser.* A 30-52

Kiiveri H. T. (1987). An incomplete data approach to the analysis of covariance structure. *Psychometrika*, 52, 539-554.

Kollmann T., Formann A. K. (1997). Using latent class models to analyse response patters in epidemiologic mail survey. On *Applications of Latent Trait and Latent Class Models in the Social Sciences* J. Rost, R. Langeheine (eds.), 345-351. Münster, Germany: Waxmann.

Knuiman M. (1978). *Covariance selection.* In R.L. Tweedie (Ed.) Proceedings of the conference on Spatial Patterns and Processes. Supplement to Advances in Applied Probability, 10, 123-130.

Koster J.T.A (1996). Markov properties of non recursive causal models. *Ann. Statist.* 24, 2148-2177.

Koster J.T.A. (1999). On the validity of Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand. J. Statist.*, 26, 413-431.

Koster J.T.A. (2002). Marginalizing and conditioning on graphical models. *Bernoulli*, 8, 817-840.

Land K. C. (1973). Identification, parameter estimation and hypothesis testing in recursive sociological models. In A. S. Goldberger and O. D. Duncan (Eds.) Structural equation models in the social science. Seminar Press, New York.

Lang J. B. (1992). Obtaining the observed information matrix for the poisson loglinear model when the EM-algorithm is used. *Biometrika*, 79, 405-407.

Lange K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society*, B, 57, 425-437.

Lange K. (1995b). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5, 1-18.

Langeheine R. (1994). Latent variable Markov models. In A. von Eye and C.C. Clogg (eds.) *Latent Variable Analysis. Application for Developmental Research.* Thousand Oaks, CA: Sage, pp 373-395.

Langeheine R., Pannekoek J., van de Pol F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociologiacal Methods and Research* 24(4), 492-516.

Langeheine R., van de Pol F. (2002). Latent Markov Chains. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 304-341. Cambridge University Press.

Lauritsen J. L. (1998). The age-crime debate: Assessing the limits of longitudinal self-report data *Soc. Forces*, 77 (1): 127-154.

Lauritzen S. L. (1996). *Graphical Models.* Oxford: Clarendon Press.

Lauritzen S. L. (2002). gRaphical models in R, *R News*, 3(2)39.

Lauritzen S., Dawid A., Larsen B., Leimer H. (1990). Independence properties of directed Markov fields. *Networks*, 20, 491-505.

Lauritzen S. L., Wermuth N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17, 31-57.

Lazarsfeld P. F. (1950). The logical and mathematical foundation of latent structure analysis. In *Studies in Social Psychology in World War II* Volume IV: *Measurement and Prediction*, S. A. Stouffer, L. Guttman, E. A. Suchman (eds) 362-412. New York: Princeton University Press.

Lazarsfeld P. F., Henry N. W (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

Lederman W. (1938). The orthogonal transformations of a factorial matrix into itself. *Psychometrika*, 3, 181-187.

Lin T. H., and Dayton C. M. (1997). Model selection information criteria for non-nested latent class models. *Journalof Educational and Behavioral Statistics*, 22, 249-264.

Lindgren G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics: Theory and Applications*, 5, 81-91.

Lombroso (1876). *L'Uomo delinquente*. Torino: Fratelli Bocca.

Louis T. A. (1982). Finding the observed information matrix when using the EM-algorithm. *Journal of the Royal Statistical Society, B, 44, 226-233*.

MacDonald I., Zucchini W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.

Marchetti G. M., Drton M. (2003). `ggm`: an `R` pachakage for Gaussian graphical models, URL: *http://cran.r-project.org/*.

Matza D. (1964). *Delinquency and Drift*. New York: Wiley.

McCutheon A. L. (2002). Basic concepts and procedures in single and multiple group latent class analysis. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 56-85. Cambridge University Press.

McHugh R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* 21, 331-347.

McHugh R. B. (1958). Note on "Efficient estimation and local identification in latent class analysis". *Psychometrika* 23, 273-274.

McLachlan G. J., Basford K. E. (1988). *Mixture Models: Inference and Application to Clustering* . New York: Marcel Dekker.

McLachlan G. J., Peel D. (2000). *Finite Mixture Models*. Wiley Series: New York.

Moran P.A.P (1961). *Path coefficient s reconsidered*. Australian Journal of Statistics, 3, 87-93.

