Graphical Models for Event History Analysis based on Local Independence

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund
vorgelegt von
Vanessa Didelez

Dortmund 2000

Gutachter:

Prof. Dr. I. Pigeot
Prof. Dr. U. Gather
Prof. Dr. S. Schach

Tag der mündlichen Prüfung:
15. Dezember 2000

To my mother

The order that our mind imagines is like a net, or like a ladder, built to attain something. But afterwards you must throw the ladder away, because you discover that, even if it was useful, it was meaningless.

Umberto Eco

Contents

Introduction **9**

1 Graphs **13**

1.1 Notation and terminology 13

1.2 Graph separation 20

2 Graphical models for random structures in time **33**

2.1 Conditional independence graphs 34

2.1.1 Conditional independence models 40

2.1.2 Dynamic interaction models 46

2.1.3 Partial correlation graphs for time series 51

2.2 Causality graphs 54

3 Local independence **57**

3.1 Basic concepts of event history analysis 58

3.2 Local independence 65

3.2.1 Local independence as irrelevance relation 69

3.2.2 Local independence and multi–state processes 75

3.2.3 Local independence for Markov processes 78

4 Local independence graphs **87**

4.1 Dynamic Markov properties 88

4.2 Factorization of the likelihood 95

4.3 Collapsibility 102

4.3.1 Weak collapsibility 103

4.3.2 Strong collapsibility 106

4.4 Local independence graphs for stopped processes . . . 107
4.5 Causal reasoning in local independence graphs . . . 111
4.5.1 Causal graphs . . . 112
4.5.2 Intervention graphs and identifiability . . . 115
4.5.3 Open questions . . . 117

5 Statistical inference for local independence graphs 123
5.1 Preliminaries . . . 124
5.2 Models for aggregated counting processes . . . 127
5.2.1 Maximum likelihood estimation . . . 129
5.2.2 LR tests for composable Markov Processes . . . 130
5.2.3 Nonparametric inference . . . 132
5.3 Regression models . . . 139
5.3.1 Maximum likelihood estimation . . . 142
5.3.2 Semiparametric estimation for multiplicative hazards . . . 146

Discussion 151

A Conditional expectation and conditional independence 155
A.1 Conditional expectation . . . 155
A.2 Conditional independence . . . 159

B Stochastic processes 163
B.1 Basic notions . . . 163
B.2 The Doob–Meyer decomposition . . . 167

C Notations 171

Bibliography 177

Acknowledgements 183

Introduction

The statistical modeling and analysis of multivariate data typically deals with complex association structures due to various direct and indirect relations among the variables. If the data structure comprises a time dimension, as for instance in event histories, the additional dynamic character of the associations has to be taken into account. A typical example for this kind of data is given by clinical studies, where the time up to a specific event is investigated, e.g. recovery, relapse, or death. A deeper understanding of the underlying physiological mechanisms mostly requires the consideration of additional time constant and time varying variables such as onset of specific side effects, medicamentation during the observation period, and other physiological indicators. In this situation, it is not sufficient to examine associations among the processes since the *direction* of these associations in time often plays an important role. A specific harmful side effect might for instance be prevented by a suitable medicamentation, but this could on the long term have a negative effect on survival because the drug possibly affects not only the side effect but indirectly also the whole subsequent development of the system. Therefore, models and methods that permit a careful analysis of these dynamic associations are called for. Obviously, such models should meet the demand of complexity and still facilitate a clear interpretation.

For general multivariate data, graphical models have been developed to satisfy the above requirements. The idea of graphical models is to represent the dependence structure of a multivariate random vector by a graph, where the vertices correspond to the variables, and the absence of edges between vertices stands for conditional or marginal independencies. As a framework for analyzing dependence structures in multivariate data sets, graphical models mainly offer three features. (1) The

graphical representation allows a direct and intuitive understanding of the possibly complex underlying dependence structure; (2) qualitative information about conditional independencies in the underlying statistical model is given a precise representation and can thus unambiguously be read off the graph; and (3) the structure of the graph yields direct information about various aspects related to the statistical analysis, e.g. collapsibility or decomposability of estimating procedures.

However, a direct application of the present theory of graphical models to event history analysis is not satisfactory: The inherent dynamic character of the associations cannot be taken into account, except under restrictive assumptions. This problem is mainly due to the convention that vertices represent variables and to the concept of conditional independence that does not include the time dimension. Therefore, we choose to base the graphical representation of dependencies in event history data on counting processes as the underlying random structure and on a dynamic association concept called *local independence*. The idea of local independence is quite simple. Consider for instance the question of how the two events of finding a job and giving birth to a child are related in a female population. One could formulate the following (maybe unrealistic) hypotheses: (1) The event of finding a job is more probable if a woman has no child than if she has, whereas (2) the event of getting a child is as probable if she has a job as if she has no job. In the described situation we say that the event of finding a job *locally depends* on the event of giving birth to a child because it depends on whether and when the latter has occurred in the past. In contrast, getting a child is *locally independent* of finding a job because it is irrelevant whether the latter has occurred in the past. The independence is called *local* since the time when the job is found and the time of birth are of course stochastically dependent due to the first hypothesis. This example already illustrates the main features of the concept of local independence as compared to conditional independence:

- It is a dynamic concept in the sense that it explicitly formulates dependencies of a present random structure on a past development.
- It is an asymmetric concept, i.e. if an event A is locally independent of B then it is not necessarily implied that B is locally independent of A.
- It encompasses more information than captured by conditional independence

structures among the corresponding times of events.

- Similar to conditional independence, local independence is a conditional concept since it always involves conditioning on the past.

The present thesis explores models defined through local independence structures induced by suitable graphs and investigates the properties of the resulting local independence graphs with particular respect to their interpretation and their implications for statistical inference. For this purpose, we proceed as follows.

The first chapter provides the necessary notation and background on graph theory. Since local independence graphs represent asymmetric dependencies the usual graph terminology has to be generalized to apply for instance to cyclic graphs. In particular, we introduce a new asymmetric graph separation which will turn out to capture precisely those separations induced by local independence structures. The properties of this separation are discussed along the lines of an axiomatic framework developed for irrelevance relations, the so–called graphoid axioms. These, too, have to be generalized to cope with asymmetric relations.
In Chapter 2, different approaches to the application of conditional independence graphs to event history data are addressed. This mainly serves to give a first idea of the specific difficulties one has to face when modeling random structures in time. Two of the presented models are already based on the idea of representing the components of multivariate time series as the vertices in the graph. The same principle underlies local independence graphs.
The notion of local independence, which is fundamental to the subsequent models, is introduced in Chapter 3. For the general definition we adopt the framework of marked point processes. In analogy to graph separation, the discussion of the properties of local independence refers to the graphoid axioms. As a special case of marked point processes we consider local independence for composable Markov processes.
Chapter 4 is the core of this thesis. Here, local independence graphs are defined, and it is shown which information on local independencies can be read off the graph. This includes statements about subprocesses, i.e. about the local independence structure resulting from discarding information on the occurrence of some of the events.

To this end, we define, in analogy to conditional independence graphs, the so–called *pairwise, local, and global dynamic Markov properties* and show their equivalence, where the asymmetric graph separation plays a key role. Further insight into the relation between local independence and conditional independence properties is gained by investigating the factorization property of the likelihood which is very similar to the well–known factorization for independence models on directed acyclic graph. The equivalence of the Markov properties is also used to derive conditions for collapsibility which is central to any complexity reduction in statistical models. Another important aspect of local independence graphs is their potential causal interpretation. It should be clear that they are a–priori no causal models. However, the graph semantics are useful for causal reasoning as will briefly be addressed in the last section.

The final chapter gives an overview over the application of different inference procedures to local independence graphs. These include non– and semiparametric approaches as well as likelihood based inference. The non–parametric models, however, are limited to the case of Markov processes. More complicated forms of dependencies among the events may instead be modeled by (semi)parametric regression models. For all these standard models, statistical tests on local independence are derived.

The thesis is concluded by a discussion of the results as well as some open questions regarding the application and generalization of local independence graphs.

Chapter 1

Graphs

We begin by establishing the terminology and conventions concerning graphs that will be used throughout the thesis. A very general introduction to the terminology is presented in Section 1.1. Special attention is given to the notion of separation in Section 1.2 where a new kind of separation is defined, δ–separation. This is suitable for cyclic graphs and has the unusual property of being asymmetric. Due to the asymmetry, δ–separation does not meet the usual axioms for independence relations so that we discuss its properties in terms of an asymmetric irrelevance relation.

1.1 Notation and terminology

The approach to apply graphical models in event history analysis presented later requires the definition of special graphs. We therefore present the required notation and some results from graph theory in necessary generality. The following definition of a graph not only allows for cycles but also for more than one edge between two vertices and thus for bidirected edges.

Definition 1.1.1 *Graph / edges / subgraph*
A *graph* is a pair $G = (V, E)$, where $V = \{1, \ldots, K\}$ is a finite set of vertices and E is a set of directed and/or undirected edges. *Undirected edges* are members of the class $E^u(V) = \{\{j, k\} | j, k \in V, j \neq k\}$. *Directed edges* belong to $E^d(V) = \{(j, k) | j, k \in V, j \neq k\}$. Thus, $E \subset E^u(V) \cup E^d(V)$.
For $A \subset V$ the *induced subgraph* G_A is defined as (A, E_A) with $E_A = E \cap (E^d(A) \cup E^u(A))$. //

In the visualization of a graph undirected edges $\{j,k\}$ are represented by lines, j — k, and directed edges (j,k) by arrows, $j \longrightarrow k$. If $(j,k) \in E$ and $(k,j) \in E$ this is shown as bidirected arrow, $j \longleftrightarrow k$. The meaning of the different edges depends on the specific statistical model represented by the graph as will become clear later. As mentioned above, the definition allows for the possibility that for two vertices j and k each of $\{j,k\}$, (j,k), and (k,j) are edges in the graph. This is unusual for most of the common graphs and we therefore mainly consider graphs, where $\{j,k\} \in E$ prevents $(j,k) \in E$ or $(k,j) \in E$ and vice versa. The definition of such graphs given below requires first the notions of paths and cycles (cf. Koster, 1996, Anderson et al., 2001).

Definition 1.1.2 *Paths / trails / cycles*
Consider a graph $G = (V, E)$.

1. An ordered $(n+1)$–tuple $(j_0, \ldots, j_n)$ of distinct vertices (except possibly $j_0 = j_n$) is called a *path of length n from j_0 to j_n* if $\{j_{i-1}, j_i\} \in E$ or $(j_{i-1}, j_i) \in E$ for all $i = 1, \ldots, n$.
   It is called *undirected* if $\{j_{i-1}, j_i\} \in E \ \forall\, i = 1, \ldots, n$.
   It is called *semidirected* if $\exists\, i : \{j_{i-1}, j_i\} \notin E$, and *directed* if the latter holds for all $i = 1, \ldots, n$.
2. A path of length n with $j_0 = j_n$ is called a *cycle*. Directed, semidirected, and undirected cycles are defined as obvious.
3. Let $\upsilon(j) = \{k \in V | \exists$ undirected path from j to k or $k = j\}$. Then, $\Upsilon(G) = \{\upsilon(j) | j \in V\}$ is the set of *(undirected) path components* of G.
4. Let $\xi(j) = \{k \in V | \exists$ path from j to k and from k to j or $k = j\}$. Then, $\Xi(G) = \{\xi(j) | j \in V\}$ is the set of *cycle components* of G.
5. A subgraph $\pi = (V', E')$ of G with $V' = \{j_0, \ldots, j_n\}$ and $E' = \{e_1, \ldots, e_n\} \subset E$ is called a *trail of length n between j_0 and j_n* if $e_i = \{j_{i-1}, j_i\}$ or $e_i = (j_{i-1}, j_i)$ or $e_i = (j_i, j_{i-1})$ for all $i = 1, \ldots, n$. //

Different undirected path components as well as different cycle components are necessarily disjoint and $\Upsilon(G)$ as well as $\Xi(G)$ form partitions of V. Note that a trail is defined as a subgraph in order to permit different trails on the same set of vertices. But this is only possible when there are more than one edges between two

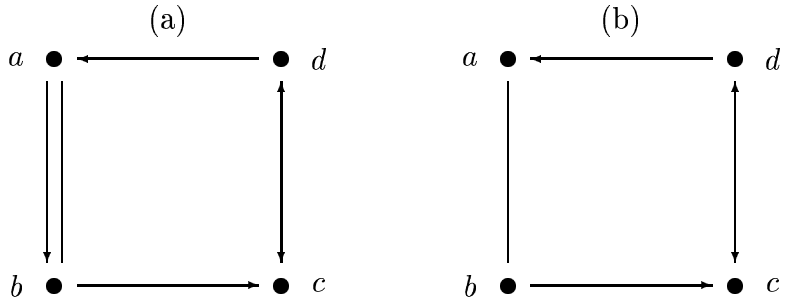
vertices. Paths and cycles may be regarded as special cases of trails.
Graphs without simultaneously undirected and directed edges between two vertices are now defined as follows.

Definition 1.1.3 *Reciprocal graph*
A graph $G = (V, E)$ with $E \cap E^d(v) = \emptyset$ for all $v \in \Upsilon(G)$ is called *reciprocal graph.*//

Note that reciprocal graphs still allow for $(j, k) \in E$ and $(k, j) \in E$, where this means that there is a bidirected edge between these vertices as for instance in the following example.

Example: Figure 1.1 (a) shows the graph $G = (V, E)$ with $V = \{a, b, c, d\}$ and $E = \{\{a, b\}, (a, b), (b, c), (c, d), (d, c), (d, a)\}$. The set of (undirected) path components is given as $\Upsilon(G) = \{\{a, b\}, \{c\}, \{d\}\}$ and the set of cycle components as $\Xi(G) = \{\{a, b, c, d\}\}$. Since $E \cap E^d(v) = \{(a, b)\} \neq \emptyset$ for $v = \{a, b\} \in \Upsilon(G)$ we can see that G is not reciprocal. Deleting the directed edge (a, b) as shown in Figure 1.1 (b) yields a reciprocal graph. //

Figure 1.1: Examples for (a) a non–reciprocal and (b) a reciprocal graph.



The following definition describes the most common graphs.

Definition 1.1.4 *Common graphs*
We say that a graph $G = (V, E)$ is

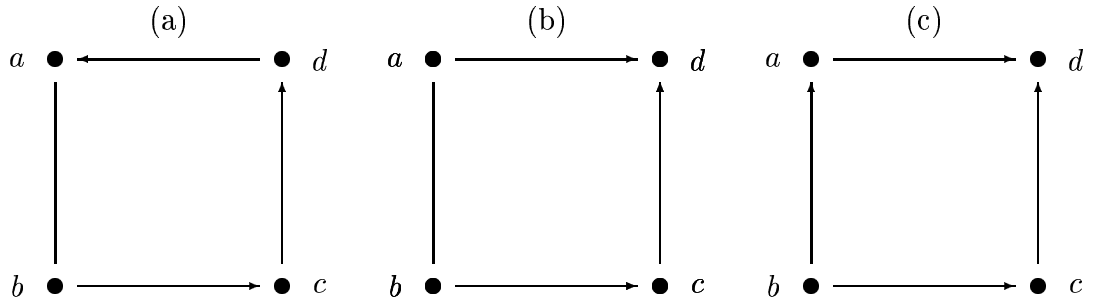
1. *undirected* if it has no directed edges, i.e. $E \cap E^d(V) = \emptyset$;
2. *directed acyclic (DAG)* if it has no undirected edges and no cycles, i.e. $\Upsilon(G) = \Xi(G) = V$;

3. *a chain graph* if it is reciprocal and does not contain any (semi)directed cycles, i.e. $\Upsilon(G) = \Xi(G)$. The sets $\Upsilon(G)$ are then called *chain components.* //

Obviously, undirected and directed acyclic graphs are special cases of chain graphs.

Example: The graph given in Figure 1.1 (a) is not reciprocal and thus no chain graph. The ones in Figure 1.1 (b) and Figure 1.2 (a) contain semidirected cycles, so that they are no chain graphs, too. In contrast, the graph in Figure 1.2 (b) is a chain graph since it is reciprocal and the only cycle is an undirected path component, i.e. $\Upsilon(G) = \Xi(G) = \{\{a, b\}, \{c\}, \{d\}\}$. Figure 1.2 (c) shows a directed acyclic graph. //

Figure 1.2: Examples for (a) a semidirected cycle, (b) a chain graph, and (c) a directed acyclic graph.



Statistical models describing the dependence structure of a set of random variables by the above common graphs are by now well known and elaborated (Lauritzen, 1996). Such graphical models formulate the distributional assumptions in terms of conditional independencies in a way that they can unambiguously be read off the corresponding graph (cf. Chapter 2). This is formalized by considering a random vector $\mathbf{X} = \mathbf{X}_V$ which is indexed by the vertex set of the graph. If $\mathbf{X}_V$ contains continuous as well as discrete components the vertex set V is divided correspondingly into disjoint sets Γ and Δ. In contrast, in event history analysis we consider ordinary random variables, for example covariates, as well as stochastic processes that describe the occurrences of events. When representing this through a graph, it is necessary to distinguish between the set V_R of vertices representing random

variables and V_P representing the processes. In addition, the dependence structure among the processes may be cyclic whereas it must not be cyclic among the random variables. This leads to the following definition of a dynamic graph, the statistical implications of which will be addressed later (cf. Chapter 4).

Definition 1.1.5 *Dynamic graph / directed graph*
Given a graph $G = (V, E)$ with $V = V_R \cup V_P$, $V_R \cap V_P = \emptyset$. Then, G is called a *dynamic graph* if

1. the subgraph G_{V_R} is a chain graph and
2. the subgraph G_{V_P} is a directed graph, i.e. $E \cap E^u(V_P) = \emptyset$ and
3. $\forall\, j \in V_R, k \in V_P : \{j, k\} \notin E$ and $(k, j) \notin E$.

If $V_R = \emptyset$ we have a simple *directed graph.* //

The graphs shown in Figure 1.1 (a) and (b) as well as Figure 1.2 (a) are no dynamic graphs for any choice of V_R and V_P. This can be seen as follows: V_R would have to include a and b since G_{V_P} has to be directed. In Figure 1.1(a) we have that for any $V_R \supset \{a, b\}$ the subgraph G_{V_R} is no chain graph. In Figures 1.1 (b) and 1.2 (a) we have that for any $V_R \supset \{a, b\}$ condition 3 of the above definition can not be satisfied or G_{V_R} is no chain graph. In contrast, the graphs in Figure 1.2 (b) with $V_R = \{a, b\}$ and (c) with $V_R = \emptyset$ are dynamic graphs.

The following terminology can be applied to general graphs.

Definition 1.1.6 *Graph terminology*
Consider a graph $G = (V, E)$. For $A \subset V$,

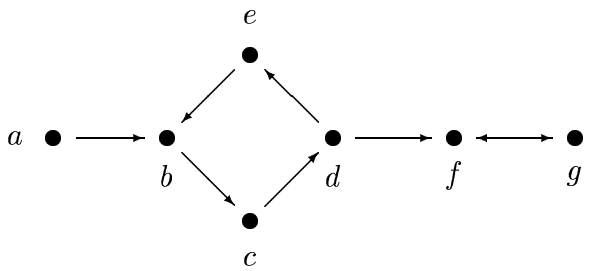
1. the set $\mathrm{pa}(A) = \{k \in V \backslash A | \exists\, j \in A : (k, j) \in E\}$ is called the *parents* of A.
2. the set $\mathrm{nb}(A) = \{k \in V \backslash A | \exists\, j \in A : \{j, k\} \in E\}$ is called the *neighbors* of A.
3. the set $\mathrm{ch}(A) = \{k \in V \backslash A | \exists\, j \in A : (j, k) \in E\}$ is called the *children* of A.
4. the set $\mathrm{bd}(A) = \mathrm{pa}(A) \cup \mathrm{nb}(A)$ is called the *boundary* of A.
5. the set $\mathrm{cl}(A) = \mathrm{bd}(A) \cup A$ is called the *closure* of A.
6. the set $\mathrm{an}(A) = \{k \in V \backslash A | \exists\, j \in A$ with a path from k to $j\}$ is called the *ancestors* of A.

7. the set $\mathrm{de}(A) = \{k \in V \backslash A | \exists\, j \in A$ with a path from j to $k\}$ is called the *descendants* of A.
8. the set $\mathrm{nd}(A) = V \backslash (\mathrm{de}(A) \cup A)$ is called the *non–descendants* of A.
9. if $\mathrm{bd}(A) = \emptyset$, then A is called *ancestral*. For general $A \subset V$, $\mathrm{An}(A)$ is the *smallest ancestral set* containing A, given as $A \cup \mathrm{an}(A)$. //

In the above definition, each set may — in case of ambiguities — be indexed by the graph from which it is constructed, for example $\mathrm{pa}_G(A)$ are the *parents of A in G*.

Example: Let us consider an example to see how the above notions apply to cyclic graphs. The graph shown in Figure 1.3 is given as $G = (V, E)$ with $V = \{a, b, \ldots, g\}$ and $E = \{(a, b), (b, c), (c, d), (d, e), (e, b), (d, f), (f, g), (g, f)\}$. As the graph contains no undirected edges, it is reciprocal. Thus, it is directed but obviously not acyclic. The cycle components are given by $\Xi(G) = \{\{a\}, \{b, c, d, e\}, \{f, g\}\}$, i.e. there are some vertices which are connected by a directed path but only in one direction. For instance, there is a path from e to g given as (e, b, c, d, f, g) whereas there is no path from g to e. Instead, we have a trail between g and e given by the vertices $\{d, e, f, g\}$ and edges $\{(g, f),\ (d, f),\ (d, e)\}$. In cyclic graphs some vertices can be parents as well as children of another vertex at a time.

Figure 1.3: A directed cyclic graph.



In Figure 1.3, for instance, we have that $\mathrm{pa}(f) = \{d, g\}$ and $\mathrm{ch}(f) = \{g\}$. This is similar for ancestors and descendants. In the example, $\mathrm{an}(d) = \{a, b, c, e\}$ and $\mathrm{de}(d) = \{b, c, e, f, g\}$. Obviously, all vertices that are members of the same cycle components are simultaneously ancestors and descendants of each other. Finally,

the only ancestral sets in this graph are $\{a\}$, $\{a, b, c, d, e\}$, and $\{a, b, \ldots, g\}$. //

As will be seen later, complete subsets of undirected graphs play a special role in the theory of graphical models, in particular the maximal ones which are defined as follows.

Definition 1.1.7 *Complete set / clique*
A set $A \subset V$ is called *complete* in an undirected graph $G = (V, E)$ if $E^u(A) \subset E$.
A complete subset A is called a *clique* if for all $j \in V \backslash A$: $E^u(A \cup \{j\}) \not\subset E$. //

In order to reduce the properties of distributions on directed graphs to the ones on undirected graphs it is often helpful to first construct a suitable undirected graph. This is mostly based on the so–called moral graph.

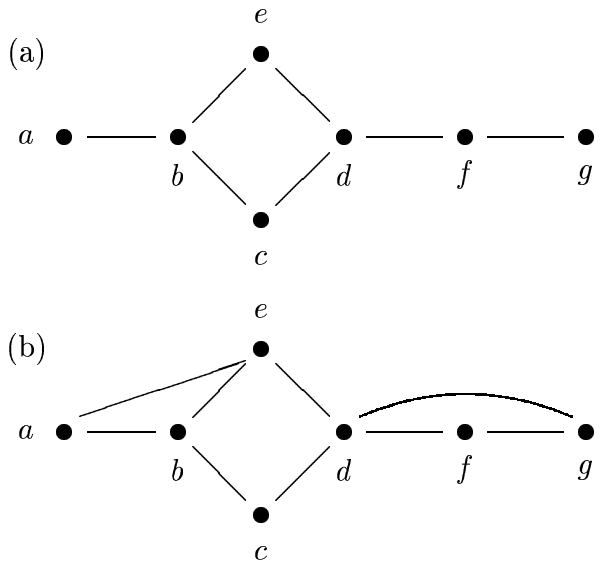
Definition 1.1.8 *Undirected version / moral graph*
Let $G = (V, E)$ be a reciprocal graph.