Nagin D.S., Farrington D.P., Moffitt T.E. (1995). Life-course trajectories of different types of offenders. *Criminology*, 33 (1) 111-139.

Nagin D.S., Land K.C. (1993). Age, Criminal Careers and population heterogeneity: Specification and estimation of a nonparametric mixed-Poisson model. *Criminology*, 31 327-362.

Noreen E. W. (1989). *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York: Wiley.

Orchard T., Woodbury M. A. (1972). A Missing Information Principle: Theory and Applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, (vol.1), 697-715.

Patterson S. (1984). *Journal of research in crime and deliquency*, Vol. 21 No. 4, 333-352.

Pearl J. (1988). *Probabilistic reasoning in Intelligent Systems*. S. Mateo, CA: Morgan Kaufmann.

Pearl J. (2000). *Causality*. New York: Oxford.

Pearl J., Paz A. (1987). Graphoids: A graph-based logic for reasoning about relevancy relations. In Boulay, B. D. editor, *Advanced in Artificial Intelligence-II*. North Holland.

Pearl J., Verma T. (1987). The logic of representing dependencies by directed graphs. In *Proc. of the conf. of the American Association of Artificial Intelligence*, 347-379.

Peirce C. S. (1884). The numerical measure of the success of predictions. *Science* 4, 453-454.

Prime J., White S., Liriano S., Patel K. (2001). *Criminal Careers of those born between 1953 and 1978*. Statistical Bulletin 4/01. London: Home Office.

Rao C. R. (1973). *Linear statistical inference and its applications*. Wiley, New York.

Raftery A. E. (1995). Bayesian Model selection in social research. In P. V. Marsden (ed.), *Sociological Methodology, 1995*. Cambridge, MA: Blackwell.

Richardson T., Spirtes P. (1999). Automated discovery of linear feedback models. In *Computation, causation, and Discovery,*(ed. C. Glymour and G. F. Cooper), pp. 253-304. MIT Press, Cambridge, MA.

Richardson T., Spirtes P. (2002). Ancestral Markov graphical models. *Annals of Statistics*, 30, 962-1030.

Rothenberg T. (1971). Identification in parametric models. *Econometrika*, 39, 577-591.

Rubin D. B. (1982). *Introduction: Research on Test Equating Sponsored by Educational Testing Service, 1978 1980*. Test Equating, New York: Academic Press, Inc. pp. 1 6. (With P.W. Holland).

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

Sewell W. H., Haller A. O., and Ohlendorf G. W. (1970). The educational and early occupational status attainment process: revisions and replications. *American Sociological Review*, 35, 1014-1027.

Skrondal A., Rabe-Hesketh S. (2004). *Generalized latent variable modelling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman and Hall, CRC Press.

Sobel M. E. (1995). Causal inference in the social and behavioral sciences. In Arminger A., Clogg C. C., Sobel E. M., editors, *Handbook of statistical modelling for the social and behavioral sciences* 1-38. Plenum Press, New York.

Speed T. P. (1978). *Relations between models for spatail data, contingency tables and Markov fields on graphs*. In R. L. Tweedie (Ed.), Prodeedings of the Conference on Spatial Patterns and Processes. Supplument to Advances in Applied Probability 10, 111-122.

Speed T. P., Kiiveri H. T. (1986). Gaussian Markov distribution over finite graphs. *Annals of Statistics* 14, 138-150.

Spearman C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201-293.

Spirtes P. (1995). Directed cyclic graphical representations of feedback models. In *Uncertainty in Artificial Intelligence* (P. Besnard and S. Hanks, eds.) 491-498. Morgan Kaufmann, San Francisco.

Spirtes P., Glymour C., Scheines R. (1993). *Causation, Prediction and Search*, New York: Springer-Verlag.

Spirtes P., Glymour C., Scheines R. (2000). *Causation, Prediction and Search*,2nd ed. Cambridge, MA: MIT Press.

Spirtes P., Richardson T., Meek C., Scheines R., Glymour C. (1998). Using path diagrams as structural equation modelling tool. *Sociological Methods and Research* 27, 182-225.

Stander J., Farrington D., Hill G. and Altham P. (1989). Markov Chain Analysis and Specialisation in Criminal Careers. *British Journal of Criminology*, 29, 317-335.

Stanghellini E. (1997). Identification of single-factor model using graphical Gaussian rules. *Biometrika*, 84, 241-244.