1. The *undirected version* of G is defined as $G^\sim = (V, E^\sim)$ with $E^\sim = \{\{j, k\} | (j, k) \in E$ or $(k, j) \in E$ or $\{j, k\} \in E\}$, i.e. $G^\sim$ is the undirected graph preserving all undirected edges and replacing all arrows by undirected edges.
2. The *moral* graph $G^m = (V, E^m)$ is given by adding all necessary undirected edges so that $\mathrm{pa}_G(v)$ is complete for all $v \in \Upsilon(G)$, i.e. the parents of children which are in the same undirected path component are *married,* and then taking the undirected version of the resulting graph, i.e. $E^m = E^\sim \cup \{\{j, k\} | \exists\, v \in \Upsilon(G) : j, k \in \mathrm{pa}_G(v)\}$. //

Note that in a directed graph $\Upsilon(G) = V$, implying that moralization is equivalent to adding edges between *vertices with common children* and substituting all directed edges by undirected ones.

Example: Figure 1.4 shows (a) the undirected version of the graph in Figure 1.3 and (b) the corresponding moral graph. The only vertices in G with unconnected parents are b and f so that two undirected edges are added to form the moral graph, where it does not matter that g is parent and child of f at the same time. //

Figure 1.4: (a) Undirected version and (b) moralized version of the graph in Figure 1.3.



1.2 Graph separation

Separating sets play a key role in the theory of graphical models. A very obvious way to define separation is given as follows but is restricted to undirected graphs.

Definition 1.2.1 *Graph separation*
Let $G = (V, E)$ be an undirected graph and $A, B, C \subset V$ pairwise disjoint. Then, C is said to *separate A and B* if every path between any two edges $j \in A$ and $k \in B$ contains elements in C. We denote graph separation by $A \perp\!\!\!\perp_G B|C$. //

As mentioned above, properties of directed graphs are often attributed to those of corresponding undirected ones e.g. when separation in DAGs is considered. This is defined as d–separation (Pearl, 1988).

Definition 1.2.2 *d–separation*
Let $G = (V, E)$ be a DAG and $A, B, C \subset V$ pairwise disjoint. Then, C is said to *d–separate A and B in G* if $A \perp\!\!\!\perp_G B|C$ in the moral subgraph on the smallest ancestral set containing A, B, and C, i.e. in $(G_{\mathrm{An}(A \cup B \cup C)})^m$. //

The original formulation of d–separation does not make use of the moralization operation, but is instead a condition on the trails between A and B. It is often more convenient to use this equivalent definition because it avoids the construction of the moral graph (for the equivalence cf. Lauritzen, 1996, p. 48). It is therefore given in the next remark.

Remark 1.2.3 *Trail condition for d–separation*
Given a DAG G, we say that a trail between j and k is *blocked by* C if it contains a vertice γ such that (1) either directed edges of the trail do not meet head–to–head at γ and $\gamma \in C$, or (2) directed edges of the trail meet head–to–head at γ and γ as well as all its descendants are no elements of C. Otherwise the trail is called *active.* Two disjoint subsets A and B are d–separated by C if all trails between A and B are blocked by C. //

The above notions of separation are well known in the literature and central to the theory of conditional independence graphs especially when simplifications of computational procedures are of interest (Pearl, 1988; Lauritzen, 1996; Cowell et al., 1999) . For our purposes, however, we need a new kind of graph separation which is on the one hand applicable to directed cyclic graphs and has on the other hand some special properties like for instance asymmetry. The proposed separation is called δ–separation and defined below. Note the similarity to d–separation.

Definition 1.2.4 *δ–separation*
Let $G = (V, E)$ be a directed graph. For $B \subset V$, let G^B denote the graph given by deleting all directed edges of G starting in B, i.e. $G^B = (V, E^B)$ with $E^B = E \backslash \{(j,k) | j \in B, k \in V\}$.
Then, we say for pairwise disjoint subsets $A, B, C \subset V$ that C *δ–separates A from B in G* if $A \perp\!\!\!\perp_G B | C$ in the undirected graph $(G^B_{\mathrm{An}(A \cup B \cup C)})^m$, i.e. the moral subgraph of $\mathrm{An}(A \cup B \cup C)$, where the directed edges starting in B have previously been deleted.
In case that A, B and C are not disjoint we define that C δ–separates A from B if $C \backslash B$ δ–separates $A \backslash (B \cup C)$ from B. The empty set is always separated from B.
Additionally, we define that the empty set δ–separates A from B if A and B are unconnected in $(G^B_{\mathrm{An}(A \cup B)})^m$. //

An example for δ–separation is given after the next proposition. Note that the moralization criterion in the above definition only applies to disjoint subsets. This is important for the proofs of the properties of δ–separation given below.

As for d–separation we can also find a trail condition to check for δ–separation. But since this may now also be applied to cyclic graphs we have to modify the original formulation and to show that both versions of δ–separations are equivalent.

Proposition 1.2.5 *Trail condition for δ–separation*
Let $G = (V, E)$ be a directed graph and A, B, C pairwise disjoint subsets of V. Define that any *allowed trail from A to B* contains no edge of the form $(b, k), b \in B, k \in V \backslash B$. For disjoint subsets A, B, C of V, we have that C δ–separates A from B if and only if all allowed trails from A to B are blocked by C.

Proof:
We have to show $A \perp\!\!\!\perp_G B|C$ in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ if and only if every allowed trail from A to B in G is blocked by C.
We start with some preliminary remarks. Let π be an allowed trail from A to B in G. Note first that the situation of $(j, k) \in E$ and $(k, j) \in E$ for j, k on π implies that there are two different trails, one for each directed edge. Second, a trail always consists of distinct vertices. From both statements it follows that cycles within a trail are not possible. Further, we know that if there is a trail π^m between A and B in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ then there exists an allowed trail π^d from A to B in G on the same vertices or containing additional ones which are in G common children of some of the vertices in π^m. In reverse, if there is an allowed trail π^d from A to B in G then either there exists a trail π^m in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ which differs from the former in that it can omit some of the vertices on π^d because they are common children of some other vertices on π^d (this is called a *short cut*). Or the trail π^d contains vertices that are not in $\mathrm{An}(A \cup B \cup C)$ from where it follows that there exists a vertex γ on π^d with $\gamma \notin \mathrm{An}(A \cup B \cup C)$ where the directed edges on this trail meet head–to–head.
Let now $A \perp\!\!\!\perp_G B|C$ in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ and π^d be an allowed trail in G from A to B. If this trail contains vertices which are not in $\mathrm{An}(A \cup B \cup C)$ then it follows from the above considerations that there exists a vertex γ on π^d where the edges meet

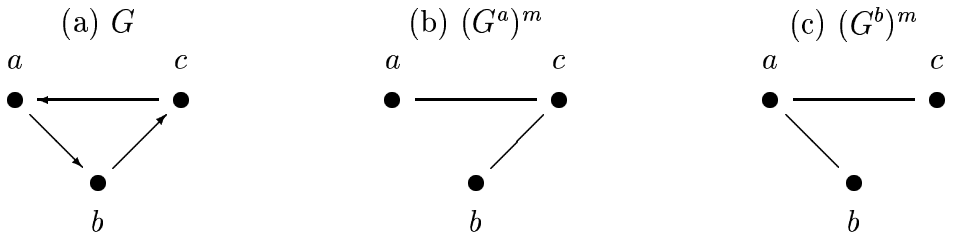
head–to–head and which is not in C as well as all its descendants. Thus, this trail is blocked by C. Otherwise, there exists a trail π^m between A and B in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ with the same vertices or some short–cuts which is intersected by C. In case that no short–cuts are possible, there are no vertices in π^d where edges meet head–to–head and therefore there has to be at least one such vertex $\gamma \in C$. In the other case, it holds for all $\gamma \in \{\mathrm{pa}(\eta)|\ \eta \text{ on } \pi^d \text{ and edges meet head–to–head at } \eta\}$, that no edges meet head–to–head in γ. Since π^m is intersected by C, it follows that at least one of the vertices in π^d where no edges meet head–to–head is element of C. Thus, this trail is also blocked by C.

Finally, let any allowed trail from A to B be blocked by C and consider a trail π^m between A and B in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$. For the corresponding allowed trail π^d it either holds that at least one of its vertices where no edges meet head–to–head is an element of C. In this case it has to be in π^m, so that we have that the latter is intersected by C. Or there exists a vertex γ on π^d where edges meet head–to–head and γ as well as all its descendants are not elements of C. Note that γ has to be an ancestor of $A \cup B \cup C$ because otherwise its parents would not be married and no elements of π^d . Thus, there is a directed path from γ to A or to B (not to C by definition of γ). Since one of these directed paths would yield an allowed trail from A to B without vertices where edges meet head–to–head it has to be intersected by C which is again not possible because then some of the descendants of γ would be elements of C. This argumentation applies to all vertices in π^d where edges meet head–to–head. Thus, any allowed trail from A to B in G which yields a path between A and B in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ always includes at least one vertex which is an element of C and where no edges meet head–to–head, and therefore any such path is intersected by C. □

Example: As mentioned above, δ–separation is not symmetric in A and B. This can already be seen in the simple graph $G = (V, E)$ with $V = \{a, b, c\}$, $E = \{(a, b), (b, c), (c, a)\}$, and $A = \{a\}$, $B = \{b\}$, $C = \{c\}$ given in Figure 1.5 (a). Here, we have that $(E^a)^m = \{\{a, c\}, \{b, c\}\}$ and $(E^b)^m = \{\{a, c\}, \{a, b\}\}$ as in Figure 1.5 (b) and (c), respectively. From the moral graphs we get that $a \perp\!\!\!\perp_G b|c$ in $(G^a)^m$ but not in $(G^b)^m$. Alternatively, there are two trails between a and b: $\{(a, b)\}$ and $\{(c, a), (b, c)\}$. Consider separating b from a, then the first trail is not allowed and

the second is blocked by c since the directed edges do not meet head–to–head in c. In contrast, if we want to separate a from b, the second path is not allowed and the first is not blocked by c. //

Figure 1.5: Example for the asymmetry of δ–sparation.



General graph separation can be regarded as irrelevance relation in the sense that A is irrelevant for B given C and vice versa if $A \perp\!\!\!\perp_G B|C$. It seems sensible to demand that such a relation satisfies specific properties, which are given by the semigraphoid and graphoid axioms (Pearl and Paz, 1987; Pearl, 1988, p. 84; Dawid, 1998). These axioms hold for instance for graph separation in undirected graphs and for d–separation (Verma and Pearl, 1990). This implies that a graphical representation is available for all relations that fulfill these axioms as for instance conditional independence (cf. Appendix A). Symmetry is one of the basic (semi)graphoid properties. Unfortunately, there is no general framework for asymmetric irrelevance relations available yet although other examples for such concepts exist (e.g. Dawid, 1979, 1980; Galles and Pearl, 1996; Cozman and Walley, 1999). Nevertheless, it seems promising to try to gain some insight into the properties of δ–separation by checking if the other (semi)graphoid properties hold. Due to the asymmetry we have two versions of the original formulation as given in the following definition.

Definition 1.2.6 *Asymmetric (semi)graphoid*
Consider a space $\mathcal{V}$ equipped with a semi–order '$\leq$', i.e. '$\leq$' is reflexive and transitive. Assume that for any elements $A, B \in \mathcal{V}$ there exists a least upper bound denoted by $A \vee B$ such that for all $C \in \mathcal{V}$ with $A \leq C$ and $B \leq C$ we have $(A \vee B) \leq C$. The largest lower bound is similarly given by $A \wedge B$. Then, $(\mathcal{V}, \leq)$ is called a *lattice*.

Further, let $(A \text{ IR } B|C)$ be a ternary relation on this lattice. The following properties are called the *asymmetric semigraphoid properties*:

Left redundancy: $A \text{ IR } B \mid A$

Right redundancy: $A \text{ IR } B \mid B$

Left decomposition: $A \text{ IR } B \mid C,\ D \leq A \Rightarrow D \text{ IR } B \mid C$

Right decomposition: $A \text{ IR } B \mid C,\ D \leq B \Rightarrow A \text{ IR } D \mid C$

Left weak union: $A \text{ IR } B \mid C,\ D \leq A \Rightarrow A \text{ IR } B \mid (C \vee D)$

Right weak union: $A \text{ IR } B \mid C,\ D \leq B \Rightarrow A \text{ IR } B \mid (C \vee D)$

Left contraction: $A \text{ IR } B \mid C$ and $D \text{ IR } B \mid (A \vee C) \Rightarrow (A \vee D) \text{ IR } B \mid C$

Right contraction: $A \text{ IR } B \mid C$ and $A \text{ IR } D \mid B \vee C \Rightarrow A \text{ IR } (B \vee D) \mid C$

If additionally the following properties hold we have an *asymmetric graphoid*:

Left intersection: $A \text{ IR } B \mid C$ and $C \text{ IR } B \mid A \Rightarrow (A \vee C) \text{ IR } B \mid (A \wedge C)$

Right intersection: $A \text{ IR } B \mid C$ and $A \text{ IR } C \mid B \Rightarrow A \text{ IR } (B \vee C) \mid (B \wedge C)$. //

While Definition 1.2.6 applies for possibly overlapping sets, the following lemma clarifies the conditions under which it suffices to consider only non–overlapping sets.

Lemma 1.2.7 *Irrelevance for disjoint sets*
Let $\mathcal{V}$ be the power set of V and $A, B, C \in \mathcal{V}$. For a ternary relation $A \text{ IR } B|C$ that holds left redundancy, decomposition, and contraction we have that

$$A \text{ IR } B \mid C \Leftrightarrow A \backslash C \text{ IR } B \mid C. \tag{1.1}$$

For a ternary relation that holds right redundancy, decomposition, and contraction we have that

$$A \text{ IR } B \mid C \Leftrightarrow A \text{ IR } B \backslash C \mid C. \tag{1.2}$$

Proof:
To see (1.1) note that it follows directly from left decomposition that $A \text{ IR } B|C \Rightarrow A \backslash C \text{ IR } B|C$. To show $A \backslash C \text{ IR } B|C \Rightarrow A \text{ IR } B|C$, note that trivially $A \backslash C \text{ IR } B|C \cup$

$(C \cap A)$. Additionally, it follows from left redundancy (i.e. $C \perp_{IR} B|C$) and left decomposition that $(C \cap A) \perp_{IR} B|C$. Left contraction now yields the desired result. Equivalence (1.2) is shown analogously. □

Properties (1.1) and (1.2) are very useful because they permit to generalize results for disjoint sets to the case of overlapping sets. The following corollary, for instance, exploits this to reformulate the intersection property.

Corollary 1.2.8 *Alternative intersection property*
Let $\mathcal{V}$ be the power set of V and $A, B, C \in \mathcal{V}$. Given an ternary relation with property (1.1), left decomposition, and left intersection. For pairwise disjoint sets $A, B, C, D \in \mathcal{V}$ it holds that

$$A \perp_{IR} B \mid (C \cup D) \text{ and } C \perp_{IR} B \mid (A \cup D) \Rightarrow (A \cup C) \perp_{IR} B \mid D. \tag{1.3}$$

With property (1.2), right decomposition, and right intersection it holds that

$$A \perp_{IR} B \mid (C \cup D) \text{ and } A \perp_{IR} C \mid (B \cup D) \Rightarrow A \perp_{IR} (B \cup C) \mid D. \tag{1.4}$$

Proof:
With property (1.1) we have that $A \perp_{IR} B|(C \cup D) \Leftrightarrow (A \cup C \cup D) \perp_{IR} B|(C \cup D)$ from where it follows with left decomposition that $(A \cup D) \perp_{IR} B|(C \cup D)$. With the same argument we get $C \perp_{IR} B|(A \cup D) \Rightarrow (C \cup D) \perp_{IR} B|(A \cup D)$. Left intersection yields $(A \cup C \cup D) \perp_{IR} B|D$ which is again equivalent to $(A \cup C) \perp_{IR} B|D$ because of (1.1). Implication (1.4) can be shown analogously. □

Finally, the following lemma is useful and easily checked.

Lemma 1.2.9 *Converse of contraction*
In Definition 1.2.6 we have that the converse of left (right) contraction follows from left (right) decomposition and left (right) weak union. //

Note that all the properties in Definition 1.2.6 hold for graph separation and d–separation (Verma and Pearl, 1990). Thus, properties (1.1) – (1.4) and the above lemma also hold with the dual formulation being redundant since both notions of separation are symmetric.

Let us now return to the special case of δ–separation. Due to the asymmetry we have to make clear how this translates to the notion of irrelevance relation.

Remark 1.2.10 *δ–separation as irrelevance relation*
The interpretation of δ–separation as irrelevance relation should be that if C δ–separates A from B in G then A is irrelevant for B given C. This is denoted by $A \,\mathrm{IR}_\delta\, B|C$. The semi–order is given by the set inclusion '$\subset$', the join and meet operations by union and intersection, respectively. //

The following proposition states which of the properties of asymmetric graphoids hold for δ–separation.

Proposition 1.2.11 *Graphoid properties for δ–separation*
Let G be a directed graph. The δ–separation satisfies the following graphoid axioms:

(1) left redundancy,

(2) left decomposition,

(3) left and right weak union,

(4) left and right contraction,

(5) left and right intersection.

Proof:
(1) We have by definition for non–disjoint sets that $A \,\mathrm{IR}_\delta\, B|A \Leftrightarrow A\backslash(B\cup A) \,\mathrm{IR}_\delta\, B|A\backslash B$ $\Leftrightarrow \emptyset \,\mathrm{IR}_\delta\, B|A\backslash B$ which is again by definition always true.
It is easily checked that for properties (2), (3), and left contraction we can assume that all involved sets are pairwise disjoint without loss of generality.
(2) It follows from $A \perp\!\!\!\perp_G B|C$ in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ that $D \perp\!\!\!\perp_G B|C$ in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ for $D \subset A$ since decomposition holds for $\perp\!\!\!\perp_G$. Further, $(G^B_{\mathrm{An}(B\cup C\cup D)})^m$ is a subgraph of $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ or has even less edges. Thus, $D \perp\!\!\!\perp_G B|C$ in $(G^B_{\mathrm{An}(B\cup C\cup D)})^m$.
(3) Left weak union can easily be shown: $A \perp\!\!\!\perp_G B|C$ in $(G^B_{\mathrm{An}(A\cup B\cup C)})^m$ implies $A \perp\!\!\!\perp_G B|(C\cup D)$ in $(G^B_{\mathrm{An}(A\cup B\cup C\cup D)})^m = (G^B_{\mathrm{An}(A\cup B\cup C)})^m$ because ordinary weak union holds for graph separation and because $\mathrm{An}(A \cup B \cup C \cup D)$ $=\mathrm{An}(A \cup B \cup C)$ for $D \subset A$. Right weak union is trivial since application of the definition for overlapping sets yields that $(C \cup D)\backslash B = C$.
(4) For left contraction we have to show that

$$A \perp\!\!\!\perp_G B \mid C \text{ in } (G^B_{\mathrm{An}(A\cup B\cup C)})^m \tag{1.5}$$

and

$$D \perp\!\!\!\perp_G B \mid (A \cup C) \text{ in } (G^B_{\text{An}(A\cup B\cup C\cup D)})^m \tag{1.6}$$

imply $(A \cup D) \perp\!\!\!\perp_G B|C$ in $(G^B_{\text{An}(A\cup B\cup C\cup D)})^m$. Since contraction holds for ordinary graph separation it is sufficient to show that $A \perp\!\!\!\perp_G B|C$ in $(G^B_{\text{An}(A\cup B\cup C\cup D)})^m$. Thus, we have to show that there is no path between A and B in the latter graph not intersected by C. Let F be the additional vertices in $(G^B_{\text{An}(A\cup B\cup C\cup D)})^m$, i.e. $F = \text{An}(D)\backslash\text{An}(A \cup B \cup C)$. If $F = \emptyset$ left contraction trivially holds. Otherwise we do not only have these additional vertices but also the additional edges due to directed edges starting in F and due to marrying parents of children in F. By definition of F we have that for all vertices $f \in F$ there exists a directed path in G from f to some $d \in D$ which is not intersected by $\text{An}(A \cup B \cup C)$. From this it follows first of all that any path between B and F in $(G^B_{\text{An}(A\cup B\cup C\cup D)})^m$ has to be intersected by $A \cup C$ because otherwise (1.6) would be violated. Thus, there is no new path between A and B through F not intersected by $A \cup C$. Secondly, for all parents k in G^B of some $f \in F$ there can be no path between k and B not intersected by $A \cup C$ in $(G^B_{\text{An}(A\cup B\cup C\cup D)})^m$ for the same reason. Thus, marrying parents of F does not yield a new path from or through $\text{pa}(F)$ to B not intersected by $A \cup C$. A path from A to B not intersected by C would therefore have to be a direct one. This is in turn not possible because of (1.5) and because A and B trivially have no common children in F (in G^B).

In order to show right contraction in full generality we have to apply the definition for overlapping sets. Let $A^* = A\backslash(B \cup C)$ and $C^* = C\backslash B$, then we have to show that from

$$A^* \perp\!\!\!\perp_G B \mid C^* \text{ in } (G^B_{\text{An}(A\cup B\cup C)})^m \tag{1.7}$$

and

$$A^*\backslash D \perp\!\!\!\perp_G D \mid (B \cup C^*)\backslash D \text{ in } (G^D_{\text{An}(A\cup B\cup C\cup D)})^m \tag{1.8}$$

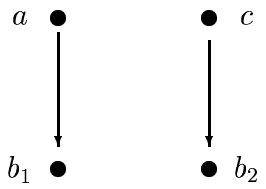
it follows that $A^*\backslash D \perp\!\!\!\perp_G (B \cup D)|C^*\backslash D$ in $(G^{B\cup D}_{\text{An}(A\cup B\cup C\cup D)})^m$. From (1.7) we get by left decomposition that $A^*\backslash D \perp\!\!\!\perp_G B|C^*$ in $(G^B_{\text{An}(A\cup B\cup C)})^m$. From this it follows by similar arguments as applied in the proof for left contraction that $A^*\backslash D \perp\!\!\!\perp_G B|C^*$ in $(G^B_{\text{An}(A\cup B\cup C\cup D)})^m$. The same separation also holds in $(G^{B\cup C}_{\text{An}(A\cup B\cup C\cup D)})^m$ since

the latter graph results by deleting some edges of the former. Thus, we have $A^*\backslash D \perp\!\!\!\perp_G B|C^*$ in $(G^{B\cup D}_{\mathrm{An}(A\cup B\cup C\cup D)})^m$. From (1.8) we get that $A^*\backslash D \perp\!\!\!\perp_G D|(B \cup C^*)\backslash D$ in $(G^{B\cup D}_{\mathrm{An}(A\cup B\cup C\cup D)})^m$ since the latter graph has again the same or less edges than the original one. With $\tilde{A} = A^*\backslash D$ we can show by the properties of ordinary graph separation (decomposition, contraction, and intersection) that $\tilde{A} \perp\!\!\!\perp_G B|C^*$ and $\tilde{A} \perp\!\!\!\perp_G D|(B\backslash D)\cup(C^*\backslash D)$ imply $\tilde{A} \perp\!\!\!\perp_G (B\cup D)|C^*\backslash D$ as desired. This can be seen as follows: With decomposition we get $\tilde{A} \perp\!\!\!\perp_G B\backslash D|(C^*\backslash D)\cup(C^*\cap D)$ and $\tilde{A} \perp\!\!\!\perp_G (C^*\cap D)|(C^*\backslash D)\cup(B\backslash D)$. Intersection and decomposition yield $\tilde{A} \perp\!\!\!\perp_G B\backslash D|C^*\backslash D$. This, together with $\tilde{A} \perp\!\!\!\perp_G D|(B\backslash D)\cup(C^*\backslash D)$ and the contraction property, provides the desired result.

(5) To see left intersection, note that all conditions for (1.1) to hold have been gathered. Thus, we can assume that A, B, C are pairwise disjoint without loss of generality. We then have to show that $A \perp\!\!\!\perp_G B|C$ and $C \perp\!\!\!\perp_G B|A$ in $(G^{B}_{\mathrm{An}(A\cup B\cup C)})^m$ imply $(A\cup C) \perp\!\!\!\perp_G B$ in the same graph. This is fulfilled by the intersection property of ordinary graph separation.

For right intersection, we can again assume that $A\cap B = \emptyset$ and $A\cap C = \emptyset$ because of (1.1). We have to show that $A \perp\!\!\!\perp_G B|C\backslash B$ in $(G^{B}_{A\cup B\cup C})^m$ and $A \perp\!\!\!\perp_G C|B\backslash C$ in $(G^{C}_{A\cup B\cup C})^m$ imply $A \perp\!\!\!\perp_G (B\cup C)$ in $(G^{B\cup C}_{A\cup B\cup C})^m$. The first two graph separations also hold in the third graph because it has the same or less edges. Applying weak union and intersection for ordinary graph separation yields the desired result. □

Figure 1.6: Counterexample for property (1.2) of δ–separation.



Note that with these results not only property (1.1) but also (1.3) hold for δ–separation. So far, however, the conditions for (1.2) and (1.4) are not satisfied because right redundancy does not hold. By definition this would imply that $A \,\mathrm{IR}_\delta\, B|B \Leftrightarrow A\backslash B \,\mathrm{IR}_\delta\, B|\emptyset$ which is only true if $A\backslash B$ and B are unconnected in $(G^{B}_{\mathrm{An}(A\cup B)})^m$. A simple counterexample is given by the graph with $V = \{a, b\}$ and $E = \{(a, b)\}$.

Additionally, property (1.2) does not hold, as can be seen by another example shown in Figure 1.6: Let $V = \{a, b_1, b_2, c\}$, $E = \{(a, b_1), (c, b_2)\}$, $A = \{a\}$, $B = \{b_1, b_2\}$, and $C = \{b_1, c\}$. Then, $A \,\text{IR}_\delta\, B \backslash C | C$ but not $A \,\text{IR}_\delta\, B | C$ since the latter only holds if $A \,\text{IR}_\delta\, B | C \backslash B$. In contrast, the converse does hold because it is a special case of the subsequent result.

Lemma 1.2.12 *Special case of right decomposition for δ–separation*
Given a directed graph G, it holds that:

$$A \,\text{IR}_\delta\, B \mid C,\ D \subset B \Rightarrow A \,\text{IR}_\delta\, D \mid (C \cup B) \backslash D$$

Proof:
Due to property (1.1) we can assume that $A \cap C = \emptyset$. Let $A^* = A \backslash B$ and $C^* = C \backslash B$. Then, we have to show that $A^* \perp\!\!\!\perp_G B | C^*$ in $(G^B_{\text{An}(A \cup B \cup C)})^m$ implies $A^* \perp\!\!\!\perp_G D | C^* \cup (B \backslash D)$ in $(G^D_{\text{An}(A \cup B \cup C)})^m$. Note that $A^* \perp\!\!\!\perp_G D | C^* \cup (B \backslash D)$ in $(G^B_{\text{An}(A \cup B \cup C)})^m$ holds due to weak union and decomposition of ordinary graph separation. Changing the graph to $(G^D_{\text{An}(A \cup B \cup C)})^m$ means that all edges that are present in G as directed edges starting in $B \backslash D$ are added. Additionally, those edges have to be added which result from vertices in $B \backslash D$ having common children with other vertices. Since all these new edges involve vertices in $B \backslash D$ there can be no additional path between A^* and D in $(G^D_{\text{An}(A \cup B \cup C)})^m$ not intersected by $C^* \cup (B \backslash D)$. □

Although (1.2) does not hold in full generality it is easily checked that (1.4) holds.

Proposition 1.2.13 *Alternative intersection property for δ–separation*
Property (1.4) holds for δ–separation, i.e.

$$A \,\text{IR}_\delta\, B \mid (C \cup D) \text{ and } A \,\text{IR}_\delta\, C \mid (B \cup D) \Rightarrow A \,\text{IR}_\delta\, (B \cup C) \mid D$$

for pairwise disjoint sets A, B, C, D.

Proof:
Given that $A \perp\!\!\!\perp_G B | (C \cup D)$ in $(G^B_{A \cup B \cup C \cup D})^m$ and $A \perp\!\!\!\perp_G C | (B \cup D)$ in $(G^C_{A \cup B \cup C \cup D})^m$, both graph separations also hold in $(G^{B \cup C}_{A \cup B \cup C \cup D})^m$. With the properties of $\perp\!\!\!\perp_G$ it follows that $A \perp\!\!\!\perp_G (B \cup C) | D$ in $(G^{B \cup C}_{A \cup B \cup C \cup D})^m$. □

Note that the above proposition does not necessarily hold if B, C, D are not disjoint. But it can be shown by a very similar proof that it remains valid if $A \cap B \neq \emptyset$ or $A \cap C \neq \emptyset$.

As mentioned above, right decomposition does not hold in general for δ–separation. A simple counterexample is for instance given by the graph $G = (V, E)$ with $V = \{a, b, c, d\}$ and $E = \{(a, c), (b, c), (b, d)\}$. Then, $\{c\}$ δ–separates $\{a\}$ from $\{b, d\}$ but it is not true that $\{c\}$ δ–separates $\{a\}$ from $\{d\}$ since in $(G^d)^m$ we have the undirected edge $\{a, b\}$ due to moralization, and the edge $\{b, d\}$ yielding a path from $\{a\}$ to $\{d\}$ that is not intersected by $\{c\}$. However, under suitable conditions we get the following result.

Proposition 1.2.14 *Right decomposition for δ–separation*
Given a directed graph $G = (V, E)$. Right decomposition as formulated in Definition 1.2.6 holds for δ–separation in the special case that $(A \cap B) \backslash (C \cup D) = \emptyset$ and

(1) either $B \perp\!\!\!\perp_\delta D | (A \cup C)$

(2) or $B \perp\!\!\!\perp_\delta A \backslash (C \cup D) | (C \cup D)$ and for all $k \in C \backslash D$ either $A \perp\!\!\!\perp_\delta \{k\} | (C \backslash \{k\}) \cup B$ or $B \perp\!\!\!\perp_\delta \{k\} | ((C \backslash \{k\}) \cup D \cup A)$.

Proof:
Due to property (1.1) we can assume that $A \cap C = \emptyset$. Since the other involved subsets are not necessarily disjoint, we have to show that $A \backslash B \perp\!\!\!\perp_\delta B | C \backslash B$ implies $A \backslash D \perp\!\!\!\perp_\delta D | C \backslash D$ for $D \subset B$. Let $A^* = A \backslash D$ and $C^* = C \backslash D$. Equivalently we then have to show that $A^* \perp\!\!\!\perp_\delta B | C^* \backslash B \Rightarrow A^* \perp\!\!\!\perp_\delta D | C^*$, where the assumption $A^* \cap B = \emptyset$ has been used.
The first step is to reduce the graph by deleting all vertices in $\mathrm{An}(A \cup B \cup C) \backslash \mathrm{An}(A \cup C \cup D)$. Let $B^* = (B \cap \mathrm{An}(A \cup C \cup D)) \backslash D$, then $A^* \perp\!\!\!\perp_\delta B | C^* \backslash B$ implies $A^* \perp\!\!\!\perp_\delta (B^* \cup D) | C^* \backslash B^*$ since $(G^{B^* \cup D}_{\mathrm{An}(A \cup C \cup D)})^m$ is just a subgraph of $(G^B_{\mathrm{An}(A \cup B \cup C)})^m$ or has even less edges. With Lemma 1.2.12 we get $A^* \perp\!\!\!\perp_\delta D | C^* \cup (B^* \backslash C^*)$. Thus, we have to show that $B^* \backslash C^*$ can be discarded in the conditioning set.
With assumption (1) it follows by definition that $B \backslash (C^* \cup D) \perp\!\!\!\perp_\delta D | (A^* \cup C^*)$. With left decomposition we get $B^* \backslash C^* \perp\!\!\!\perp_\delta D | (A^* \cup C^*)$. Application of left intersection and left decomposition (according to (1.3)) to this and $A^* \perp\!\!\!\perp_\delta D | C^* \cup (B^* \backslash C^*)$ yields

the desired result.
In case of (2) we allow $B^*\backslash C^*$ to be relevant for D given $A^* \cup C^*$. With Lemma 1.2.12 applied like above we get $A^* \operatorname{IR}_\delta B^*\backslash C^* | (C^* \cup D)$ so that any trail between A^* and $B^*\backslash C^*$ ending with a directed edge headed in $B^*\backslash C^*$ is blocked by $C^* \cup D$. From the first part of (2) it follows that $B^*\backslash C^* \operatorname{IR}_\delta A^* | (C^* \cup D)$ so that any trail between A^* and $B^*\backslash C^*$ ending with a directed edge headed in A^* is also blocked by $C^* \cup D$. With the second part of (2) we finally get that any other trail between A^* and $B^*\backslash C^*$ is also blocked by $C^* \cup D$. Therefore, $A^* \perp\!\!\!\perp_G B^*\backslash C^* | (C^* \cup D)$ in $(G^D_{\mathrm{An}(A\cup C\cup D)})$ must hold. With $A^* \perp\!\!\!\perp_G D | C^* \cup (B^*\backslash C^*)$ in $(G^D_{\mathrm{An}(A\cup C\cup D)})$ as shown above, application of the contraction and decomposition property of ordinary graph separation yields $A^* \perp\!\!\!\perp_G D | C^*$ in $(G^D_{\mathrm{An}(A\cup C\cup D)})$ as desired. □

With the preceding propositions we have shown that almost all properties of an asymmetric graphoid hold for δ–separation and that conditions can be found for right decomposition. Similar properties and conditions can be shown for the concept of local independence, which is asymmetric, too, in Chapter 3.

In contrast, conditional independence and graph separation as well as d–separation are symmetric and hold all the graphoid axioms, as already mentioned. This is the basis for constructing meaningful conditional independence graphs which are considered in the following chapter.

Chapter 2

Graphical models for random structures in time

Before introducing graphical models based on the concept of local independence, we revise in this chapter some other approaches to the application of graphical models to event history data and time series as can be found in the literature. Typically, these are closely related to the classical conditional independence graphs. This overview serves as a first idea about the specific difficulties encountered when modeling random structures in time.

In Section 2.1, we start by presenting models that are based on conditional independence graphs, i.e. the graphs describe conditional independencies induced by suitable Markov properties. The vertices may either represent random variables, e.g. the components of multivariate survival times (Section 2.1.1 and 2.1.2), or whole time series as proposed by Dahlhaus (2000) and addressed in Section 2.1.3. These approaches are mainly based on undirected graphs so that the dynamic character cannot be the main focus of the representation. However, since in event history and time series analysis the dynamic development is of particular interest, directed graphs seem more natural to represent the independence structure. In this context we can find two approaches. The first is an application of directed acyclic graphs to multivariate survival times that occur in a fixed order (cf. Section 2.1.1). The second is based on chain graphs, where each chain component represents one discrete point in time (Section 2.1.2) as proposed by Lynggaard and Walther (1993).

A basically different approach consists in representing dynamic independencies instead of conditional independencies. This has been explored for time series with a discrete time parameter by Eichler (1999, 2000) and is addressed in Section 2.2. The dynamic independencies are closely related to the notion of Granger–noncausality (Granger, 1969) wherefore the graphs have been termed *causality graphs.* Note, however, that this refers to a different concept of causality than usually discussed in the context of graphical models (Pearl, 1995; Lauritzen, 2000) which is addressed in Section 4.5.

2.1 Conditional independence graphs

In order to present conditional independence models for time series and event history data, we give a brief and rather informal introduction to conditional independence graphs. For a deeper insight we refer to the monographs by Edwards (2000), Lauritzen (1996), or Whittaker (1990). In general, a graph induces a statistical model for a multivariate random vector $\mathbf{X}_V = (X_1, \ldots, X_K)$, $V = \{1, \ldots, K\}$, by postulating that subvectors $\mathbf{X}_A$, $\mathbf{X}_B$, and $\mathbf{X}_C$, satisfy specific conditional independence relations whenever the subsets $A, B, C \subset V$ fulfill corresponding separations in the graph $G = (V, E)$. The precise formulation of the Markov properties depends on the type of graph.

Undirected conditional independence graphs

Undirected graphs are the most intuitive and therefore treated first. The basic notion of conditional independence is denoted by $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_C$, or briefly $A \perp\!\!\!\perp B|C$, meaning that $\mathbf{X}_A$ is conditionally independent of $\mathbf{X}_B$ given $\mathbf{X}_C$ (Dawid, 1979). Some background on conditional independence is given in Appendix A.

Definition 2.1.1 *Undirected conditional independence graph*
Let $G = (V, E)$ be an undirected graph. A distribution P for a multivariate random vector $\mathbf{X}_V$ is said to be *G–Markovian* if it satisfies for all disjoint subsets $A, B, C \subset V$

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C \text{ whenever } A \perp\!\!\!\perp_G B|C \text{ in } G. \tag{2.1}$$

Property (2.1) is called the *global undirected Markov property.* //

The following result is central to most simplifications of statistical inference procedures for conditional independence graphs.

Proposition 2.1.2 *Factorization*
Consider an undirected graph G and a G–Markovian distribution P. If P admits a positive and continuous density f w.r.t. a product measure μ, we have that (2.1) is equivalent to the existence of the following factorization of the density:

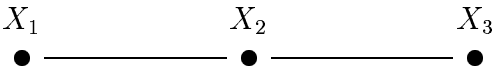
$$f(\mathbf{x}) = \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C), \tag{2.2}$$

where $\mathcal{C}$ is the set of cliques of G and ψ_C, $C \in \mathcal{C}$, are some positive functions on the space of $\mathbf{X}_C$.

Proof: Lauritzen (1996, p. 36). □

Example: Consider the graph in Figure 2.1. With (2.1) we have $X_1 \perp\!\!\!\perp X_3 | X_2$. Assuming the existence of a positive density it follows from the above proposition that $f(x_1, x_2, x_3) = \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)$. Such a factorization is for instance given by $f(x_1, x_2, x_3) = f(x_1|x_2)f(x_2)f(x_3|x_2)$. //

Figure 2.1: Example for an undirected conditional independence graph.



Directed acyclic graphs
Another type of conditional independence graph is given by directed acyclic graphs. The corresponding Markov properties yield, in specific cases, conditional independencies that could not be read off an undirected graph, so that DAG models constitute a real extension to the latter.

Definition 2.1.3 *DAG model*
Let $G = (V, E)$ be a DAG. A distribution P for a multivariate random vector $\mathbf{X}_V$ is said to be *G–Markovian* if it satisfies

$$X_k \perp\!\!\!\perp \mathbf{X}_{\mathrm{nd}\setminus\mathrm{pa}(k)} | \mathbf{X}_{\mathrm{pa}(k)}. \tag{2.3}$$

Property (2.3) is called the *local directed Markov property.* //

The local directed Markov property implies even more conditional independencies than apparent at first sight. It can be shown that (2.3) is equivalent to the *global* directed Markov property.

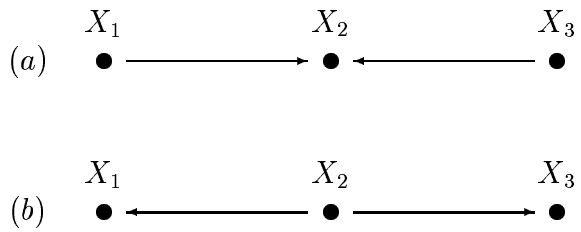
Proposition 2.1.4 *Global directed Markov property*
Consider a distribution P for a multivariate random vector $\mathbf{X}_V$ and a DAG $G = (V, E)$. Then, P being G–Markovian is equivalent to the following implication: For all disjoint subsets $A, B, C \subset V$ with $A \perp\!\!\!\perp_G B|C$ in $(G_{\mathrm{An}(A\cup B\cup C)})^m$ it holds that $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B|\mathbf{X}_C$.

Proof: Lauritzen (1996, p. 51). □

Note, that separation in DAGs can also be verified with the d–separation criterion mentioned in the first chapter since this is equivalent to the separation in the moral graph.

Example: A very simple example illustrates the logic of moralizing directed acyclic graphs when looking at separation.

Figure 2.2: Examples for directed conditional independence graph.



The graph in Figure 2.2 (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 anymore, 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 distributions of subvectors can be retained, therefore requires that parents of common children are linked. Thus, the moral graph for Figure 2.2 (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 2.2 (a). In contrast, the graph in Figure 2.2 (b) encodes the same conditional independencies as the undirected graph in Figure 2.1 which is the moral graph of the former. //

Similar to the undirected Markov property, the local directed Markov property (2.3) results in a specific factorization of the density.

Proposition 2.1.5 *Recursive factorization*
Consider a DAG G and a multivariate distribution P. If P admits a density f, then (2.3) is equivalent to the following recursive factorization:

$$f(\mathbf{x}) = \prod_{k \in V} f(x_k | \mathbf{x}_{\mathrm{pa}(k)}), \tag{2.4}$$

where $f(x_k|\mathbf{x}_{\mathrm{pa}(k)})$ denotes the conditional density of X_k given the parents in G.

Proof: Lauritzen (1996, p. 51). □

Most of the properties of DAG models have been developed and applied in the field of artificial intelligence and expert systems (Pearl, 1988; Cowell et al., 2000). This is mainly due to the above factorization which 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. Then, the resulting multivariate distribution holds the global directed Markov property. The specification of a complex multivariate distribution through univariate regressions induced by a DAG can be regarded as *local modeling* because the single regression models typically involve considerably less variables than the whole multivariate vector.

Graphical chain models

The most general class of conditional independence graphs is the one of graphical chain models. These comprise undirected conditional independence graphs and

DAGs as special cases. Chain graphs have been defined in the preceding chapter as reciprocal graphs without (semi)directed cycles.

Definition 2.1.6 *Chain graph Markov property*
Let $G = (V, E)$, $V = \{1, \ldots, K\}$, be a chain graph. For disjoint subsets $A, B, C \subset V$, let $(G_{\mathrm{An}(A \cup B \cup C)})^m$ denote the moral graph on the smallest ancestral set containing $A \cup B \cup C$. If the distribution P of a multivariate random vector $\mathbf{X}_V = (X_1, \ldots, X_K)$ satisfies for any disjoint subsets $A, B, C \subset V$:

$$A \perp\!\!\!\perp_G B | C \text{ in } (G_{\mathrm{An}(A \cup B \cup C)})^m \;\Rightarrow\; \mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C, \tag{2.5}$$

then P is said to satisfy the *(global) chain graph Markov property.* //

The chain graph Markov property (2.5) is very similar to the global directed Markov property (cf. Proposition 2.1.4). To see the difference, recall that an ancestral set in a chain graph includes vertices which are connected by undirected or semidirected paths, in particular all the neighbors of the involved variables.

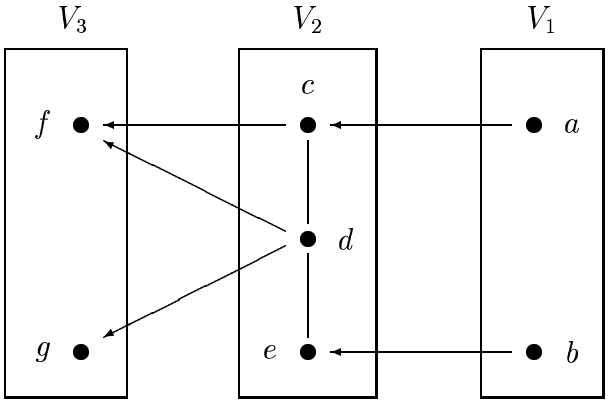
Example: Consider the chain graph in Figure 2.3. The chain components are given by $\Upsilon(G) = \{\{a\}, \{b\}, \{c, d, e\}, \{f\}, \{g\}\}$. These can also be grouped as blocks $V_1 = \{a, b\}$, $V_2 = \{c, d, e\}$, and $V_3 = \{f, g\}$ forming a partition of V, as shown by the boxes in Figure 2.3. Each block may contain one or more chain components as long as there are no directed edges within one block. The ordering of the blocks underlies the restriction that all directed edges point from prior levels to higher ones. We do not discuss all conditional independencies that can be read off the graph, but instead highlight some important and maybe unapparent characteristics. The chain graph Markov property induces for instance the following conditional independencies:

(1) It holds that $\mathbf{X}_{\{a,b\}} \perp\!\!\!\perp \mathbf{X}_{\{f,g\}} | \mathbf{X}_{\{c,d,e\}}$. This can be seen from the corresponding moral graph $(G_{\mathrm{An}(\{a,b\} \cup \{c,d,e\} \cup \{f,g\})})^m = (V, E^m)$ with $\mathrm{An}(\{a, b\} \cup \{c, d, e\} \cup \{f, g\}) = V$ and $E^m = \{\{a, b\}, \{a, c\}, \{b, e\}, \{c, d\}, \{c, f\}, \{d, e\}, \{d, f\}, \{d, g\}\}$. Another way to verify the claimed independence is to show that in this special case $V_1 \perp\!\!\!\perp V_3 | V_2$ since there is no directed arrow from any variable in V_1 to the variables in V_3. In general, the dependence structure among the blocks of a chain graph may be represented by a directed acyclic graph as addressed below.

(2) We further have that $X_c \perp\!\!\!\perp X_e | \mathbf{X}_{\{a,b,d\}}$. This may again be verified in the corresponding moral graph $(G_{\mathrm{An}(\{a,b\}\cup\{c,d,e\})})^m$ with $\mathrm{An}(\{a,b\} \cup \{c,d,e\}) = \{a,b,c,d,e\}$ and $E^m = \{\{a,b\},\{a,c\},\{b,e\},\{c,d\},\{d,e\}\}$. From this it follows more specifically that $X_c \perp\!\!\!\perp X_e | \mathbf{X}_{\{a,d\}}$ as well as $X_c \perp\!\!\!\perp X_e | \mathbf{X}_{\{b,d\}}$. The set $\{f,g\}$ can be discarded since it contains only descendants of the involved variables. Another argument for the above conditional independence is that for each block V_j those conditional independencies, that can be read off the undirected subgraph G_{V_j}, hold conditional on all prior blocks $V_1, \ldots, V_{j-1}$, i.e. G_{V_j} describes the independence structure of the conditional distribution of $\mathbf{X}_{V_j}$ given $\mathbf{X}_{V_1 \cup \ldots \cup V_{j-1}}$.

(3) Finally, we have for instance that $X_a \perp\!\!\!\perp X_g | X_d$. The relevant moral graph $(G_{\mathrm{An}(\{a,d,g\})})^m$ where off this independence can be read consists of the vertices $\mathrm{An}(\{a,\, d,\, g\}) = \{a,b,c,d,e,g\}$ and the undirected edges $E^m = \{\{a,b\},\, \{a,c\},\, \{b,e\},\, \{c,d\},\, \{d,e\},\, \{d,g\}\}$. The vertex f can be discarded since it is a pure descendant of the others, whereas $\{b,e\}$ are ancestors of d and g so that they are included in the ancestral set. Moreover, we have by the same argument that $X_g \perp\!\!\!\perp \mathbf{X}_{\{a,b,c,e\}} | X_d$ similar to the local directed Markov property (2.3) for DAGs. //

Figure 2.3: Example of a chain graph.



The idea of chain graphs with blocks $V_1, \ldots, V_J$ as described above is to reveal the conditional independence structures of different marginal and conditional distributions. We can read off the graph the conditional independencies in the (marginal) distribution of $\mathbf{X}_{V_1 \cup \ldots \cup V_j}$, $j = 1, \ldots, J$, as well as those in the distributions of $\mathbf{X}_{V_j}$

conditional on $\mathbf{X}_{V_1 \cup \ldots \cup V_{j-1}}$. A reformulation of the chain graph Markov property which is better suited for the interpretation of the graphs reads as follows.