Stanghellini E., Wermuth N. (2004). On the identification of path analysis models with one hidden variable. *Under revision.*

Steiger J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, 44, 157-167.

Suppes P., Zanotti M. (1981). When are probabilistic explanations possible? *Sythese*, 148, 191-199.

Tanner M.A. (1996). *Tools for statistical inference.* New York: Springer.

Tibshirani R., Hstie T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.

Tinbergen J. (1937). *An Econometric Approach to Business Cycle Problems.* Hermann, Paris.

Titterington D. M, Smith A. F. M., Markov U. E. (1985). *Statistical analysis of Finite Mixture Distributions.* New York: John Wiley.

Thurstone L.L. (1938). *Primary mental abilities.* Chicago: University Press.

Uebersax J.S. (1997). Analysis of student problem behaviors with latent trait, latent class, and related probit mixture models. In: Rost J, Langeheine R, eds. *Applications of Latent Trait and Latent Class Models in the Social Sciences.* New York, NY: Waxmann; pp. 188-195.

van der Heijden P., 't Hart H., Dessens J. (1997). A parametric bootstrap procedure to perform statistical tests in a LCA of anti-social behaviour. In: Rost J, Langeheine R, eds. *Applications of Latent Trait and Latent Class Models in the Social Sciences.* New York, NY: Waxmann; p.196-208.

Vermunt J. K., Magidson J. (2000). *Latent Gold User's Guide.* Belmont: Statistical Innovations Inc.

Vermunt J. K., Magidson J. (2002). Latent Class Cluster Analysis. In *Applied Latent Class Analysis*, Hagenaars J. A., McCutCheon A. L., pp 89-106. Cambridge University Press.

Vicard P. (2000). On the identification of a single-factor model with correlated residuals. *Biometrika*, 87, 199-205.

Von Eye A., Clogg C. C. (1994). *Latent variable analysis: Applications for developmental research*. Thousand Oaks, CA: Sage.

Wermuth N. (1976a). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32, 95-108.

Wermuth N. (1976b). Model search among multiplicative models. *Biometrics*, 32, 253-263.

Wermuth N. (1980). Linear recursive equations, covariance selection, and path analysis. *Amer. Statist. Assoc.*,75, 963-972.

Wermuth N. (1999). Effects of an unobserved confounder on a system with an intermediate outcome. ZUMA-Arbeitsbericht 99/07.

Wermuth N. (2003). Analysing social science data with graphical Markov models. In *Highly Structured Stochastic Systems*, P. Green, N. Hjort, Richardson S. (eds.), pp. 45-52. Oxford University Press.

Wermuth N., Cox D. R. (1998). On association models defined over independence graphs. *Bernoulli* 4, 477-496.

Wermuth N., Cox D. R. (2003). On modified triangular systems. *http://psystat.sowi.uni-mainz.de/wermuth/pdfs/papmodtri.pdf*

Wermuth N., Cox D. R. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society*, B, to appear.

Wermuth N., Lauritzen S. (1990). On substantive research hypothesis, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society, Series, B*, 52, 21-50.

Whittaker J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Wiggins L. M (1973). *Panel Analysis: Latent Probability Models for Attitudes and Behavior Processes*. Amsterdam: Elsevier.

Wiley D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger and O. D. Duncan, eds., *Structural Equation Models in the Social Sciences.* New York: Academic Press, pp. 69-83.

Wilson E. B. (1928). On hierarchical correlation systems. *Proceedings, National Academy of Science*, 14, 283-291.

Wolfgang M., Figlio R. and Sellin T. (1972). *Delinquency in a birth cohort.* Chicago: University of Chicago Press.

Wolfe J. H. (1970). Pattern clustering by multivariate cluster analysis. *Multivariate Behavioural research*, 5, 329-350.

Wright S. (1921). Correlation and Causation. *Journal of Agricultural Research*, 20, 557-585.

Wright S. (1923). The theory of path coefficients: A reply to Niels'criticism. *Genetics* 8, 239-255.

Wright S. (1934). The method of path coefficients. *Annals of Statistics*, 5, 161-215.

Wu C. F. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11, 95-103.

Wu L. L., Tuma N. B. (1990). Local Hazard Models. *Sociological Methods*, 20, 141-180.

Zellner A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* 57, 348-368.