Corollary 2.1.7 *Chain graph Markov property (revisited)*
Consider a chain graph $G = (V, E)$ with blocks $V_1, \ldots V_J$ as described above. If the distribution P of $\mathbf{X}_V$ admits a positive density f w.r.t. a product measure μ we have that the chain graph Markov property is equivalent to the following three properties:

(1) Let $D = (V^D, E^D)$ be a directed graph with $V^D = \{1, \ldots, J\}$ and $E^D = \{(i,j) | \exists\, k \in V_i$ and $l \in V_j$ such that $(k,l) \in E\}$. Then, D reflects the dependence structure among the blocks, i.e. similar to (2.3) we have

$$V_j \perp\!\!\!\perp \{V_i |\, i \in V^D \backslash \mathrm{pa}_D(j)\} \mid \{V_k | k \in \mathrm{pa}_D(j)\}.$$

(2) The conditional distribution of $\mathbf{X}_{V_j}$ given $(\mathbf{X}_{V_1 \cup \ldots \cup V_{j-1}})$ holds the undirected Markov property (2.1) w.r.t. the subgraph G_{V_j}.

(3) With $C_j = \bigcup_{i=1}^{j} V_i$ it holds for all $k \in V_j$ that

$$X_k \perp\!\!\!\perp \mathbf{X}_{C_j \backslash \mathrm{cl}(k)} \mid \mathbf{X}_{\mathrm{bd}(k)}$$

for all $j = 1, \ldots, J$.

Proof: Frydenberg (1990b). □

Note that in the preceding example we considered three conditional independence statements corresponding to (1) – (3) of the above proposition. For a detailed discussion of further properties of chain graph models confer e.g. Wermuth and Lauritzen (1990) and Frydenberg (1990b).

2.1.1 Conditional independence models

Based on the above graphical models, we now consider a specific setting for event history data. It is assumed that for any individual a fixed number of survival times $T_1, \ldots, T_K$ is observed, where T_k denotes the time of the occurrence of a specific event e_k. The main two approaches to a graphical representation of the conditional

independence structure of the survival times are: (1) undirected conditional independence graphs, where the essential problem consists in finding an appropriate multivariate distribution, and (2) DAGs, where the dependencies have to be specified through appropriate regression models according to (2.4). Both approaches have some disadvantages as illustrated in the following. The problem of censored data, although obviously relevant in this setting, is not explicitly addressed.

Undirected conditional independence graphs

In principle, it is possible to represent conditional independencies of the joint distribution of the survival times $T_1, \ldots, T_K$ by an undirected graph. The basic problem, here, lies in finding an appropriate class of multivariate distributions which permits to parameterize all conditional independencies induced by the graph. The most popular distributional assumption for graphical models is the conditional Gaussian distribution (CG–distribution) which is investigated in this context by Lauritzen and Wermuth (1989) and Frydenberg and Lauritzen (1989). However, the CG–distribution is obviously not suitable for survival times which are usually measured continuously taking only positive values. Nevertheless, it is difficult to find in the literature a more general class of multivariate distributions which allows for a flexible modeling of conditional independence structures. The most promising approach is investigated by Caputo (1998) and is based on a general way of constructing a multivariate distribution from almost arbitrary marginal distributions. The underlying idea is proposed by Koehler and Symanowski (1995). Its main advantage consists in that the properties of the corresponding classes of multivariate distributions are almost independent of the choice of marginal distributions, i.e. the association structure is driven by separate parameters and based on a specific copula (Heinicke, 1999). It is therefore possible to investigate the independence structure without referring to some specific distributional assumptions. This even broadens the potential field of application beyond survival analysis. Caputo (1998) indicates the restrictions on the parameters induced by the Markov properties of undirected graphs and shows how ML–estimation can be performed and simplified by decomposition in a subclass of these multivariate distributions.

However, this approach has some drawbacks, too. It turns out that in the classes of distributions constructed in accordance with Koehler and Symanowski (1995),

marginal and conditional independencies are equivalent. This constitutes a severe limitation to the practical suitability since important forms of dependencies cannot be modeled, as for instance Simpson's paradox (Simpson, 1951). Further, Heinicke (1999) shows that among other unpleasant properties the association parameters are very difficult to interpret. No simple relation between the parameters and well–known association measures, such as marginal and partial correlations, can be established. In addition, situations can be found where the ML–estimator does not exist.

A flexible modeling of the association structure is required for the application to graphical models. However, the only model class known to fulfill this requirement, besides the CG–distributions, is the one based on the Koehler and Symanowski–principle. Since this class has severe limitations no sensible and implementable approach to undirected conditional independence graphs for multivariate survival data seems available at present.

Models based on directed acyclic graphs

From the factorization (2.4) it can be seen that DAGs are suitable for the representation of the conditional independence structure when the variables can be semi ordered like a data generating process such that the parents of a vertice in a graph represent the direct influences, the ancestors the indirect influences and the descendants the consequences, whereas variables that are neither ancestors nor descendants of each other are independent. As mentioned above, the construction of a multivariate distribution only requires the specification of appropriate univariate regression models which is a well–known method in survival analysis. Directed acyclic graphs may therefore be used for event history analysis if the events occur in an almost fixed order, i.e. when the semi order induced by the DAG is compatible with the order of occurrences of the events. The idea is to specify the regression model for a survival time T_k through the corresponding hazard rate conditional on the previous events which are potential explanatory variables for T_k. In order to formalize this approach we first define some basic notions of survival analysis.

Definition 2.1.8 *Survival function / hazard rate*

Let T be a continuous nonnegative random variable with probability distribution P

and assume that the density f exists. The *survival function* $S(t)$ is given as

$$S(t) = P(T > t), \quad t \geq 0.$$

The *hazard rate* $\alpha(t)$ of T is given as

$$\alpha(t) = \lim_{h \downarrow 0} \frac{1}{h} P(t \leq T < t + h | T \geq t), \quad t \geq 0, \tag{2.6}$$

which equals in the absolutely continuous case $f(t)/S(t)$. //

Note that the distribution function of T is uniquely determined by the hazard rate through the relation

$$S(t) = \exp\left(- \int_0^t \alpha(s)\, ds \right), \quad t \geq 0,$$

or, equivalently, through

$$f(t) = \alpha(t) \exp\left(- \int_0^t \alpha(s)\, ds \right), \quad t \geq 0.$$

The hazard rate can be regarded as describing the dynamic development conditional on the past. For instance, in the case of two survival times T_1 and T_2 with $P(T_1 < T_2) = 1$ and joint density $f(t_1, t_2)$ we may specify the joint distribution through

$$f(t_1, t_2) = \begin{cases} f(t_1) f(t_2 | T_1 = t_1), & t_1 < t_2 \\ 0, & \text{otherwise.} \end{cases}$$

Correspondingly, the conditional hazard rate for T_2 reads as

$$\begin{aligned} \alpha(t_2 | t_1) &= \lim_{h \downarrow 0} \frac{1}{h} P(t_2 \leq T_2 < t_2 + h | T_2 \geq t_2, T_1 = t_1) \\ &= \frac{f(t_2 | t_1)}{S(t_2 | t_1)}, \quad 0 \leq t_1 < t_2, \end{aligned} \tag{2.7}$$

yielding the conditional density

$$f(t_2 | t_1) = \alpha(t_2 | t_1) \exp \int_{t_1}^{t_2} \alpha(s | t_1)\, ds, \quad 0 \leq t_1 < t_2.$$

Note that the above definition of a conditional hazard rate is indeed only sensible if $P(T_1 < T_2) = 1$. Otherwise, the condition $T_1 = t_1$ for $t_1 > t_2$ in the conditional probability $P(t_2 \leq T_2 < t_2 + h | T_2 \geq t_2, T_1 = t_1)$ would contain information on the development of the process beyond t_2 which could lead to degenerate hazard rates. The dependence structure among several survival times $T_1, \ldots, T_K$ may thus be represented by a DAG with the restriction that if $j \in \mathrm{an}(k)$ in the graph then $P(T_j < T_k) = 1$. This is specified in the following definition.

Proposition 2.1.9 *DAG model for survival times*
Let $\mathbf{T} = (T_1, \ldots, T_K)$ be a multivariate vector of nonnegative random variables and $G = (V, E)$, $V = \{1, \ldots, K\}$, a DAG. Assume that for any pair $j, k \in V$ such that $j \in \mathrm{an}(k)$ we have $P(T_j < T_k) = 1$. Then, it holds that

$$T_k \perp\!\!\!\perp \mathbf{T}_{\mathrm{nd}\setminus\mathrm{cl}(k)} \mid \mathbf{T}_{\mathrm{pa}(k)} \quad \forall\, k \in V,$$

if and only if

$$f(\mathbf{t}) = \prod_{k \in V} f(t_k | \mathbf{t}_{\mathrm{pa}(k)}),$$

where

$$f(t_k | \mathbf{t}_{\mathrm{pa}(k)}) = \alpha(t_k | \mathbf{t}_{\mathrm{pa}(k)}) \exp \int_{t^*}^{t_k} \alpha(s | \mathbf{t}_{\mathrm{pa}(k)})\, ds, \quad 0 \leq t^* < t_k, \tag{2.8}$$

with $t^* = \max(t_{\mathrm{pa}(k)})$.

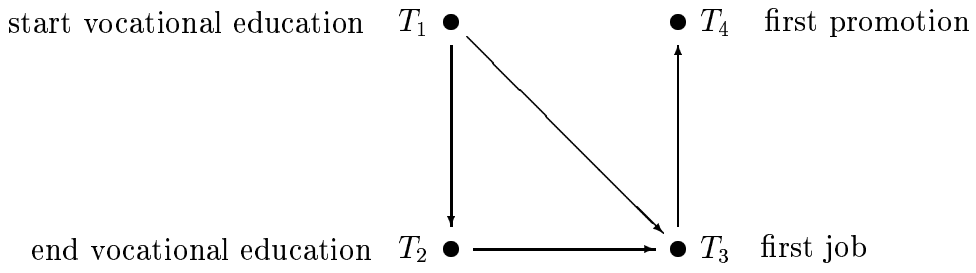
Proof:
The first part of the statement follows from the equivalence of the local Markov property for DAGs (2.3) and the factorization property (2.4). Further, the factors may be expressed through the hazard rates as in (2.8) due to the assumptions that ancestors have to occur before their descendants so that the conditional hazard given precedent survival times are well defined. □

With the above proposition we have that model specification may be performed by choosing appropriate regression models based on conditional hazard rates given the times of previous events. Some specific regression models are for instance addressed in Chapter 5. From these, statistical tests may be derived to select those previous events on which the hazard of a considered event depends in order to identify the set of parents for each vertex. Note that due to the assumption of a fixed order of the survival times we implicitly use an asymmetric dependence concept which determines the direction of the edges. This is extended and formalized in Chapter 3.

Example: Let us elucidate the foregoing proposition by a fictitious example. Consider for instance the following survival times relevant in life histories: $T_1 \mathrel{\hat{=}}$ 'beginning of vocational education', $T_2 \mathrel{\hat{=}}$ 'end of vocational education', $T_3 \mathrel{\hat{=}}$ 'beginning of first job', $T_4 \mathrel{\hat{=}}$ 'first promotion'. Although it has to be supposed that these variables

cannot be measured continuously, we treat them as if they were continuous. It is almost sure that the life history events described by these survival times occur in the stated order. We might then assume the following conditional (in)dependence structure: T_2 usually depends on T_1 since it cannot be assumed that the duration of the vocational education follows the same distribution for those who start late as compared to those who start early. Further, T_3 may depend on the duration of the education, thus on both T_1 and T_2. In contrast, it might be reasonable to assume that the time of the first promotion only depends on the performance in the job so that $T_4 \perp\!\!\!\perp \mathbf{T}_{\{1,2\}} | T_3$. These assumptions result in the conditional independence graph given in Figure 2.4. //

Figure 2.4: Fictitious conditional independence graph for $T_1 \hat{=}$ 'beginning of vocational education', $T_2 \hat{=}$ 'end of vocational education', $T_3 \hat{=}$ 'beginning of first job', $T_4 \hat{=}$ 'first promotion'.



Directed acyclic graphs for multivariate survival times can be generalized to include for instance time constant covariates, which would be regarded as prior to the other variables in the graph, as well as situations, where events do not occur in a fixed order as far as the corresponding survival times are independent given the preceding ones.

However, for even more general situations with recurrent events, no fixed order of occurrences, or a different number and type of events for the sample units, it becomes complicated to apply the above approach and to interpret the resulting graphs in a reasonable manner.

2.1.2 Dynamic interaction models

While the preceding subsection deals explicitly with event history data, in the present and the next subsection, we mainly consider time series. These are typically modeled with a discrete time parameter $\mathcal{T} = \mathbb{Z}$ so that it is straightforward to use chain graphs for the representation of the independence structure, where each discrete point in time forms a chain component. The same principle can be applied to longitudinal data including survival times (e.g. Klein et al., 1993).
An explicit theory for the modeling of multivariate time series through chain graphs based on the CG–distribution is developed by Lynggaard and Walther (1993). The corresponding models are called *dynamic interaction models.* In the following, we do not consider the problems arising from the specific distributional assumption nor the estimation and prediction task, but restrict ourselves to an outline of the idea of dynamic interaction models.

First, note that the underlying graphs have to take into account that at any discrete point t in time the same kind of variables are measured. This is formalized as follows.

Definition 2.1.10 *Dynamic chain graph*
Let $\mathcal{D}_T = \{V_0, \ldots, V_T\}$ be a dependence chain, where V_t are time indexed copies of a given set of vertices V. Then, $G^{\mathcal{D}_T} = (V^T, E^{\mathcal{D}_T})$ with $V^T = \bigcup_{t=0}^{T} V_t$ is called a *dynamic chain graph.* //

Regarding the independence structure of the multivariate time series, the authors propose to assume certain stationarity properties.

Definition 2.1.11 *Stationarity conditions*
Let P be the distribution of a multivariate stochastic process $\{\mathbf{Y}(t) |\ t \in \mathcal{T}\}$. This is said to be of *M-th order Markovian* if

$$P(\mathbf{Y}(t)|\mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) = P(\mathbf{Y}(t)|\mathbf{Y}(t-1), \ldots, \mathbf{Y}(t-M)) \qquad (2.9)$$

for all $t \in \mathcal{T}$. Further, P is said to have *stationary dynamics* if in addition to (2.9) it fulfills

$$P(\mathbf{Y}(t_1)|\mathbf{Y}(t_1-1), \ldots, \mathbf{Y}(t_1-M)) = P(\mathbf{Y}(t_2)|\mathbf{Y}(t_2-1), \ldots, \mathbf{Y}(t_2-M)) \quad (2.10)$$

for all $t_1 \neq t_2$, $t_1, t_2 \in \mathcal{T}$. //

The property of stationary dynamics implies for instance that the conditional means and covariances are time independent, whereas the marginal parameters may depend on time. Note that in case of longitudinal data from a large sample of individuals stationarity is typically not required. For time series analysis, however, it is a usual assumption. Combining the notion of a dynamic graph with the stationarity property yields the following definition of a dynamic graphical interaction model.

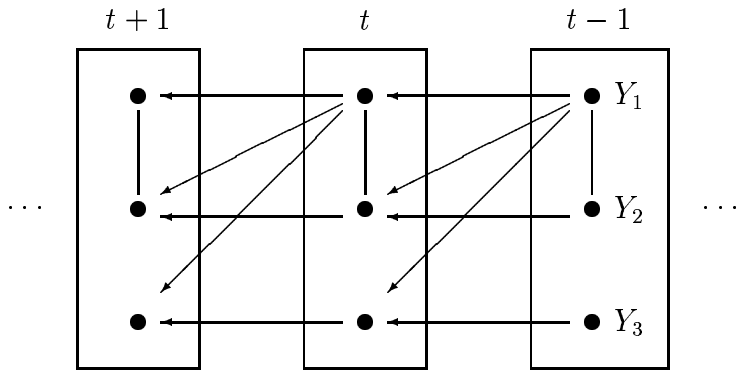
Definition 2.1.12 *Dynamic graphical interaction model*
Let $G^{\mathcal{D}_T} = (V^T, E^{\mathcal{D}_T})$ be a dynamic chain graph and $\{\mathbf{Y}(t) |\ t \in \mathcal{T}\}$ a multivariate stochastic process such that the components of $\mathbf{Y}(t)$ can be identified with V_t, i.e. $\mathbf{Y}(t) = \mathbf{Y}_{V_t}$. Consider a class of distributions $\mathcal{P}$, where any $P \in \mathcal{P}$ is of M-th order Markovian with $M < T$ and satisfies the stationarity condition (2.10). Then, $\mathcal{P}$ is called *dynamic graphical interaction model on* $G^{\mathcal{D}_T}$ if the joint density of any T successive points in time $t_1, \ldots, t_T \in \mathcal{T}$ given by

$$f(\mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_T}) = \prod_{t=t_1}^{t_T} f(\mathbf{y}_t | \mathbf{y}_{t-M}, \ldots, \mathbf{y}_{t-1}), \tag{2.11}$$

fulfills the chain graph Markov properties (2.5) induced by $G^{\mathcal{D}_T}$. //

Note that due to (2.9) and (2.10) it suffices to consider dynamic graphs with as many components as the order of the process. It further follows from the stationarity properties that every chain component is a copy of a fixed undirected graph and the dependence structure between the vertices V_t and $(V_{t-M}, \ldots, V_{t-1})$ is the same for all $t \in \mathcal{T}$. An example for the graph of a dynamic interaction model of first order is given in Figure 2.5.

Figure 2.5: Conditional independence graph of a dynamic interaction model with first order Markovian distribution.



The specific shape of the graph is one of the two main differences between general chain graphs and dynamic interaction models. The second difference lies in the stationarity assumptions which not only imply that the dependence structure between present and past is stationary but also that the conditional distributions are the same at different points in time as postulated by (2.10). The factorization (2.11) of the density may therefore be regarded as analogy to the one for DAGs (2.4) except that each factor constitutes the same multivariate regression model.

In the following, we describe some relations between dynamic interaction models and vector auto regressive (VAR) models. This also serves as prerequisite for the subsequent sections.

Definition 2.1.13 *VAR–model*
Let $\{\mathbf{Y}(t)|\ t \in \mathcal{T}\}$ be a q–variate stochastic process. A *VAR–model (of M-th order, VAR(M))* is given as a parametric class of distributions $\mathcal{P} = \{P_{a,A^1,\ldots,A^M} | a \in \mathbb{R}^{q\times 1}, A^m \in \mathbb{R}^{q\times q}, m = 1,\ldots,M\}$ such that under $P_{a,A^1,\ldots,A^M}$ it holds that

$$E(\mathbf{Y}(t)|\mathbf{Y}(t-1) = \mathbf{y}_{t-1},\ldots,\mathbf{Y}(t-M) = \mathbf{y}_{t-M}) = a + \sum_{m=1}^{M} A^m \mathbf{y}_{t-m}.$$

If, in addition, the conditional distribution $P_{a,A^1,\ldots,A^M}(\mathbf{Y}(t)|\mathbf{Y}(t-1) = \mathbf{y}_{t-1},\ldots,\mathbf{Y}(t-M) = \mathbf{y}_{t-M})$ is the q–variate normal distribution with mean $a + \sum_{m=1}^{M} A^m \mathbf{y}_{t-m}$ and covariance matrix Σ_ϵ, then we have a *Gaussian VAR(M)–model.* Further, if the class is restricted to those distributions, where Σ_ϵ is positive definite, we speak of a *regular Gaussian VAR(M)–model.* //

A VAR–model as defined above is obviously of M-th order Markovian with stationary dynamics. For simplicity we consider in the next corollary only VAR(1)–models. The restrictions induced by a dynamic graphical interaction model are then given as follows.

Corollary 2.1.14 *Dynamic graphical VAR–model*
Let $\mathcal{P} = \{P_{a,A}\}$ be a regular Gaussian VAR(1)–model. Let further $G^{\mathcal{D}_T} = (V^T, E^{\mathcal{D}_T})$ be a dynamic chain graph. Then, $P_{a,A} \in \mathcal{P}$ fulfills the chain graph Markov properties of $G^{\mathcal{D}_T}$ iff

(1) for all $t \in \mathcal{T}$, $j, k \in V_t$: $\{j,k\} \notin E^{\mathcal{D}_T} \Rightarrow (\Sigma_\epsilon^{-1})_{j,k} = 0$ and

(2) for all $t \in \mathcal{T}$, $j \in V_{t-1}$ and $k \in V_t$: $(j,k) \notin E^{\mathcal{D}_T} \Rightarrow (\Sigma_\epsilon^{-1} A)_{k,j} = 0$.

Proof:
The first condition implies that whenever there is no undirected edge between two vertices $j, k \in V_t$ within a block then

$$Y_j(t) \perp\!\!\!\perp Y_k(t) | \mathbf{Y}_{V_t \setminus \{j,k\}}, \mathbf{Y}(t-1). \tag{2.12}$$

This conditional independence holds because it is equivalent to the parameter restriction in (1) (Lynggaard and Walther, 1993, p. 81). The second condition states that whenever there is no directed edge from a vertex j in a preceding block to a vertex k in the following block, then we have

$$Y_k(t) \perp\!\!\!\perp Y_j(t-1) | \mathbf{Y}_{V_t \setminus \{k\}}, \mathbf{Y}_{V_{t-1} \setminus \{j\}}, \tag{2.13}$$

because this is in turn equivalent to the parameter restriction in (2) (Lynggaard and Walther, 1993, p. 81). To see this, note that by the properties of conditional distributions from multivariate Gaussian distributions we have $A = Cov(\mathbf{Y}(t), \mathbf{Y}(t-1)) Var(\mathbf{Y}(t-1))^{-1}$.
Finally, it can be shown that the pairwise conditional independencies (2.12) and (2.13) induced by missing edges in the graph are equivalent to the global chain graph Markov property for Gaussian distributions with positive definite covariance matrices (Frydenberg, 1990b). □

Due to the stationarity assumptions (2.9) and (2.10) the relevant chain graphs in the above corollary are actually those where the subgraphs $G_{V_{t-1} \cup V_t}$ have the same structure for all $t \in \mathcal{T}$.

Let us further point out a property which is important for the correct interpretation of dynamic chain graphs: It follows from condition (2) in the above corollary that in a VAR(1)–model the Matrix A alone is not responsible for the presence or absence of directed edges from earlier chain components to later ones. Instead we have that a zero entry $A_{k,j} = 0$ yields

$$Y_k(t) \perp\!\!\!\perp Y_j(t-1) \mid \mathbf{Y}_{V \setminus \{j\}}(t-1), \tag{2.14}$$

which differs from (2.13) by the conditioning set. For a missing directed edge from $j \in V_{t-1}$ to $k \in V_t$ it is instead required that in addition to $A_{k,j} = 0$ there exists no $i \in V$ such that $A_{i,j} \neq 0$ and $(\Sigma_\epsilon^{-1})_{k,i} \neq 0$, i.e. $Y_k(t)$ should not be partially correlated with $Y_i(t)$ given the past if $Y_i(t)$ is affected by $Y_j(t-1)$.
A deeper insight into this specific interpretational aspect can be gained by considering the so–called Granger–causality. According to this concept, a process is regarded as cause for another process if the precision of the prediction w.r.t. the latter decreases when the former is discarded. More formally, we have the following definition.

Definition 2.1.15 *Granger–causality*
Let Ω_t be the information set containing 'all the relevant information in the universe' up to and including time t. Consider the optimal (minimum mean squared error) one step prediction process of $\mathbf{Y}(t+1)$ given Ω_t, denoted by $\mathbf{Y}_t(t+1|\Omega_t)$, and let the corresponding forecast mean squared error be denoted by $\Sigma_{Y_t}(t+1|\Omega_t)$. The process $\mathbf{X} = \{\mathbf{X}(t)|t \in \mathcal{T}\}$ is said to Granger–cause $\mathbf{Y} = \{\mathbf{Y}(t)|t \in \mathcal{T}\}$ if the difference

$$\Sigma_{Y_t}(t+1|\Omega_t \backslash \{\mathbf{X}(s)|s \leq t\}) - \Sigma_{Y_t}(t+1|\Omega_t)$$

is positive definite. //

For VAR(M)–models it can be shown that Y_j is noncausal for Y_k based on the information in the whole process $\mathbf{Y}$ if and only if $A_{k,j}^m = 0$ for all $m = 1, \ldots, M$ (Tjøstheim, 1981). As argued above, this cannot be read off the dynamic interaction graph by verifying whether there are directed edges from vertices j in earlier blocks to k in later ones. An exception is given when the models are restricted by the assumption that Σ_ϵ is a diagonal matrix, i.e. there is no association among the components left after conditioning on the past. In Figure 2.5, for instance, we can see that Y_3 is no Granger–cause for $\mathbf{Y}_{\{1,2\}}$ since $\mathbf{Y}_{\{1,2\}}(t) \perp\!\!\!\perp Y_3(t)|\mathbf{Y}(t-1)$ as well as $\mathbf{Y}_{\{1,2\}}(t) \perp\!\!\!\perp Y_3(t-1)|\mathbf{Y}_{\{1,2\}}(t-1)$. Another possibility to read associations like Granger–causality off the graph is to define a new type of graph where missing arrows stand for conditional independencies like (2.14) instead of (2.13). This has been proposed by Andersson et al. (2001) for chain graphs and by Eichler (2000) for time series. The latter approach is addressed in Section 2.2.

Note that in the framework of Granger–causality, it makes no sense to consider whether a component Y_j, $j \in B \subset V$, is a Granger–cause for Y_k, $k \in B$, w.r.t. a subprocess $\mathbf{Y}_B$, since the definition explicitly states that all relevant information has to be included, i.e. in particular $\mathbf{Y}_{V \setminus B}$. Thus, Granger–causes have to be 'direct' causes, i.e. a process which affects another one only through mediation of other processes is not regarded as cause. In practice, however, it might be important to asses the effect even if it is only mediated through other processes because sometimes we cannot intervene in the 'direct' causes but only in the 'indirect' ones. This is considered in more detail in Chapter 4.

2.1.3 Partial correlation graphs for time series

This section discusses graphical models which are still based on the idea of reflecting conditional independencies. The difference to the above concepts consists in that the vertices now represent the components of multivariate time series. The corresponding models have been termed partial correlation graphs by Dahlhaus (2000) since the main tool for identifying the conditional independence structure only applies to linear dependencies. However, since these are equivalent to conditional independencies for the Gaussian case, we restrict ourselves to this distributional assumption. As shown by Dahlhaus (2000), the well–known results for undirected conditional independence graphs can be carried forward to the case where the vertices represent the components of multivariate time series. We now give a brief outline of this approach without going into the technical details (cf. Dahlhaus, 2000; Eichler, 1999). Again, we have a discrete time parameter $\mathcal{T} = \mathbb{Z}$.

Let $\mathbf{Y} = \mathbf{Y}_V = \{\mathbf{Y}_V(t) | t \in \mathcal{T}\}$ be a multivariate time series with components indexed by $V = \{1, \ldots, K\}$. In order to define partial (un)correlation of Y_j and Y_k, $j, k \in V$, we have to consider subprocesses where the linear effect of the remaining components has been removed: The *partial error process* $Y_{j|V\setminus\{j,k\}}(t)$ is given by

$$Y_{j|V\setminus\{j,k\}}(t) = Y_j(t) - \mu_j^* - \sum_{s\in\mathcal{T}} \sum_{i\in V\setminus\{j,k\}} \phi_j^*(t-s) Y_i(s),$$

where μ_j^* and $\phi_j^*(u)$ are the values minimizing

$$E\left(Y_j(t) - \mu_j - \sum_{s\in\mathcal{T}}\sum_{i\in V\setminus\{j,k\}} \phi_j(t-s)Y_i(s)\right)^2 .$$

Both, $\phi_j^*(u)$ and μ_j^* can be expressed through the spectral matrix $f^*(u)$ of the process $\mathbf{Y}$ (Brillinger, 1981).

Definition 2.1.16 *Partial correlation for time series*
Consider a multivariate time series $\{\mathbf{Y}_V(t)|t\in\mathcal{T}\}$. We say that Y_j and Y_k, $j,k\in V$, are *partially uncorrelated* given the remaining components $\mathbf{Y}_{V\setminus\{j,k\}}$ and write

$$\{Y_j\} \perp\!\!\!\perp \{Y_k\} \mid \{\mathbf{Y}_{V\setminus\{j,k\}}\}$$

if the partial error processes $Y_{j|V\setminus\{j,k\}}(t)$ and $Y_{k|V\setminus\{j,k\}}(t+u)$ are uncorrelated at all lags $u\in\mathbb{Z}$. //

It can be shown that partial correlation for time series as defined above satisfies the graphoid axioms, where for the intersection property we have to assume that the eigenvalues of the spectral matrix $f^*(u)$ are positive and bounded (Dahlhaus, 2000). This ensures that none of the components of $\mathbf{Y}$ are pure linear transformations of other components and may therefore be regarded as analogy to the assumption of a strictly positive density which ensures for conditional independence that the intersection property holds (cf. Appendix A; Lauritzen, 1996, p. 29). Once the validity of the axioms has been established, it is straightforward how to define partial correlation graphs for time series and to show the separation theorem.

Definition 2.1.17 *Partial correlation graph*
Let $G=(V,E)$ be an undirected graph and $\mathbf{Y}_V=\{\mathbf{Y}_V(t)|t\in\mathcal{T}\}$ a multivariate time series. Then, G is said to be the *partial correlation graph of* $\mathbf{Y}_V$ if the following implications hold

$$\{j,k\}\notin E \;\Rightarrow\; \{Y_j\} \perp\!\!\!\perp \{Y_k\} \mid \{\mathbf{Y}_{V\setminus\{j,k\}}\}$$

for all $j,k\in V$. //

Corollary 2.1.18 *Separation in partial correlation graphs*
Let $G = (V, E)$ be the partial correlation graph of $\mathbf{Y}_V$ and assume that the eigenvalues of its spectral matrix are positive and bounded. For disjoint subsets $A, B, C \subset V$ it holds

$$A \perp\!\!\!\perp_G B \mid C \;\Rightarrow\; \{\mathbf{Y}_A\} \perp\!\!\!\perp \{\mathbf{Y}_B\} \mid \{\mathbf{Y}_C\}.$$

Proof:
This can be shown exactly like the equivalence of pairwise and global Markov property for conditional independence graphs by exploiting the graphoid axioms (cf. Lauritzen, 1996; Dahlhaus, 2000). □

Note that the separation theorem also holds when the joint distribution of the time series it not Gaussian. In this case, the graph indeed only represents partial uncorrelation, whereas in the Gaussian case it represents conditional independencies. In general, the uncorrelatedness of $Y_{j|V\setminus\{j,k\}}(t)$ and $Y_{k|V\setminus\{j,k\}}(t+u)$ at all lags u is equivalent to corresponding zero entries in the partial cross spectra. This suggests a simple method for identifying the partial correlation graph of a multivariate time series: First, estimate the spectral matrix $f^*(u)$ by smoothing the periodogram and then test whether the functions occurring in the rescaled inverse are zero. An application may be found in Dahlhaus (2000).

If $\mathbf{Y}$ is a regular Gaussian VAR(M)–process with Σ_ϵ diagonal, it can be shown for its partial correlation graph G that an edge is absent, i.e. $\{j,k\} \notin E$, if and only if (1) $A^m_{j,k} = A^m_{k,j} = 0$ and (2) there exists no $i \in V$ such that $A^m_{i,k} \neq 0 \wedge A^m_{i,j} \neq 0$. Restating this in terms of Granger–causality reveals that $\{Y_j\} \perp\!\!\!\perp \{Y_k\} | \{\mathbf{Y}_{V\setminus\{j,k\}}\}$ if and only if (1) neither Y_j is a Granger–cause of Y_k nor vice versa and (2) for all other components Y_i, $i \in V\setminus\{j,k\}$, either Y_k or Y_j is no Granger–cause of Y_i. The latter condition can be justified by the same principle which motivates the moralization in DAGs (cf. the global directed Markov property given in Proposition 2.1.4): Conditioning on a common consequence of marginally independent Y_j and Y_k may induce an association.

A drawback of partial correlation graphs is that they contain no information on the dynamics of the process. In particular, since the time dimension plays an important

role in interpreting the missing edges, as described above in terms of Granger–causality, it is desirable to include this information in the graphical representation. This idea underlies the approach presented in the following section.

2.2 Causality graphs

Causality graphs have been introduced by Eichler (1999, 2000). As for partial correlation graphs, the vertices represent the components of a multivariate time series. But instead of considering only undirected graphs, the information on associations like Granger–causality is included by suitably defined directed edges. This motivates the name of these models, while the author is aware of the fact that (any version of) Granger–causality is only a measure for the association between lagged variables. As we have seen in the case of VAR–models, there are two basic types of association relevant for time series models. First, there is the dependence of the 'presence on the past'. Second, we have the correlation at any point in time that cannot be ruled out by conditioning on the past. Therefore, causality graphs allow for multiple edges between two vertices: undirected and directed ones, where the latter may have both directions and are then considered as two different edges. Thus, the general definition of a graph given in the first section has to be applied. In order to formalize the 'dependence of the presence on the past' we define the (strict) past $\bar{\mathbf{Y}}(t)$ of $\mathbf{Y}(t)$ as $\bar{\mathbf{Y}}(t) = \{\mathbf{Y}(s) | s < t\}$.

Definition 2.2.1 *Noncausality*
Consider a multivariate time series $\{\mathbf{Y}_V(t) | t \in \mathcal{T}\}$. We say that Y_j is *noncausal for* Y_k if for all $t \in \mathcal{T}$

$$Y_k(t) \perp\!\!\!\perp \bar{Y}_j(t) \mid \bar{\mathbf{Y}}_{V\setminus\{j\}}(t). \tag{2.15}$$

This is denoted by $Y_j \not\to Y_k\,[\mathbf{Y}_V]$, or briefly $j \not\to k\,[V]$.
Further, we say that Y_j and Y_k are *instantaneously noncausal* if for all $t \in \mathcal{T}$

$$Y_j(t) \perp\!\!\!\perp Y_k(t) \mid \bar{\mathbf{Y}}(t), \mathbf{Y}_{V\setminus\{j,k\}}(t). \tag{2.16}$$

This is denoted by $Y_j \not\sim Y_k\,[\mathbf{Y}_V]$, or briefly $j \not\sim k\,[V]$. //

Note that noncausality (2.15) equals the conditional independence statement (2.14) given earlier in the context of dynamic interaction models. It follows for Gaussian VAR–models that noncausality is determined by zero entries in the matrices A^m, $m = 1, \dots M$, whereas instantaneous noncausality can be parameterized by corresponding zero entries in the inverse conditional covariance matrix Σ_ϵ^{-1}.

A causality graph shows noncausality by missing directed edges and instantaneous noncausality by missing undirected edges as follows.

Definition 2.2.2 *Causality graph*
The causality graph of a multivariate time series $\{\mathbf{Y}_V(t)|t \in \mathcal{T}\}$ is given as a graph $G = (V, E)$ with vertices $V = \{1, \dots, K\}$ such that

$$(j,k) \notin E \;\Leftrightarrow\; Y_j \not\to Y_k\,[\mathbf{Y}_V]$$

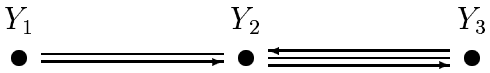
and

$$\{j,k\} \notin E \;\Leftrightarrow\; Y_j \not\sim Y_k\,[\mathbf{Y}_V]$$

for all $j, k \in V$, $j \neq k$. //

Example: An example for a causality graph is given in Figure 2.6. Here, the set of vertices is given by $V = \{1,2,3\}$ and the edges by $E = \{\{1,2\}, \{2,3\}, (1,2), (2,3), (3,2)\}$. This implies the noncausalities $\{2\} \not\to \{1\} \mid \{3\}$, $\{3\} \not\to \{1\} \mid \{2\}$,and $\{1\} \not\to \{3\} \mid \{2\}$. In addition, the missing undirected edge indicates an instantaneous noncausality, i.e. $\{1\} \not\sim \{3\} \mid \{2\}$. For a Gaussian VAR(1)–model the induced parameter restrictions are as follows: $A_{1,2} = A_{1,3} = A_{3,1} = 0$ and $(\Sigma_\epsilon)^{-1}_{1,3} = 0$. //

Figure 2.6: Example for a causality graph.



Causality graphs are thus an appropriate graphical representation for the conditional independence structures as they occur in Gaussian VAR(M)–models or for the partial correlation structure if the normal distribution is not justified. This is not

possible with the classical chain graphs as shown above in the context of dynamic interaction models since these usually have more edges and encode different conditional independencies due to the difference between (2.13) and (2.15). The same problem is considered by Andersson et al. (2001) who therefore define so–called *alternative Markov properties,* where the essential difference is that property (2.13) is replaced by (2.15) (see also the notion of summary graphs introduced by Cox and Wermuth, 1996, p. 204). Thus, causality graphs could be represented as chain graphs similar to dynamic interaction models by interpreting the edges according to the alternative Markov properties. Note, that a parametrization other than for the Gaussian case is quite difficult as indicated by Andersson et al. (2001). The notations and results of Andersson et al. can be applied to causality graphs in order to explore their relation to partial correlation graphs as well as their separation properties. Since an appropriate treatment of these properties requires substantial notational effort, we desist from going into details, here, and refer to Eichler (2000).

Chapter 3

Local independence

For general random structures involving the time dimension, we may translate the notion of noncausality considered in the foregoing chapter more appropriately into the statement that 'the past of Y_j is not informative for the presence of Y_k as long as we know the past of Y_k and of the other relevant components'. The concept of local independence presented in this chapter formalizes this idea for quite general processes, in particular for event history data, where an important aspect is that a *continuous time* dimension is considered. Processes with a continuous state space which vary steadily are obviously difficult to measure in continuous time so that a discrete time approach seems more reasonable. In contrast, point processes which provide the theoretic framework for event history data may indeed often be regarded as measured in continuous time and this entails several specific properties and problems not encountered in discrete time which are addressed in this and the following chapters.

In this chapter, local independence is presented with special regard to marked point processes. Thus, in the first section we review some basic concepts of this theory as far as needed to understand the subsequent parts. Some additional background is given in Appendix B. Section 3.2 gives the general definition of local independence. Its properties are further discussed along the lines of asymmetric irrelevance relations already introduced in the first chapter for δ–separation. Then, it is shown how local independence applies to multi–state processes and, finally, we consider the special case of Markov processes.

3.1 Basic concepts of event history analysis

Event history analysis is concerned with analyzing the occurrence and interaction of different events in time and their possible dependence on time varying or time constant covariates. Often, the events can be described as transitions between different states as for example the event of loosing a job, which can be regarded as change from the state employed to the state unemployed. Here, the set of interesting events or states is assumed to be finite and rather 'small' and we restrict ourselves to time constant covariates or to covariates which can themselves be modeled as events or changes of state. A suitable theoretical framework to describe this data situation is given by the so–called marked point processes (Arjas, 1989).

Definition 3.1.1 *Event history / marked point process*
Let $\mathcal{E} = \{e_1, \ldots, e_K\}$, $K < \infty$, denote the *mark space,* i.e. the set containing all events of interest, and $\mathcal{T}$ the *time space* in which the observations take place. If time is measured continuously we usually have that $\mathcal{T} = [0, \tau)$ or $\mathcal{T} = [0, \tau]$ where $\tau = \infty$ is possible.
Consider a measurable space $(\Omega, \mathcal{F})$. An *event history* or *marked point process* is then given by a set $\{(T_s, E_s) | s = 1, \ldots, S\}$ of pairs of random variables on $(\Omega, \mathcal{F})$ where $T_s \in \mathcal{T}$, $0 < T_1 < T_2 \ldots < T_S$, are the times of occurrences of the respective events $E_s \in \mathcal{E}$. //

Note that the number of possible events K and the actually observed set of events may be different for each individual, i.e. S may be random. The theory of marked point processes also applies to the more general situation of an infinite mark space $\mathcal{E}$. Since this is not considered here, we mainly refer to marked point processes as event histories.
Usually an additional vector $(X_1, \ldots, X_R)$ of time constant random variables is recorded describing the information available at time $t = 0$. Most of the following concepts and results can be generalized to this situation but we do not take it explicitly into account.

Examples: In survival analysis one is interested in the single event of death, so that the data for a single person consists of some covariates, for example age, gender and some kind of exposition, and the time of death which is often measured as time

from a defined starting point such as a surgery. In addition to observing the time of death one could measure the general health status with states good, medium, bad, and e.g. the smoking status with states smoker and non–smoker, where transitions from each status to the others and back are possible. //

There are further possibilities of formalizing the random structure of event histories, the same information as given by the times of the events is for example given by the time between the events with a defined starting time. It is evident that there are several ways of looking at event history data and consequently different possible concepts of dependence. Application of graphical models to this type of data is therefore not straightforward, as already illustrated in the previous chapter. Here, we pursue the approach that consists in assessing how one event affects the probability for the occurrence of another. Consider for example two events e_1 and e_2. If e_2 is more likely to occur (or less likely) when e_1 has already occurred than if it has not, one can say that the occurrence of e_2 depends on the occurrence of e_1. This intuitive notion of dependence of events has been called *local dependence* by Schweder (1970), generalized by Aalen (1987) and applied for instance by Aalen et al. (1980) and Gottard (1998). To make it more rigorous in the next section we need some notation and definitions concerning stochastic processes. Additional terminology and results can be found in Appendix B.

In event history analysis a special kind of stochastic processes plays an important role. These are multivariate counting processes where each component counts the number of occurrences of one specific event.

Definition 3.1.2 *Counting process*
A stochastic process N with with state space $\mathcal{S} = \{0, 1, 2, \ldots\}$, zero at time zero, paths which are right–continuous, where the left–limits exist, piecewise constant, and non–decreasing with jumps $\Delta N(t) = N(t) - N(t^-) = 1$ is called a *counting process.*
A multivariate process $\mathbf{N} = (N_1, \ldots, N_K)$, where the components N_k are counting processes, $k = 1, \ldots, K$, is called *multivariate counting process* if for the continuous time situation no two components jump at the same time. //

A given finite state process Y can be associated with a multivariate counting process $\mathbf{N} = (N_{qr}|q, r \in \mathcal{S})$ where $N_{qr}(t)$ denotes the number of transitions from state q to state r before or at time t. In terms of events $\mathcal{E} = \{e_1, \ldots, e_K\}$ occurring for a single observational unit the information about the event times is analogously contained in the multivariate counting process $\mathbf{N} = (N_1, \ldots, N_K)$ with $N_k(t)$ the number of events e_k that have occurred prior to or at time t. More formally this is defined as follows.

Definition 3.1.3 *Counting process for marked point processes*
Let $\{(T_s, E_s)|s = 1, \ldots, S\}$ be a marked point process as in Definition 3.1.1. Then, the associated *mark specific counting processes* are given by

$$N_k(t) = \sum_{s=1}^{S} \mathbf{1}\{T_s \leq t; E_s = e_k\}, \qquad k = 1, \ldots, K.$$

Alternatively we also write $N(t; e_k)$, $e_k \in \mathcal{E}$, which is easier to read when the subscript gets more complicated. //

The distribution of the marks and their occurrences is often described in terms of the infinitesimal development of the corresponding counting process given the past of the process, i.e. by their intensity process defined further below. This means that we need a notion suited to describe the evolution of the process in time. It is common to take this to be the *internal filtration* of a stochastic process Y which is given as $\{\mathcal{F}_t|t \geq 0\}$ with $\mathcal{F}_t = \sigma\{Y(s)|s \leq t\}$. Note that $\mathcal{F}_t \subset \mathcal{F}$, where $(\Omega, \mathcal{F})$ is the measurable space on which Y is defined. The left hand limit is given as $\mathcal{F}_{t-} = \sigma\{Y(s)|s < t\}$. In case of a marked point process we have that the internal filtration is given by $\mathcal{F}_t = \sigma\{(T_s, E_s)|T_s \leq t, E_s \in \mathcal{E}\}$ which is equal to $\sigma\{(N_1(s), \ldots, N_K(s))|s \leq t\}$. The internal filtration is sometimes augmented by an additional σ–field $\mathcal{F}_0$ generated by random variables realized at time $t = 0$, for instance by the vector $(X_1, \ldots, X_R)$ of time independent covariates. But if not stated otherwise, we use the internal filtration. Further, we assume that all considered filtrations satisfy the 'usual conditions' (cf. Appendix B).
A stochastic process can now be decomposed under quite general conditions into a $\mathcal{F}_t$–predictable part, i.e. a process that is predictable from the information in $\mathcal{F}_{t-}$, and a 'residual' part which forms a martingale. The former is called its compensator.

Definition 3.1.4 *Compensator*
Given a stochastic process Y on $(\Omega, \mathcal{F}, P)$ adapted to a filtration $\{\mathcal{F}_t\}$. Then, any $\mathcal{F}_t$–predictable, cadlag, and finite variation process Λ such that $M = Y - \Lambda$ is a local $\mathcal{F}_t$–martingale, zero at time zero, is called $\mathcal{F}_t$*–compensator* of Y. //

The decomposition $Y = \Lambda + M$ is called the *Doob–Meyer decomposition.* If such a process Λ exists it is unique. The conditions for the existence of a compensator are rather general and they hold for any counting processes (cf. Fleming and Harrington, 1991, p. 61).

Remark 3.1.5 *Compensator for counting processes*
Let N be a counting process. The compensator $\Lambda(t)$ is given by $\int_0^t \Lambda(ds)$ with

$$\Lambda(dt) = E(N(dt) \mid \mathcal{F}_{t^-}).$$

The compensator $\Lambda(t)$ may thus be regarded as the expected number of events that occur before or at time t given the strict pre–t history (cf. Fleming and Harrington, 1991, p. 38). In this sense, it provides a kind of 'short–term' prediction for the counting process. //

In the following we mainly consider processes where the compensator has a certain smoothness property.

Definition 3.1.6 *Local characteristic / intensity process*
Let Λ be the compensator of a stochastic process Y. The compensator Λ is *absolutely continuous,* if there exists a process $\lambda = \{\lambda(t) | t \in \mathcal{T}\}$ with

$$\Lambda(t) = \int_0^t \lambda(s)\, ds.$$

The process $\lambda(t)$ is called *local characteristic.* If Y is a counting process λ is also called the *intensity process* and Λ the *cumulative intensity.* //

As mentioned above, the compensator has to be $\mathcal{F}_t$–predictable, i.e. it is determined by all events that have occurred up to but not at time t. Thus, it is important to note that compensator and intensity process depend on the considered filtration. If the information about the past is restricted or extended worse or better 'predictions' are possible, respectively. This is the objective of the following remark.

Remark 3.1.7 *Innovation theorem*
Let $\mathcal{F}_t$ be the internal history of Y and Λ its $\mathcal{F}_t$–compensator. In general, if one considers a smaller filtration $\{\hat{\mathcal{F}}_t | t \in \mathcal{T}\}$ with $\hat{\mathcal{F}}_t \subset \mathcal{F}_t$ for all $t \in \mathcal{T}$, i.e. if less information about the past is included than given by the process Y itself, the corresponding $\hat{\mathcal{F}}_t$–compensator $\hat{\Lambda}$ may differ from Λ. By the *innovation theorem* we have the heuristic relation

$$\hat{\Lambda}(dt) = E(\Lambda(dt)|\hat{\mathcal{F}}_{t^-}) \quad \forall\, t \in \mathcal{T}.$$

For the exact statement as well as the required assumptions concerning the involved processes and filtrations see Brémaud (1981, pp. 83). //

In case of a marked point process with mark specific counting processes we also have mark specific compensators $\Lambda_k(t)$ and intensities $\lambda_k(t)$, $e_k \in \mathcal{E}$. Again, we write $\Lambda(t; e_k)$ and $\lambda(t; e_k)$ in case that the subscript gets more complicated. All the above notions apply analogously to these mark specific processes. But note that the processes are defined w.r.t. to the internal history of the whole multivariate process $\{N_k | k = 1, \ldots, K\}$ and not w.r.t. the internal history of the single mark specific counting process N_k which would contain less information on the past. If one is interested in the latter the innovation theorem has to be applied. More generally, this theorem may be used to determine the intensity of an event given that only a subset of the previous events have been observed. Note that this is a special filtering problem (Brémaud, 1981, pp. 83) which has been treated for marked point processes by Arjas et al. (1992). Let $\mathcal{E}$ be the set of all possible events and $\hat{\mathcal{E}} \subset \mathcal{E}$ the set of all observable events. Consequently, $\hat{\mathcal{F}}_t = \sigma\{(T_s, E_s) | T_s \leq t, E_s \in \hat{\mathcal{E}}\}$ is the filtration generated by the observable marked points and $\hat{\mathcal{F}}_t \subset \mathcal{F}_t, t \in \mathcal{T}$. In oder to determine the $\hat{\mathcal{F}}_t$–intensities $\hat{\lambda}(t; \hat{e})$, $\hat{e} \in \hat{\mathcal{E}}$, we need to specify the distribution of the unobserved events given the observed. This is simplified by introducing the additional notion of a history process.

Definition 3.1.8 *History process*
Consider a marked point process $\{(T_s, E_s) | s = 1, \ldots, S\}$ with mark space $\mathcal{E} = \{e_1, \ldots, e_K\}$. The *pre–t history process* H_t is defined as the set of those marked points which occurred before or at time t, i.e.

$$H_t = \{(T_s, E_s) \mid T_s \leq t, s = 1, \ldots, S\}.$$

As for filtrations, H_{t^-} denotes the *strict pre–t history process.* Additionally, H_t^A, $A \subset \{1, \ldots, K\}$, defined as

$$H_t^A = \{(T_s, E_s) \mid T_s \leq t \text{ and } \exists\, k \in A : E_s = e_k, s = 1, \ldots, S\}$$

denotes *the history process restricted to the events in A.*
Any set of marked points for which it holds that $t_s = t_u, s \neq u$, implies $e_s = e_u$ can be a history, i.e. a realization of H_t. Let thus $I\!H$ be the set of all histories. We denote the corresponding σ–field of Borel sets by $\mathcal{H}$. //

Note that the different filtrations can be regarded as being generated by the history process, i.e. $\mathcal{F}_t^A = \sigma\{H_t^A\}$, $A \subset \{1, \ldots, K\}$, such that if $\hat{\mathcal{E}} = \{e_k \in \mathcal{E} | k \in A\}$ then $\hat{\mathcal{F}}_t = \sigma\{H_t^A\}$.

As outlined by Arjas et al. (1992) the basic problem in determining $\hat{\lambda}(t; \hat{e})$ consists in calculating the conditional distribution of the underlying pre–t history given the observations in $\hat{\mathcal{F}}_t$. For this purpose, let

$$\hat{\pi}_t(B) = P(H_t \in B \mid \hat{\mathcal{F}}_t), \quad B \in \mathcal{H}, t > 0.$$

As stated by the innovation theorem, the intensities of the observed events can now be written as

$$\hat{\Lambda}(dt; \hat{e}) = \int_{I\!H} \hat{\pi}_{t^-}(dH)\Lambda(dt, \hat{e}), \quad \hat{e} \in \hat{\mathcal{E}},$$

which in the absolutely continuous case reduces to

$$\hat{\lambda}(t; \hat{e}) = \int_{I\!H} \hat{\pi}_{t^-}(dH)\lambda(t, \hat{e}), \quad \hat{e} \in \hat{\mathcal{E}}. \tag{3.1}$$

Note that in both formulae above, Λ and λ, respectively, depend on the integrand H since they are conditional on the strict pre–t history. The conditional distribution $\hat{\pi}$ may be calculated either explicitly using the Bayes formula or via an integral equation. The former requires the specification of a parametric model. The latter shows how the probability $\hat{\pi}$ is updated whenever an observable marked point $(t_s, \hat{e}_s)$, $\hat{e}_s \in \hat{\mathcal{E}}$, occurs. This result is quoted in the following theorem for the absolutely continuous case.

Theorem 3.1.9 *Partially observed marked point process*
Consider a marked point process and its corresponding counting processes. Assume

that the sample paths of $N = \sum_{k=1}^{K} N_k$ are a.s. finite–valued, $\Lambda = \sum_{k=1}^{K} \Lambda_k$ is absolutely continuous and that the sample paths of $\lambda = \sum_{k=1}^{K} \lambda_k$ are uniformly bounded on finite intervals. Then, the conditional distribution $\hat{\pi}_t$ is uniquely determined through the recursive equation

$$\begin{aligned}\hat{\pi}_t(B) &= \mathbf{1}\{\emptyset \in B\} \\ &\quad + \int_0^t \int_{I\!H} \hat{\pi}_{s^-}(dH) \sum_{e \in \mathcal{E}} [\mathbf{1}\{H \cup \{(s,e)\} \in B\} - \mathbf{1}\{H \in B\}] \lambda(s;e) ds \\ &\quad + \sum_{\hat{e} \in \hat{\mathcal{E}}} \int_0^t [\hat{Z}_s(\hat{e}) - \hat{\pi}_{s^-}(B)](N(ds;\hat{e}) - \hat{\lambda}(s;\hat{e})\, ds), \qquad B \in \mathcal{H} \end{aligned} \tag{3.2}$$

where

$$\hat{Z}_t(\hat{e}) = \frac{\int_{I\!H} \hat{\pi}_{t^-}(dH) \mathbf{1}\{(H \cup \{(t,\hat{e})\} \in B\} \lambda(t;\hat{e})}{\int_{I\!H} \hat{\pi}_{t^-}(dH) \lambda(t;\hat{e})}, \quad \hat{e} \in \hat{\mathcal{E}}.$$

Proof: Arjas et al. (1992) □

The recursive equation (3.2) can be understood as follows: Assume that the probability $\hat{\pi}_{t_0}(B)$ at a specific time t_0 is known. If we continue to observe the partial marked point process then we collect further information for instance when observing the next event $\hat{e} \in \hat{\mathcal{E}}$ at time $t_1 > t_0$. This additional information consists of knowing that nothing observable has occurred between t_0 and t_1 and the event at t_1. Consequently the probability for B is updated on the one hand due to the non–occurrence in (t_0, t_1) and on the other hand due to the occurrence of $\{(t_1, \hat{e})\}$. The *innovation gain,* i.e. the change in the intensity due to observing what happens at t, is given as the difference $\hat{Z}_t(e) - \hat{\pi}_{t^-}(B)$ weighted according to whether an event occurs at t or not, $t_0 < t \leq t_1$. Note that $N(dt;\hat{e})$ in (3.2) equals one if and only if the event $\hat{e}$ occurs at t and zero otherwise.

Let us consider the updating of $\hat{\pi}_t(B)$ in more detail assuming first (1) that no observable event occurs at t and then (2) that an observable event occurs at t. In case of (1) we have that $T_n < t < T_{n+1}$, where T_n denotes the time of the latest observable event before t. Then, it follows from Theorem 3.1.9 that

$$\begin{aligned}\hat{\pi}_t(B) &= \hat{\pi}_{T_n}(B) \\ &\quad + \int_{T_n}^t \int_{I\!H} \hat{\pi}_{s-}(dH) \sum_{e \in \mathcal{E}} [\mathbf{1}\{H \cup \{(s,e)\} \in B\} - \mathbf{1}\{H \in B\}] \lambda(s;e)\, ds \end{aligned}$$

$$
\begin{aligned}
&- \int_{T_n}^{t} \int_{I\!H} \hat{\pi}_{s-}(dH) \sum_{\hat{e} \in \hat{\mathcal{E}}} \mathbf{1}\{H \cup \{(s,\hat{e})\} \in B\} \lambda(s;\hat{e})\, ds \\
&+ \int_{T_n}^{t} \hat{\pi}_{s-}(B) \sum_{\hat{e} \in \hat{\mathcal{E}}} \hat{\lambda}(s;\hat{e})\, ds \\
= \;& \hat{\pi}_{T_n}(B) \\
&+ \int_{T_n}^{t} \int_{I\!H} \hat{\pi}_{s^-}(dH) \left\{ \sum_{e \in \mathcal{E} \setminus \hat{\mathcal{E}}} \mathbf{1}\{H \cup \{(s,e)\} \in B\} \lambda(s;e) \right. \\
&\left. -\mathbf{1}\{H \in B\} \lambda(s) + \hat{\pi}_{s-}(B) \sum_{\hat{e} \in \hat{\mathcal{E}}} \hat{\lambda}(s;\hat{e}) \right\} ds, \qquad (3.3)
\end{aligned}
$$

where we exploited (3.1) which occurs in the denominator of $\hat{Z}_t(e)$ and $\lambda(t) = \sum_k \lambda_k(t)$. The last equation (3.3) shows how the non–occurrences of observable events contribute to the conditional probability for B.
In case of (2), assume that the event $\hat{e}$ occurs at $t = T_{n+1}$. We then have

$$
\hat{\pi}_{T_{n+1}}(B) \;=\; \frac{\int_{I\!H} \hat{\pi}_{T_{n+1}^-}(dH) \mathbf{1}\{H \cup \{(\hat{T}_{n+1},\hat{e})\} \in B\} \lambda(T_{n+1};\hat{e})}{\int_{I\!H} \hat{\pi}_{T_{n+1}^-}(dH) \lambda(T_{n+1};\hat{e})}, \qquad (3.4)
$$

which can be regarded as a continuous time version of the Bayes formula.

In the following section we define local independence for marked point processes with the aim to describe the independence structure among events. The above considerations are used to find conditions for the independence structure among a subset of events to stay the same when the remaining events are discarded. The definition of local independence is based on the mark specific counting processes and compensators and thus assumes that they exist. Further basic results concerning likelihood based inference for marked point processes are given in Chapter 5.

3.2 Local independence

Following the definition given by Aalen (1987) the above mentioned intuitive notion of dependence of events is now formulated for general stochastic processes allowing for a Doob–Meyer decomposition. We first consider a bivariate process $\mathbf{Y} = (Y_1, Y_2)$ and the respective componentwise $\mathcal{F}_t$–compensators Λ_k with local characteristics λ_k,

$k = 1, 2$. Note that the latter two are assumed to be w.r.t. the internal filtration of the whole bivariate process $\mathbf{Y}$. Let $\mathcal{F}_t^k = \sigma\{Y_k(s)|s \leq t\}$, $k = 1, 2$, be the componentwise filtrations. As mentioned above, the $\mathcal{F}_t^k$–compensators $\tilde{\Lambda}_k$ may in general differ from Λ_k since the former are restricted to the information given by the history of the k-th component alone instead of both components. However, the equality $\tilde{\Lambda}_k$=Λ_k implies that the other component carries no information for the infinitesimal development of Y_k given its own past. This is the key to the following definition which was originally given by Aalen (1987) but restricted to the situation where an intensity process exists. We state it for the more general case of a compensator which is not necessarily absolutely continuous.

Definition 3.2.1 *Local independence (bivariate)*
Let Y_1 and Y_2 be two stochastic processes on $(\Omega, \mathcal{F}, P)$. Let further $\mathcal{F}_t^k = \sigma\{Y_k(s)|0 \leq s \leq t\}$, $k = 1, 2$, and $\mathcal{F}_t = \mathcal{F}_t^1 \vee \mathcal{F}_t^2 = \sigma\{\mathcal{F}_t^1 \cup \mathcal{F}_t^2\}$, $t \in \mathcal{T}$. Assume that the processes permit a Doob–Meyer decomposition with respect to $\mathcal{F}_t$ and that the $\mathcal{F}_t$–martingales $Y_k - \Lambda_k$, $k = 1, 2$, are orthogonal.
Then, Y_1 is said to be *locally independent* of Y_2 over $\mathcal{T}$ if $\Lambda_1(t)$ is measurable w.r.t $\mathcal{F}_t^1$ for all $t \in \mathcal{T}$. Otherwise we speak of local dependence. //

The process Y_1 being locally independent of Y_2 is symbolized by $Y_2 \not\to Y_1$ which can also be read as Y_2 being irrelevant for Y_1. The interpretation of local independence as irrelevance property is discussed in the next section. As already suggested earlier, local independence may be regarded as generalization of Granger–noncausality to the continuous time situation by interpreting the compensator as short–term prediction (cf. Remark 3.1.5). A brief discussion of this aspect is given by Florens and Fougère (1996).

As can easily be checked, local independence needs be neither symmetric, reflexive nor transitive. However, we make the following assumption.

Assumption 3.2.2 *Reflexivity*
Since in most practical situations a process depends at least on its own past we only consider stochastic processes where local dependence is reflexive. //

With regard to the interpretation of local independence one could heuristically say that if $Y_2 \not\to Y_1$ then the presence of Y_1 is conditionally independent of the past

of Y_2 given the past of Y_1 symbolizing this by $Y_1(t) \perp\!\!\!\perp \mathcal{F}^2_{t^-} | \mathcal{F}^1_{t^-}$. However, this is in general a stronger property than local independence but it holds for marked point processes as shown in the next chapter. The relation between local and conditional independence is also explored in Section 3.2.3 for Markov processes.
Note that in special situations there are trivial forms of local dependence. In survival analysis, for instance, each internal covariate process is locally dependent of the survival process since failure implies that the intensities for any further transition are zero.

Let us now consider local independence for counting processes which are of special interest when considering marked point processes. If $N_2 \not\to N_1$, where N_k, $k = 1, 2$, are counting processes, we have that the compensator Λ_1 of N_1 is a function of N_1 itself, and not of N_2, i.e. the 'infinitesimal prediction' of the current development of N_1 cannot be improved by including N_2. In the absolutely continuous case we even have by the innovation theorem for the $\mathcal{F}^1_t$–intensity $\tilde{\lambda}_1(t)$ of N_1 that $\tilde{\lambda}_1(t) = \lambda_1(t)$ if $N_2 \not\to N_1$. Thus, the intensity process remains the same if the included information is changed from $\mathcal{F}_t$ to $\mathcal{F}^1_t$ which contains only the information about the development of N_1, i.e. about whether and when event e_1 has occurred up to time $t \in \mathcal{T}$.

An important aspect of Definition 3.2.1 is the assumption that the martingales $M_1 = Y_1 - \Lambda_1$ and $M_2 = Y_2 - \Lambda_2$ are orthogonal. The implications are elucidated in the following remark.

Remark 3.2.3 *Assumption of orthogonality*
The assumption of orthogonality means that the innovations of the two processes are unrelated given the past. This can be seen as follows. Formally, we have by this assumption that $\langle M_1, M_2 \rangle = 0$ so that $Cov(M_1(dt), M_2(dt) | \mathcal{F}_{t-}) = 0$. From this it follows for the increments of the counting processes $N_1(dt)$ and $N_2(dt)$ that

$$\begin{aligned} Cov(N_1(dt), N_2(dt) | \mathcal{F}_{t^-}) &= E(N_1(dt)N_2(dt) | \mathcal{F}_{t^-}) - E(N_1(dt) | \mathcal{F}_{t^-}) E(N_2(dt) | \mathcal{F}_{t^-}) \\ &= E(\Lambda_1\Lambda_2(dt) + \Lambda_1 M_2(dt) + \Lambda_2 M_1(dt) + M_1 M_2(dt) \mid \mathcal{F}_{t^-}) - \Lambda_1\Lambda_2(dt) \\ &= \Lambda_1(dt) E(M_2(dt) \mid \mathcal{F}_{t^-}) + \Lambda_2(dt) E(M_1(dt) \mid \mathcal{F}_{t^-}) + E(M_1 M_2(dt) \mid \mathcal{F}_{t^-}) \\ &= 0, \end{aligned}$$

where the last transformation is due to the predictability of the compensators, the properties of martingales and the orthogonality assumption. Since $N_k(dt)$, $k = 1, 2$, can be regarded as Bernoulli variables and since it holds that two Bernoulli variables are stochastically independent iff their covariance is zero, we have that the increments of the considered counting processes are conditionally independent given the history $\mathcal{F}_{t-}$. The orthogonality condition seems reasonable, since one could hardly speak of independence if the same innovations fed the two processes even if the compensators are independent. In the continuous time situation and assuming the existence of an intensity process, i.e. for Λ absolutely continuous, the orthogonality of $M_k = N_k - \Lambda_k$, $k = 1, 2$, is attained precisely when the processes cannot jump simultaneously. This can e.g. be made sure by a suitable choice of the mark space for the corresponding marked point process, i.e. if events may occur at the same time with a positive intensity, then a corresponding mark has to be added. In this case one could of course not assume any kind of independence between these processes. //

Example: An example for local independence has be investigated by (Aalen et al., 1980). The two events were (1) first occurrence of a specific kind of chronic skin disease and (2) beginning of the menopause. The findings indicate, as expected, that 'menopause' is locally independent of 'skin disease' but not vice versa. The authors found enough evidence for concluding that the skin disease is more probable for women who have started their menopause. //

The extension of the definition of local independence to a collection of more than two processes deserves special attention since it has to be made clear which history a subprocess is independent of. Note that for a multivariate process $\mathbf{Y} = (Y_1, \ldots, Y_K)$ the $\mathcal{F}_t$–compensator is itself vector–valued and given by $\Lambda = (\Lambda_1, \ldots, \Lambda_K)$, where Λ_k is the $\mathcal{F}_t$–compensator of Y_k. The compensator Λ_A of a subprocess $\mathbf{Y}_A$, $A \subset \{1, \ldots, K\}$, is given in a similar manner.

Definition 3.2.4 *Local independence (multivariate)*
Consider a multivariate process $\mathbf{Y} = (Y_1, \ldots, Y_K)$ with individual filtrations $\mathcal{F}_t^k = \sigma\{Y_k(s) | 0 \leq s \leq t\}$ and joint filtrations $\mathcal{F}_t^A = \bigvee_{k \in A} \mathcal{F}_t^k$ corresponding to subprocesses $\mathbf{Y}_A$, $A \subset V = \{1, \ldots, K\}$. For $A, B, C \subset V$ assume that $\mathcal{F}_t^{A \cup B \cup C}$–compensators Λ_A and Λ_B exist such that $\mathbf{Y}_A - \Lambda_A$ and $\mathbf{Y}_B - \Lambda_B$ are orthogonal

$\mathcal{F}_t^{A\cup B\cup C}$–martingales.
We then say that a *subprocess* $\mathbf{Y}_B$ *is locally independent of* $\mathbf{Y}_A$ *given* $\mathbf{Y}_C$ if all $\mathcal{F}_t^{A\cup B\cup C}$–compensators Λ_k, $k \in B$, are measurable with respect to $\mathcal{F}_t^{B\cup C}$. This is denoted by $\mathbf{Y}_A \not\to \mathbf{Y}_B|\mathbf{Y}_C$ or briefly $A \not\to B|C$. Otherwise, $\mathbf{Y}_B$ is *locally dependent* of $\mathbf{Y}_A$ given $\mathbf{Y}_C$, i.e. $A \to B|C$. //

We have, again, as heuristic interpretation of $A \not\to B|C$ that the prediction of the infinitesimal development of $\mathbf{Y}_B$ cannot be improved by including information on the past of $\mathbf{Y}_A$ given that the past of $\mathbf{Y}_B$ and $\mathbf{Y}_C$ is already known, i.e. A is in a sense discussed below irrelevant for B given C. Note that the past of the considered process, here $\mathcal{F}_{t^-}^B$, is always included in the condition which is justified by the reflexivity Assumption 3.2.2.

3.2.1 Local independence as irrelevance relation

The interpretation of the set C in the local independence statement $A \not\to B|C$ is important for the graphical representation of local independencies as addressed in Chapter 4. If $A \not\to B|C$ we would expect that in a graphical representation C is a separating set for A and B. As shown in that chapter, this can be achieved through directed graphs and δ–separation. The discussion of the properties of such a graphical representation is simplified by first considering in more detail the properties of local independence as an irrelevance relation.

An obvious way to translate local independence as irrelevance relation is by letting $A \,\text{IR}\, B|C$ stand for $A \not\to B|C$, $A, B, C \subset V$. In this case we have that $'A \leq B' \hat{=} A \subset B$, $'A \vee B' \hat{=} A \cup B$ and $'A \wedge B' \hat{=} A \cap B$. Local independence does not satisfy all the asymmetric graphoid properties. For the decomposition as well as for intersection we need further conditions. But let us first consider those properties that always hold.

Proposition 3.2.5 *Graphoid properties of local independence*
The following properties hold for local independence according to Definition 3.2.4:

(1) left redundancy,

(2) left decomposition,

(3) left and right weak union,

(4) left and right contraction,

(5) right intersection.

Proof:

(1) Left redundancy holds since obviously the $\mathcal{F}_t^{A\cup B}$–compensators of $\mathbf{Y}_B$ are $\mathcal{F}_t^{A\cup B}$–measurable, i.e. if the past of $\mathbf{Y}_A$ is known, then the past of $\mathbf{Y}_A$ is of course irrelevant.

(2) Left decomposition holds since the $\mathcal{F}_t^{A\cup B\cup C}$–compensators $\Lambda_k(t)$, $k \in B$, are $\mathcal{F}_t^{B\cup C}$–measurable by assumption so that the same must hold for the $\mathcal{F}_t^{B\cup C\cup D}$–compensators $\Lambda_k(t)$, $k \in B$, for $D \subset A$.

(3) Left and right weak union also trivially hold since adding information on the past of components that are already uninformative (left) or included (right) does not change the compensator.

(4) Left contraction holds since we have that the $\mathcal{F}_t^{A\cup B\cup C\cup D}$–compensators Λ_k, $k \in B$, are by assumption $\mathcal{F}_t^{A\cup B\cup C}$–measurable and these are again by assumption $\mathcal{F}_t^{B\cup C}$–measurable.

Right contraction is not so easily seen. Here, we have to show that the $\mathcal{F}_t^{A\cup B\cup C\cup D}$–compensators Λ_k, $k \in B \cup D$, are $\mathcal{F}_t^{B\cup C\cup D}$–measurable. This holds for $k \in D$ by assumption. For $k \in B$, consider first the case that $D \not\to B|(A \cup C)$. Then, we have with left contraction that $(A \cup D) \not\to B|C$ showing that the assertion is true. In case that $D \to B|(A \cup C)$, we have that the assumption of $A \not\to B|C$ implies that D cannot affect A because if it was a 'common cause' of A and B then there could not be this local independence. Further, C cannot contain any 'common consequences' of A and D because this would in turn induce a dependence between A and D contradicting $A \not\to B|C$. Further, we already know that A does not affect D. Therefore, the additional information $\mathcal{F}_{t-}^D$ on the past of process $\mathbf{Y}_D$ cannot generate an effect of $\mathbf{Y}_A$ on $\mathbf{Y}_B$.

(5) The property of right intersection can be checked by noting that in the definition of local independence (Definition 3.2.4) the filtration w.r.t. which the intensity process should be measurable is always generated at least by the process itself. □

Note that with the above proposition we have that (1.1) holds, i.e. $A \not\to B|C \Leftrightarrow A\backslash C \not\to B|C$. In contrast, it is clear by the definition of local independence that

right redundancy does not hold because otherwise any process would always be locally independent of any other process given its own past. It follows that property (1.2) does not hold. This is also clear because $A \not\to B\backslash C|C$ does not contain any information about $\Lambda_{B\cap C}$ at all, so that it cannot imply $A \not\to B|C$. It is easily seen that the converse is always true, i.e. $A \not\to B|C \Rightarrow A \not\to B\backslash C|C$. This is in turn a special case of the following result which parallels Lemma 1.2.12.

Lemma 3.2.6 *Special case of right decomposition for local independence*
The following implication holds for local independence:

$$A \not\to B \mid C,\ D \subset B \Rightarrow A \not\to D \mid (C \cup B)\backslash D$$

Proof:
If the $\mathcal{F}^{A\cup B\cup C}$–compensator Λ_B is $\mathcal{F}^{B\cup C}$–measurable, so is any subprocess Λ_D, $D \subset B$. □

Let us now consider the property of left intersection. For symmetric irrelevance relations it is the property of intersection that makes a semigraphoid a graphoid. Its importance for the equivalence of pairwise, local and global Markov properties in undirected conditional independence graphs is well–known (cf. Lauritzen, 1996). As shown later, it is of similar importance for the local independence graphs. In the following proposition we formulate conditions for left intersection to hold.

Proposition 3.2.7 *Left intersection for local independence*
Consider the irrelevance relation defined through local independence as formulated in Definition 3.2.4. Under the assumption that

$$\mathcal{F}_t^A \cap \mathcal{F}_t^B = \mathcal{F}_t^{A\cap B} \quad \forall\, A, B \subset V, \quad \forall\, t \in \mathcal{T} \tag{3.5}$$

the property of left intersection according to Definition 1.2.6 holds.

Proof:
Left intersection assumes that the $\mathcal{F}_t^{A\cup B\cup C}$–compensators $\Lambda_k(t)$, $k \in B$, are $\mathcal{F}_t^{B\cup C}$– as well as $\mathcal{F}_t^{A\cup B}$–measurable. With (3.5) we get that they are $\mathcal{F}_t^{B\cup(A\cap C)}$–measurable which yields the desired result. □

Property (3.5) formalizes the intuitive condition that the considered subprocesses of the system are different enough to ensure that common events are necessarily due to common components. Note that under (3.5), property (1.3), too, holds for local independence.

The remaining property, right decomposition, requires special consideration because it makes a statement about the irrelevance of a process $\mathbf{Y}_A$ after discarding part of the possibly *relevant* information $\mathbf{Y}_{B\setminus D}$. If the irrelevance of $\mathbf{Y}_A$ is due to knowing the past of $\mathbf{Y}_{B\setminus D}$ then it will not necessarily be irrelevant anymore if the latter is discarded. Right decomposition therefore only holds under specific restrictions on the relation between the potentially irrelevant process and the relevant process to be discarded. The first restriction exploits property (3.5) to show that under an additional condition right decomposition also holds for local independence.

Proposition 3.2.8 *First condition for right decomposition of local independence*
Given (3.5) then right decomposition according to Definition 1.2.6 holds for local independence if the following additional conditions are true:

$$(B \cap A)\setminus(C \cup D) = \emptyset \quad \text{and} \quad B \not\to D \mid A \cup C. \tag{3.6}$$

Proof:
Due to (1.1) we can assume that $A \cap C = \emptyset$. Let $B^* = B\setminus D$. Note first that because of Lemma 3.2.6 we have $A \not\to B|C \Rightarrow A \not\to D|C \cup B^*$. By assumption and left decomposition it holds that $B^* \not\to D|A \cup C$. Thus, we can apply (1.3) which yields $A \cup B^* \not\to D|C$, where $A \cap B^* = \emptyset$ has been used. With left decomposition we get the desired result. □

Another situation where right decomposition holds is given in the following proposition and is restricted to counting processes.

Proposition 3.2.9 *Second condition for right decomposition of local independence*
Consider a marked point process with the assumptions of Theorem 3.1.9 and the irrelevance relation defined through local independence as formulated in Definition 3.2.4. Let $A, B, C, D \subset V$ with $(B \cap A)\setminus(C \cup D) = \emptyset$. Right decomposition, i.e.

$$A \not\to B \mid C,\ D \subset B \ \Rightarrow\ A \not\to D \mid C$$

(cf. Definition 1.2.6), holds under the conditions that

$$B \not\to A \backslash (C \cup D) \mid (C \cup D) \tag{3.7}$$

and

$$A \not\to \{k\} \mid C \cup B \text{ or } B \not\to \{k\} \mid (C \cup D \cup A) \tag{3.8}$$

for all $k \in C \backslash D$.

Proof:
Let us first consider the simple situation where $A = \{a\}, B = \{b, d\}, C = \{c\}, D = \{d\}$, $V = A \cup B \cup C \cup D$. From the assumptions it follows for the $\mathcal{F}_t^{abcd}$–intensities that

- $\lambda_a(t)$ is $\mathcal{F}_t^{acd}$–measurable,
- $\lambda_b(t)$ is $\mathcal{F}_t^{bcd}$–measurable,
- $\lambda_c(t)$ is either $\mathcal{F}_t^{acd}$– or $\mathcal{F}_t^{bcd}$–measurable,
- and $\lambda_d(t)$ is $\mathcal{F}_t^{bcd}$–measurable.

We have to show that the $\mathcal{F}_t^{acd}$–intensity $\hat{\lambda}_d(t)$ is $\mathcal{F}_t^{cd}$–measurable. With the innovation theorem and (3.1) it holds that

$$\hat{\lambda}_d(t)dt = \int_{I\!H} \hat{\pi}_{t^-}(dH)\lambda_d(t)dt,$$

where $\lambda_d(t)dt = P(N(dt;d) = 1|\mathcal{F}_{t^-}) = P(N(dt;d) = 1|\mathcal{F}_{t^-}^{bcd})$ is independent of $\mathcal{F}_t^a$ by assumption. Thus we have to show that $\hat{\pi}_t(dH) = P(H_t \in dH|\mathcal{F}_t^{acd}) = P(H_t \in dH|\mathcal{F}_t^{cd})$. This can be done by verifying that neither the innovation gains due to the occurrences of observable events as given by (3.4) nor the probability given that no observable event occurs as given by (3.3) depend on the history of $N(t;a)$, i.e. on H_t^a. Note that the integration over $I\!H$ in both expressions is in fact an integration over the space of the unobservable histories $I\!H^b = \{\tilde{H} \in I\!H | \tilde{H}^{acd} = \hat{H}^{acd}\}$ since the remaining histories are fixed to what has been observed and is therefore denoted by $\hat{H}_t^{acd}$ in order to make the distinction clearer. For the same reason we only consider probabilities for sets $\tilde{B} \in \mathcal{H}^b$.

Assume that an observable event occurs at t. If this event equals a we get according to (3.4)

$$\begin{aligned}\hat{\pi}_t(\tilde{B}) &= \frac{\int_{I\!H^b} \hat{\pi}_{t^-}(dH^b)\mathbf{1}\{H^b \in \tilde{B}\}\lambda_a(t)}{\int_{I\!H^b} \hat{\pi}_{t^-}(dH^b)\lambda_a(t)} \\ &= \int_{I\!H^b} \hat{\pi}_{t^-}(dH^b)\mathbf{1}\{H^b \in \tilde{B}\} = \hat{\pi}_{t^-}(\tilde{B}), \quad \tilde{B} \in \mathcal{H}^b,\end{aligned}$$

since $\lambda_a(t)dt = P(N(dt;a) = 1|H^b_{t^-} \cup \hat{H}^{acd}_{t^-}) = P(N(dt;a) = 1|\hat{H}^{acd}_{t^-})$ cancels out. Thus, the probability for $\tilde{B}$ remains unchanged when the event a occurs.
In addition, we have to show that the innovation gain due to one of the other observable events is also independent of the information whether and when an event a has previously occurred. Consider the case that c occurs at time t then

$$\begin{aligned}\hat{\pi}_t(\tilde{B}) &= \frac{\int_{I\!H^b} \hat{\pi}_{t^-}(dH^b)\mathbf{1}\{H^b \in \tilde{B}\}\lambda_c(t)}{\int_{I\!H^b} \hat{\pi}_{t^-}(dH^b)\lambda_c(t)} \\ &= \begin{cases} \hat{\pi}_{t^-}(\tilde{B}), & \lambda_c(t) \text{ is } \mathcal{F}^{acd}_t\text{–measurable}, \\ P(H_t \in \tilde{B}|\hat{H}^{cd}_t), & \lambda_c(t) \text{ is } \mathcal{F}^{bcd}_t\text{–measurable}, \end{cases}\end{aligned}$$

where the second part follows because the foregoing ratio is independent of $\hat{H}^a$. Finally, if d occurs at time t we have that

$$\begin{aligned}\hat{\pi}_t(\tilde{B}) &= \frac{\int_{I\!H^b} \hat{\pi}_{t^-}(dH^b)\mathbf{1}\{H^b \in \tilde{B}\}\lambda_d(t)}{\int_{I\!H^b} \hat{\pi}_{t^-}(dH^b)\lambda_d(t)} \\ &= P(H_t \in \tilde{B} \mid \hat{H}^{cd}_t), \end{aligned} \tag{3.9}$$

since $\lambda_d(t)$ is $\mathcal{F}^{bcd}_t$–measurable.
For the calculation of $\hat{\pi}_t(\tilde{B})$ on intervals, where no observable events occur, as given in (3.3), we note that $\lambda_b(t)$ as well as $\lambda_d(t)$ do not depend on the history $\mathcal{F}^a_t$. As to the contribution of $\lambda_a(t)$ and $\lambda_c(t)$, we have shown above that the occurrence of a leaves $\hat{\pi}_t(\tilde{B})$ unchanged. Either the same holds for c or $\lambda_c(t)$ does not depend on the history $\mathcal{F}^a_t$.
The foregoing argumentation extends easily to the case of arbitrary disjoint subsets sets A, B, C of $\{1, \ldots, K\}$ and $D \subset B$, where $B' = B \backslash D$ takes the role of $\{b\}$. Condition (3.8) ensures that C can be partitioned into $C = C_1 \cup C_2$ such that $\lambda_{c_1}(t)$ is $\mathcal{F}^{ACD}_t$–measurable for all $c_1 \in C_1$ and $\lambda_{c_2}(t)$ is $\mathcal{F}^{BCD}_t$–measurable for all $c_2 \in C_2$.

In case that the sets A, B, C and D overlap let $A^* = A \backslash (C \cup D)$, $C^* = C \backslash D$ and $B^* = B \backslash D$. The conditions of this proposition, (3.7) and (3.8), then imply that $B^* \not\to A^* | (C^* \cup D)$ and $\forall\, k \in C^* : A^* \not\to \{k\} | (C^* \cup B)$ or $B^* \not\to \{k\} | (C^* \cup A^* \cup D)$. The property of right decomposition is equal to the implication $A^* \not\to B | C^* \backslash B^* \Rightarrow A^* \not\to D | C^*$. Since the last transformation involves only disjoint sets the same argumentation as above may again be applied. □

Note that the conditions (3.6) – (3.8) given for right decomposition of local independence parallel those given in Proposition 1.2.14 for δ–separation.

3.2.2 Local independence and multi–state processes

Local independence has been formulated for general processes allowing for a Doob–Meyer decomposition. As mentioned above, we mainly aim at applications in event history analysis. This is not as restrictive as it may sound since there is often a sensible one–to–one relation between a marked point process and a multi–state process. In order to analyze the local independence structure of a marked point process it may in most situations be even helpful to base the statistical model on a multi–state process which can itself again be represented as a marked point process and which captures some aspects of the dependence structure through a suitable choice of the state space. There is, however, no general method how to proceed since the appropriate model heavily depends on the actual data situation, the substantial background, and on the question of interest. We therefore just give a brief example and address some aspects that may be important in most situations that we have in mind, where the local independence structure might be of interest.

To illustrate the foregoing remarks consider the example of a marked point process with three marks $\mathcal{E} = \{e_1, e_2, e_3\}$. Assume that the events can occur in any order but each only once and that the corresponding counting processes N_1, N_2, N_3 have absolutely continuous compensators. Let T_1, T_2 and T_3 be the random times of their occurrences. For the local independence structure it is of interest to find out if, for instance, the intensity for e_1 at t depends on whether and when the other events have previously occurred. Event e_1 being locally independent of the other events would implies that its intensity remains the same in the following different

situations: (1) $T_1 < \min(T_2, T_3)$, (2) $T_2 < T_1 < T_3$ or $T_3 < T_1 < T_2$, and (3) $\max(T_2, T_3) < T_1$. These situations could also be looked at as defining different states of a three dimensional process $\mathbf{Y}$ with state space $\mathcal{S} = \{0,1\}^3$ where the k-th component indicates whether e_k, $k = 1,2,3$, has occurred before or at time t. Each transition can now be regarded as an event of its own resulting in another marked point process $\{T_s, \tilde{E}_s | s = 1, \ldots, S\}$ with the same points in time T_s but a different mark space $\tilde{\mathcal{E}} = \{(\mathbf{y}, \mathbf{y}') | \mathbf{y}, \mathbf{y}' \in \mathcal{S} \text{ and } \exists\, k \in \{1,2,3\} : y_k = 0 \wedge y'_k = 1 \wedge y_j = y'_j, j \neq k\}$. The set $\tilde{\mathcal{E}}$ describes all possible transitions assuming that no different events occur at the same time. The corresponding counting processes $\tilde{N}(t; (\mathbf{y}, \mathbf{y}'))$ can be indexed by k if $y_k = 0 \wedge y'_k = 1$ indicating that the transition is due to a change in the k-th component (or that e_k occurs). They are given as

$$\tilde{N}_k(t; (\mathbf{y}, \mathbf{y}')) = \sum_{s \leq t} \mathbf{1}\{\mathbf{Y}(s^-) = \mathbf{y}, \mathbf{Y}(s) = \mathbf{y}'\},$$

for $y_k = 0$, $y'_k = 1$, and $y_j = y'_j$, $j \neq k$, $k = 1,2,3$. If for instance $k = 1$, $\mathbf{y} = (0,0,1)$, and $\mathbf{y} = (1,0,1)$ we may equivalently define

$$\tilde{N}_1(t; ((0,0,1),(1,0,1))) = \mathbf{1}\{T_3 < T_1 \leq t \leq T_2\}.$$

The counting processes $N_k(t)$ for the original marked point processes are given as appropriate sums over the $\tilde{N}_k(t; (\mathbf{y}, \mathbf{y}'))$, e.g.

$$N(t; e_1) = \sum_{j,l \in \{0,1\}} \tilde{N}_1(t; ((0,j,l),(1,j,l))).$$

Analogously, the original intensity processes can be expressed via those corresponding to $\tilde{\mathcal{E}}$, e.g.

$$\lambda_1(t) = \begin{cases} \tilde{\lambda}_1(t; ((0,0,0),(1,0,0))), & t \leq \min(T_2, T_3), \\ \tilde{\lambda}_1(t; ((0,1,0),(1,1,0))), & T_2 < t \leq T_3), \\ \tilde{\lambda}_1(t; ((0,0,1),(1,0,1))), & T_3 < t \leq T_2), \\ \tilde{\lambda}_1(t; ((0,1,1),(1,1,1))), & \max(T_2, T_3) < t. \end{cases} \tag{3.10}$$

Further and most important for our purposes, the local independence structure of $\mathbf{N}(t) = (N_1(t), N_2(t), N_3(t))$ carries over to $\tilde{\mathbf{N}}(t) = (\tilde{N}_k(t; \tilde{e}) | k = 1,2,3, e \in \tilde{\mathcal{E}})$ in the sense that if $N_k \not\to N_j | \mathbf{N}_{V \setminus \{j,k\}}$ then $\tilde{N}_k \not\to \tilde{N}_j | \tilde{\mathbf{N}}_{V \setminus \{j,k\}}$. For instance if e_1 is locally independent of e_2, i.e. $\{2\} \not\to \{1\} | \{3\}$ we have by definition that $\lambda_1(t)$ is $\mathcal{F}^{13}$–measurable. This implies that $\tilde{\lambda}_1(t; ((0,0,0),(1,0,0)) = \tilde{\lambda}_1(t; ((0,1,0),(1,1,0))$ and

$\tilde{\lambda}_1(t;((0,0,1),(1,0,1)) = \tilde{\lambda}_1(t;((0,1,1),(1,1,1))$ as can be seen from (3.10) since the information whether $t \leq T_2$ or $t > T_2$ is not relevant. Thus, each counting process N_k of the original marked point process may be identified with a multivariate counting process $(\tilde{N}_k(t;\tilde{e})|\tilde{e} \in \tilde{\mathcal{E}})$ corresponding to the transitions of the associated multi–state process $\mathbf{Y}$.

The above example can be generalized and specialized in several ways. For instance, if all the intensities possibly depend on whether other events have occurred but not on the specific times of these events we have a Markov process which is treated in more detail in the next section. It might also be possible that the marks $\mathcal{E}$ of the original process occur more often than once. The associated multi–state process could then be given as the multivariate counting process itself, $\mathbf{Y} = \mathbf{N} = (N_1, \ldots, N_K)$, with state space $\mathcal{S} = \mathbb{N}_0^K$ so that e.g. the state $\mathbf{Y}(t) = (3,2,0)$ would imply that e_1 has occurred three times, e_2 two times, and e_1 never before t. All states are then transient, where the only possible transitions are into states where one component has increased by one. The different intensities are then not only those of N_k, $k = 1, \ldots K$, but those describing all the changes of states of $\mathbf{Y}$. A multi–state process of this generality is difficult to analyze so that in practical situations restrictions are typically required.

In other situations it may seem appropriate to group together some of the events in $\mathcal{E}$. The original marked point process is then equivalently represented by a multi–state process possibly with less than k components and recurrent states. Consider the example of e_1 ='becoming ill' and e_2 ='recover'. Both events are trivially locally dependent of each other: one can only recover if one has been ill before, and one can only become ill if one has previously recovered. The corresponding multi–state process $\mathbf{Y}$ may then comprise a component Y_j that indicates the health status with the two recurrent states 'healthy' and 'ill'. This type of multi–state processes, where a component not necessarily indicates whether and how often an event has occurred but instead the alternation between several events, is also treated in the next section assuming that the whole process is Markovian.

Formally, the relation between a marked point process and an equivalent multi–state

process consists in that both generate the same history as indicated in the following definition.

Definition 3.2.10 *Associated multi–state process*
Consider a marked point process $\{(T_s, E_s)|s = 1, \ldots, S\}$ with $E_s \in \mathcal{E} = \{e_1, \ldots, e_K\}$ (cf. Definition 3.1.1). Let $\mathbf{N} = (N_1, \ldots, N_K)$ be the corresponding mark specific counting processes. A multivariate stochastic process $\mathbf{Y} = (Y_1, \ldots, Y_J)$ with finite state space $\mathcal{S}$ and corresponding counting processes for the transitions $\tilde{N}_{qr}$, $q, r \in \mathcal{S}$, is called a *multi–state process associated to the marked point process* if $\sigma\{\mathbf{N}(s)|s \leq t\} = \sigma\{\tilde{N}_{qr}(s)|s \leq t; q, r \in \mathcal{S}, q \neq r\}$ for all $t \in \mathcal{T}$. //

As explained above, the associated multi–state process is not necessarily unique. A more precise definition and example for the relation of marked point processes and a multi–state process is given in the following subsection.

3.2.3 Local independence for Markov processes

Local independence was originally defined by Schweder (1970) for the special case of Markov processes. Since this is a very broad class of processes often used for modeling random structures in time, we now illustrate some results concerning local independence by restricting ourselves to this class. In contrast to Definition 2.1.11, we now give a general definition including the continuous time situation, but we restrict ourselves to Markov processes of first order.

Definition 3.2.11 *Finite Markov process*
Let $\mathcal{S}$ be a finite set of states. A stochastic process $Y = \{Y(t)|t \in \mathcal{T}\}$ is called a *finite (first order) Markov process* with state space $\mathcal{S}$ if for all $n \in \mathbb{N}$ and $t_n \geq t_{n-1} > t_{n-2} > \cdots > t_0 \in \mathcal{T}$:

$$P(Y(t_n) = y_n \mid Y(t_{n-1}) = y_{n-1}, Y(t_{n-2}) = y_{n-2}, \ldots, Y(t_0) = y_0)$$
$$= P(Y(t_n) = y_n \mid Y(t_{n-1}) = y_{n-1})$$

for all $y_n, \ldots, y_0 \in \mathcal{S}$.
A Markov process is called *stationary* if

$$P(Y(t+s) = y' \mid Y(s) = y) = P(Y(t) = y' \mid Y(0) = y) \quad \forall\, t, t+s \in \mathcal{T}.$$

//

A Markov process is characterized by its transition intensities which are defined as follows.

Definition 3.2.12 *Transition intensities*
Let Y be a finite state Markov process. The *transition intensities* $\alpha_{qr}(t)$, $t \in \mathcal{T}$, of Y are defined as

$$\alpha_{qr}(t) = \lim_{h \downarrow 0} \frac{1}{h} P\left(Y(t+h) = r \mid Y(t) = q\right), \qquad q \neq r \in \mathcal{S}.$$

//

In the following we show how a Markov process can be reformulated as a multivariate counting process.

Definition 3.2.13 *Markov process as counting process*
Let $Y = \{Y(t) | t \in \mathcal{T}\}$ be a finite state Markov process with state space $\mathcal{S}$ and with transition intensities $\alpha_{qr}(t)$, $q, r \in \mathcal{S}$. The *associated multivariate counting process* is given by $\mathbf{N} = (N_{qr} | q, r \in \mathcal{S}, q \neq r)$ with

$$N_{qr}(t) = \sum_{s \leq t} \mathbf{1}\{Y(s^-) = q \text{ and } Y(s) = r\}, \quad q \neq r.$$

With $\mathcal{F}_t = \sigma\{Y(s) | s \leq t\}$, its $\mathcal{F}_t$–compensator (if it exists) has components

$$\Lambda_{qr}(t) = \int_0^t Z_q(s) \alpha_{qr}(s)\, ds, \tag{3.11}$$

where $Z_q(t) = \mathbf{1}\{Y(t^-) = q\}$. The compensator exists if the involved transition intensities $\alpha_{qr}(t)$ exist (Andersen et al., 1993, p. 94). //

From (3.11) it can be seen that the compensator is given as the expected number of transitions from q to r before or at time t and the intensities are given by

$$\lambda_{qr}(t) = Z_q(t) \alpha_{qr}(t), \quad q, r \in \mathcal{S}, q \neq r. \tag{3.12}$$

To develop the ideas of Schweder (1970) some smoothness conditions with regard to the transition intensities are necessary. We assume that the transition intensities $\alpha_{qr}(t)$ exist, i.e.

$$\alpha_{qr}(t) < \infty \qquad \forall\, q \neq r, \tag{3.13}$$

and are continuous and bounded functions of t on any closed interval in $\mathcal{T}$.

The Markov process Y may be vector valued but a more general way of defining that a process consists of a set of different components is via the *compositioning* of the process.

Definition 3.2.14 *Composable Markov process*
Let $V = \{1, \ldots, K\}$, $K \geq 2$, and assume that there are K spaces $\mathcal{S}_k, k \in V$, with $|\mathcal{S}_k| \geq 2$, and that there exists a one-to-one mapping f of $\mathcal{S}$ onto $\times_{k \in V} \mathcal{S}_k$ so that elements $y \in \mathcal{S}$ can be identified with elements $(y_1, \ldots, y_K) \in \times_{k \in V} \mathcal{S}_k$. Then, a Markov process $\mathbf{Y}$ is a composable process with components $Y_1, \ldots, Y_K$ given by $f(\mathbf{Y}(t)) = (Y_1(t), \ldots, Y_K(t))$ if for all $A \subset V$, $|A| \geq 2$,

$$\lim_{h \downarrow 0} \frac{1}{h} P\left(\bigcap_{k \in A} \{Y_k(t+h) \neq y_k\} | \bigcap_{k \in A} \{Y_k(t) = y_k\} \right) = 0 \tag{3.14}$$

for all $y_k \in \mathcal{S}_k, k \in V$, and $t \in \mathcal{T}$. We then write $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$. //

The definition implies that for a composable process the probability that more than one component changes in a short period of length h is of magnitude $o(h)$. It is, thus, justified to regard the processes as *composed* of different components because any change of state can be represented as a change in only one of the components. Note that the compositioning is not necessarily unique. If for example $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ then $\mathbf{Y} \sim (\mathbf{Y}_A, \mathbf{Y}_B)$ with $A \subset V$ and $B = V \backslash A$.
In the following, we restrict ourselves to composable finite Markov processes (CFMP).

Definition 3.2.15 *Composable finite Markov process, CFMP*
Let $\mathbf{Y}$ be a Markov process with finite state space $\mathcal{S}$ and with the property (3.13). If $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ then it is called a *composable finite Markov process.* //

From (3.13) it follows that the transition probabilities of a CFMP fulfill

$$\lim_{h \downarrow 0} P\left(\mathbf{Y}(t+h) = \mathbf{y}' \mid \mathbf{Y}(t) = \mathbf{y}\right) = \begin{cases} 1, & \mathbf{y} = \mathbf{y}' \\ 0, & \mathbf{y} \neq \mathbf{y}' \end{cases} \quad \forall t \in \mathcal{T}.$$

Further, the transition intensities of a CFMP have the following properties.

Corollary 3.2.16 *Transition intensities for CFMP*
Let $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ be a CFMP. In the following we write $\alpha(t; (\mathbf{y}, \mathbf{y}'))$ instead of $\alpha_{\mathbf{y}\mathbf{y}'}(t)$ for notational convenience.
The intensity $\alpha(t; (\mathbf{y}, \mathbf{y}'))$ for any $\mathbf{y} \neq \mathbf{y}'$ is given by

$$\alpha(t; (\mathbf{y}, \mathbf{y}')) = \begin{cases} \alpha_k(t; (\mathbf{y}, y'_k)), & y_k \neq y'_k \text{ and } \mathbf{y}_{-k} = \mathbf{y}'_{-k} \\ 0, & \text{else,} \end{cases},$$

where $\mathbf{y}_{-k} = \mathbf{y}_{V \setminus \{k\}}$. Thus, the intensity equals 0 if $\mathbf{y}$ and $\mathbf{y}'$ differ on more than one component, and

$$\alpha_k(t; (\mathbf{y}, y'_k)) = \lim_{h \downarrow 0} \frac{1}{h} P\left(Y_k(t+h) = y'_k \mid \mathbf{Y}(t) = \mathbf{y}\right).$$

Proof:
The result is an immediate consequence of the definition of a CFMP. □

The dependence structure of the components $(Y_1, \ldots, Y_K)$ is thus determined by the quantities $\alpha_k(t; (\mathbf{y}, y'_k)), \mathbf{y} \in \mathcal{S}, y'_k \in \mathcal{S}_k, k \in V$. The intuitive notion of local independence and the definition given by Schweder (1970) concern the possibility that these transition intensities do not depend on all components of their first arguments $\mathbf{y}$. We now show the equivalence of this definition to the local independence of stochastic processes as given in Definition 3.2.1.

Proposition 3.2.17 *Local independece in a CFMP*
Let $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ be a CFMP. Then, Y_j is locally independent of Y_k, $k \neq j$, if and only if $\alpha_j(t; (\mathbf{y}, y'_j))$ is constant in the k-th component y_k of the first argument $\forall\, \mathbf{y}_{-k} \in \mathcal{S}_{-k}$ and $y'_j \in \mathcal{S}_j$, $y'_j \neq y_j$.

Proof:
The associated multivariate counting process for a single component Y_j, $j = 1, \ldots, K$, is given by the set of counting processes each counting a change of state where the destination state differs from the origin in the j-th component, i.e. $\mathbf{N}_j(t) = (N(t; (y_j, y'_j)) | y_j, y'_j \in \mathcal{S}_j, y_j \neq y'_j)$ with

$$N(t; (y_j, y'_j)) = \sum_{s \leq t} \mathbf{1}\{Y_j(s^-) = y_j, Y_j(s) = y'_j\}.$$

Let $\mathcal{F}_t$ be the filtration generated by the whole process $\mathbf{Y}$. Then, the $\mathcal{F}_t$–intensities of $\mathbf{N}_j$ are given by $\lambda_j(t) = (\lambda(t;(y_j,y_j'))|y_j, y_j' \in \mathcal{S}_j, y_j \neq y_j')$ with

$$\lambda(t;(y_j,y_j')) = \sum_{\mathbf{y}_{-j} \in \mathcal{S}_{-j}} Z_{\mathbf{y}}(t)\alpha_j(t;(\mathbf{y},y_j')), \qquad j = 1,\ldots,K. \tag{3.15}$$

From this and recalling (3.12) as well as $Z_{\mathbf{y}}(t) = \mathbf{1}\{\mathbf{Y}(t^-) = \mathbf{y}\}$, it is obvious that the following statements are equivalent: (1) $\alpha_j(t;(\mathbf{y},y_j'))$ is constant in $y_k \; \forall \; \mathbf{y}_{-k} \in \mathcal{S}_{-k}$ and all $y_j' \in \mathcal{S}_j$, $y_j' \neq y_j$, and (2) $\lambda_j(t)$ is $\mathcal{F}_t^{-k}$ measurable for all $t \in \mathcal{T}$. □

Note that the condition (3.5) needed for left intersection always holds for CFMPs by definition since the components are not allowed to jump at the same time.

Example: Consider as an example a CFMP with four components $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, where Y_1 describes the employment status of women with states $y_1 = 0 \mathrel{\hat{=}}$ 'not employed', $y_1 = 1 \mathrel{\hat{=}}$ 'employed', Y_2 the fertility with states $y_2 \mathrel{\hat{=}}$ 'number of children', Y_3 the martial status with states $y_3 = 0 \mathrel{\hat{=}}$ 'unwed', $y_3 = 1 \mathrel{\hat{=}}$ 'married', $y_3 = 2 \mathrel{\hat{=}}$ 'divorced', and Y_4 indicating whether the person still lives with her parents or not with states $y_4 = 0 \mathrel{\hat{=}}$ 'not at home' and $y_4 = 1 \mathrel{\hat{=}}$ 'at home'. In this situation one might hypothesize the following local independencies: Given the information about whether the woman is employed and married, knowing that she lives at home or not adds no new information w.r.t. the intensity of getting a child, i.e. $Y_4 \not\to Y_2|\mathbf{Y}_{\{1,3\}}$. Further, knowing the number of children, the employment status as well as the information about where the women lives adds no information w.r.t. the intensity for getting married or divorced, i.e. $Y_1 \not\to Y_3|\mathbf{Y}_{\{2,4\}}$ and $Y_4 \not\to Y_3|\mathbf{Y}_{\{1,2\}}$. Since the considered life history events are not determined by one another we may assume that condition (3.5) holds and get by the property of left intersection for local independence that $\mathbf{Y}_{\{1,4\}} \not\to Y_3|Y_2$. This also follows directly from the implications for the transitions intensities. With the first local independence it holds that

$$\alpha_2(t;((y_1,x,y_3,0),x+1)) = \alpha_2(t;((y_1,x,y_3,1),x+1)),$$

for $y_1 = 0,1,2$, $y_3 = 0,1$, and for all $x \in \mathbb{N}_0$. From the second pair of local independencies we have

$$\alpha_3(t;((0,x,0,0),1)) = \alpha_3(t;((1,x,0,0),1)) = \\ \alpha_3(t;((0,x,0,1),1)) = \alpha_3(t;((1,x,0,1),1))$$

for all $x \in \mathbb{N}_0$, and analogous equalities hold for transitions between the states 'married' and 'divorced'. Thus we can write $\alpha_3(t;((y_1,x,0,y_4),1)) = \alpha_3(t;((x,0),1))$ etc. With regard to the corresponding formulation via the intensity processes for all possible transitions (y_3, y_3') in component Y_3, which are $(0,1)$, $(1,2)$, and $(2,1)$, we have that

$$\lambda_3(t;(y_3,y_3')) = \sum_x \alpha_3(t;((x,y_3),y_3'))\mathbf{1}\{Y_2(t^-) = x, Y_3(t^-) = y_3\}$$

which is obviously measurable w.r.t. $\mathcal{F}_t^{\{23\}}$ as claimed in the above proposition. //

An interesting result for local independence for CFMPs concerns the relation between local and conditional independence. The following lemma has already been shown by Schweder (1970).

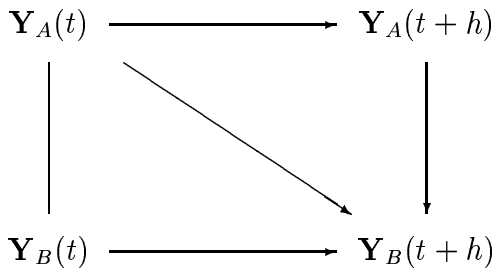
Lemma 3.2.18 *Conditional independence for partition of CFMP*
Let $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ be a CFMP and $A, B \subset V = \{1, \ldots, K\}$ with $A \cap B = \emptyset$ and $A \cup B = V$. If $\mathbf{Y}_A$ is locally independent of $\mathbf{Y}_B$, i.e. $B \not\to A$, then the following conditional independencies hold

$$\mathbf{Y}_A(t+h) \perp\!\!\!\perp \mathbf{Y}_B(t) \mid \mathbf{Y}_A(t) \quad \forall\, h > 0; \;\; t, t+h \in \mathcal{T}. \tag{3.16}$$

Further, $\mathbf{Y}_A(t)$ is a Markov process with transition intensities $\alpha_k(t;(\mathbf{y}_A, y_k'))$, $k \in A$.

Proof:
This follows from Theorems 1 and 2 of Schweder (1970). □

Figure 3.1: Conditional independence graph given the local independence $B \not\to A$.



The above lemma implies that the independence structure of $\mathbf{Y}_A$ and $\mathbf{Y}_B$ may be depicted in a graphical chain model as given in Figure 3.1. In this graph, the only

conditional independence is the one given in (3.16). No other conditional independencies hold in general if $\mathbf{Y}_B$ locally depends on $\mathbf{Y}_A$. In particular, it does not hold that $\mathbf{Y}_B(t+h) \perp\!\!\!\perp \mathbf{Y}_A(t+h) | \mathbf{Y}(t)$ in the considered situation due to marginalizing w.r.t. $\mathbf{Y}(s)$, $t < s < t+h$.

Generalizing the above lemma to the case where (A, B) is no partition of V yields the next proposition.

Proposition 3.2.19 *Conditional independence for subsets of CFMP*
Let $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ be a CFMP and $A, B \subset V$ with $A \cap B = \emptyset$ and $A \cup B \neq V$. Define $C = V \backslash (A \cup B)$. Assume that $B \not\to A|C$.

(1) If $C \not\to A \mid B$ then property (3.16) still holds.

(2) If $B \not\to C \mid A$ then we have the following conditional independence:

$$\mathbf{Y}_A(t+h) \perp\!\!\!\perp \mathbf{Y}_B(t) \mid \mathbf{Y}_{A \cup C}(t) \quad \forall\, h > 0; t, t+h \in \mathcal{T}.$$

Proof:
The first part follows immediately from the implication $(B \not\to A|C) \wedge (C \not\to A|B) \Rightarrow (B \cup C \not\to A)$ (left intersection), the above Lemma 3.2.18, and the property of decomposition for conditional independence.
In order to show the second part, note that $B \not\to A|C$ and $B \not\to C|A$ implies $B \not\to A \cup C$ by right intersection for local independence. Therefore, we get with the Lemma 3.2.18

$$\mathbf{Y}_{A \cup C}(t+h) \perp\!\!\!\perp \mathbf{Y}_B(t) \mid \mathbf{Y}_{A \cup C}(t) \quad \forall\, h > 0; t, t+h \in \mathcal{T}.$$

Applying again decomposition for conditional independence yields the desired result. □

It remains the situation where $\mathbf{Y}_A$ locally depends on $\mathbf{Y}_C$ and $\mathbf{Y}_C$ locally depends on $\mathbf{Y}_B$. Then, no conditional independence statement similar to (3.16) is possible because marginalizing w.r.t. possible intermediate transitions in $\mathbf{Y}_C(s)$, $t < s < t+h$, may induce a dependence between $\mathbf{Y}_A(t+h)$ and $\mathbf{Y}_B(t)$ which is not 'absorbed' by $\mathbf{Y}_A(t)$. For the situation where a continuous time CFMP can only be observed in

discrete time, say at $t_1, t_2, \ldots, t_n$, Lemma 3.2.18 and the foregoing proposition imply that the conditional independence structure of $\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_n)$ cannot be captured by a graphical chain model. At least, the characteristic independencies could not be read off a conditional independence graph since there would typically be several dependencies on past components as well as relations among the components at a given discrete point in time due to marginalizing over the time in between similar to the instantaneous causality defined in the previous chapter.

Further results on the properties of composable Markov processes are referred to the next chapter after introducing local independence graphs.

Chapter 4

Local independence graphs

Let us now turn to the graphical representation of dependencies among events. It is the aim of such a graphical representation to capture as much information about the dependence structure as possible so that important properties of the underlying statistical model can directly be read off the graph. As we have seen in the previous chapters, event history data is conveniently modeled through the mark specific point processes and the distribution is characterized by the compensators of these processes. In addition, local independence is defined through the compensators. It is thus self–evident to base the graphical representation on the local independencies so that, compared to conditional independence graphs, the compensators now take the role of the conditional probabilities.

In conditional independence graphs the representation is based on the Markov properties of the underlying statistical model. But when considering local independence structures we have to deal with an asymmetric concept and possibly with cycles so that new properties analogous to the Markov properties for conditional independence graphs are required. These are introduced in Section 4.1 below and called *dynamic* Markov properties. Again in analogy to conditional independence graphs, we explore in Section 4.2 the factorization property of the likelihood for marked point processes with a given local independence graph. It is shown that some interesting results relating local and conditional independencies can be obtained from this factorization. In Section 4.3, the topic of collapsibility is addressed, i.e. conditions for a subprocess to have the structure of the corresponding subgraph. In

Section 4.4, we briefly indicate a simplification of the graphical representation using stopped processes for situations like e.g. survival analysis where we typically have a final absorbing state. Finally, we sketch in Section 4.5 how local independence graphs may be used for causal reasoning, emphasizing that they are a–priori no causal models.

4.1 Dynamic Markov properties

An obvious way of representing the local independence structure of a multivariate process by a graph is to use an arrow as symbol for local dependence yielding a local independence graph. This parallels the definition of the pairwise Markov property for conditional independence graphs. The local independence graph thus consists of vertices representing the components of the multivariate process and of directed edges, where $(j,k) \in E$ and $(k,j) \in E$ is allowed and represented as bidirected edge. The following pairwise dynamic Markov property is at the same time the defining property for local independence graphs.

Definition 4.1.1 *Local independence graph*
Let $\mathbf{Y}_V = (Y_1, \ldots, Y_K)$ be a multivariate process. Let further $G = (V, E)$ be a directed graph, $V = (1, \ldots, K)$. Then, G is called the *local independence graph of* $\mathbf{Y}_V$ if for all $j, k \in V$

$$(j,k) \notin E \quad \Leftrightarrow \quad Y_j \not\to Y_k \mid \mathbf{Y}_{V\setminus\{j,k\}}. \tag{4.1}$$

Property (4.1) is called the *pairwise dynamic Markov property.* Alternatively we also write $\{j\} \not\to \{k\} | V\setminus\{j,k\}$. //

The above definition entails the question if more properties than the one of pairwise local independence can be read off the graph. In analogy to the classical graphs one could be interested in 'local' and 'global' dynamic Markov properties. Let us first formulate the former one.

Definition 4.1.2 *Local dynamic Markov property*
Let $G = (V, E)$ be a directed graph. For a multivariate process $\mathbf{Y}_V$ the property

$$\forall\, k \in V: \quad V\setminus \mathrm{cl}(k) \not\to \{k\} \mid \mathrm{pa}(k), \tag{4.2}$$

where $\mathrm{cl}(k) = \mathrm{pa}(k) \cup \{k\}$ (cf. Definition 1.1.6), is called the *local dynamic Markov property w.r.t. G.* //

The above property (4.2) could for instance be violated if two components in $\mathrm{pa}(k)$ are a.s. identical. Let for instance $Y_1 = Y_2$ a.s. Then, it holds for any Y_3 that $Y_1 \not\to Y_3|Y_2$ as well as $Y_2 \not\to Y_3|Y_1$ but not necessarily $(Y_1, Y_2) \not\to Y_3$. This, however, is prevented for marked point processes by the orthogonality Assumption 3.2.3 because for counting processes which are a.s. identical the martingales resulting from the Doob–Meyer decomposition are not orthogonal (Andersen et al., 1993, pp. 73).

As shown below, the local dynamic Markov property (4.2) follows from the pairwise (4.1) under the assumption of property (3.5). This implication is a special case of a more general dynamic Markov property, namely the *global dynamic Markov property.* The global property is related to the notion of separation in graphs or equivalently in the multivariate random systems depicted by the graph. In a local independence graph, separation would entail the following property: If $A \not\to B|V\backslash(A \cup B)$, i.e. if there is a local independence w.r.t. the whole available information, a separating set $C \subset V\backslash(A \cup B)$ is given if $A \not\to B|C$, i.e. the local independence is preserved even when ignoring the information in the subprocess $\mathbf{Y}_{V\backslash(A\cup B\cup C)}$. This implies that the information in $\mathcal{F}_t^{B\cup C}$ is sufficient to assess the local independence $\mathbf{Y}_A \not\to \mathbf{Y}_B|\mathbf{Y}_C$. Clearly, this property is essential for the possibility to reduce complexity in a multivariate setting. The concept of δ–separation meets these requirements which can heuristically be explained as follows. Assuming that $A \not\to B|C$ corresponds to the past $\mathcal{F}_{t^-}^A$ of $\mathbf{Y}_A$ being irrelevant for the infinitesimal prediction of $\mathbf{Y}_B(t+h)$ given the past $\mathcal{F}_{t^-}^C$ of $\mathbf{Y}_C$. Thus, we have to consider the relation between $\mathbf{Y}_B(t+h)$ and $\mathcal{F}_{t^-}^A$ as well as $\mathcal{F}_{t^-}^C$ which can approximately be represented by a DAG. Thus, all 'descendants' of $\mathbf{Y}_B(t+h)$, i.e. the future, are to be discarded. Consequently, all directed edges starting in B are deleted when checking for δ–separation. Further, for any instance $s < t$ in the past the independence structure among $\mathbf{Y}(s)$ and $\mathbf{Y}(s+h)$ may also be represented as a DAG with no edge (j,k) between $Y_j(s)$ and $Y_k(s+h)$ for $\{j\} \not\to \{k\}|V\backslash\{j,k\}$. Thus, by conditioning on the past we have to moralize the graph according to the moralization of a DAG.

The global dynamic Markov property therefore reads as follows.

Definition 4.1.3 *Global dynamic Markov property*
Let $\mathbf{Y}_V$ be a multivariate stochastic process and $G = (V, E)$ a directed graph. For disjoint subsets $A, B, C \subset V$ the property

$$C \ \delta\text{–separates } A \text{ from } B \text{ in G} \quad \Rightarrow \quad A \not\to B \mid C. \tag{4.3}$$

is called the *global dynamic Markov property w.r.t. G.* //

We have seen that local independence and δ–separation essentially hold the same properties of asymmetric graphoids. Thus, all prerequisites for the main result in this section have been gathered.

Theorem 4.1.4 *Equivalence of dynamic Markov properties*
Let $\{(T_s, E_s)|s = 1, \ldots, S\}$ be a marked point process with $\mathcal{E} = \{e_k|k = 1, \ldots, K\}$ and with local independence graph $G = (V, E)$, $V = \{1, \ldots, K\}$. Under the assumptions of Theorem 3.1.9 and assuming (3.5), the pairwise, local and global dynamic Markov properties are equivalent.

Proof:
The structure of the proof corresponds to the one given by Lauritzen (1996, p. 34) for the equivalence of the Markov properties in undirected conditional independence graphs. Due to the asymmetry of local independence the present proof is slightly more complicated.
It is easily checked that (4.3) $\Rightarrow$ (4.2) $\Rightarrow$ (4.1): First, pa(k) always δ–separates $V\backslash(\text{pa}(k) \cup \{k\})$ from $\{k\}$ in G. Second, (4.2) $\Rightarrow$ (4.1) holds by left weak union and let decomposition.
Now, assume that (4.1) holds and that C δ–separates A from B in the local independence graph. We have to show that $A \not\to B|C$, i.e. the $\mathcal{F}_t^{A\cup B\cup C}$–compensators $\Lambda_k(t)$, $k \in B$, are $\mathcal{F}_t^{B\cup C}$–measurable. The proof is via backward induction on the number $|C|$ of vertices in the separating set. If $|C| = |V| - 2$ then both, A and B consist of only one element and (4.3) trivially holds. If $|C| < |V| - 2$ then either A or B consist of more than one element.
Let us first consider the case that A, B, C is a partition of V and none of them is empty. If $|A| > 1$ let $\alpha \in A$. Then, by left weak union and left decomposition (or converse contraction) of δ–separation we have that $C \cup (A\backslash\{\alpha\})$ δ–separates $\{\alpha\}$

from B, i.e.

$$\{\alpha\} \perp\!\!\!\perp_{\delta} B \mid C \cup (A \backslash \{\alpha\})$$

and $C \cup \{\alpha\}$ δ–separates $A \backslash \{\alpha\}$ from B in G, i.e.

$$A \backslash \{\alpha\} \perp\!\!\!\perp_{\delta} B \mid (C \cup \{\alpha\}).$$

Therefore, we have by the induction hypothesis that

$$\{\alpha\} \not\perp\!\!\!\perp B \mid C \cup (A \backslash \{\alpha\}) \text{ and } A \backslash \{\alpha\} \not\perp\!\!\!\perp B \mid (C \cup \{\alpha\}).$$

From this it follows with the modified version of left intersection as given in (1.3) (which can be applied because of the assumption that (3.5) holds) that $A \not\perp\!\!\!\perp B | C$ as desired.

If $|B| > 1$ let $\{\beta\} \in B$. With Lemma 1.2.12 we have that $C \cup (B \backslash \{\beta\})$ δ–separates A from $\{\beta\}$, i.e.

$$A \perp\!\!\!\perp_{\delta} \{\beta\} \mid C \cup (B \backslash \{\beta\})$$

and $C \cup \{\beta\}$ δ–separates A from $B \backslash \{\beta\}$, i.e.

$$A \perp\!\!\!\perp_{\delta} B \backslash \{\beta\} \mid (C \cup \{\beta\}).$$

Therefore, we have again by the induction hypothesis that

$$A \not\perp\!\!\!\perp \{\beta\} \mid C \cup (B \backslash \{\beta\}) \text{ and } A \not\perp\!\!\!\perp B \backslash \{\beta\} \mid (C \cup \{\beta\}). \tag{4.4}$$

From the first part of (4.4) it follows trivially by definition that $A \not\perp\!\!\!\perp \{\beta\} | (C \cup B)$. Applying right contraction to this and the second part of (4.4) yields $A \not\perp\!\!\!\perp B | (C \cup \{\beta\})$ which implies again by definition $A \not\perp\!\!\!\perp B | C$. Note that this argumentation could not be applied to the first part of the proof because $A \not\perp\!\!\!\perp B | (C \cup \{\alpha\})$ does not imply $A \not\perp\!\!\!\perp B | C$.

Let us now consider the case that $A, B, C \subset V$ are disjoint but no partition of V. First, we assume that they are a partition of $\mathrm{An}(A \cup B \cup C)$, i.e. that $A \cup B \cup C$ is an ancestral set. Let $\gamma \in V \backslash (A \cup B \cup C)$, i.e. γ is no ancestor of $A \cup B \cup C$. Thus, every allowed trail (cf. Proposition 1.2.5) from γ to B is blocked by $A \cup C$ since any such trail includes an edge (k, b) for some $b \in B$ where no edges meet head–to–head in k and $k \in A \cup C$. Therefore, we get

$$\{\gamma\} \perp\!\!\!\perp_{\delta} B \mid (A \cup C).$$

Application of left contraction, weak union, and decomposition for δ–separation yields

$$A \,\mathrm{IR}_\delta\, B \mid (C \cup \{\gamma\}).$$

It follows with the induction hypothesis that

$$A \not\to B \mid (C \cup \{\gamma\}) \text{ as well as } \{\gamma\} \not\to B \mid (A \cup C).$$

With left intersection as given by (1.3) in Corollary 1.2.8 and left decomposition for local independence we get the desired result.
Finally, let A, B, C be disjoint subsets of V and $A \cup B \cup C$ not necessarily an ancestral set. Choose $\gamma \in \mathrm{an}(A \cup B \cup C)$ and define $\tilde{G}^B = G^B_{\mathrm{An}(A \cup B \cup C)}$. Since $A \perp\!\!\!\perp_G B | C$ in $(\tilde{G}^B)^m$ we know from the properties of ordinary graph separation that

$$\begin{aligned} &(1) \quad \text{either } \{\gamma\} \perp\!\!\!\perp_G B \mid (A \cup C) \text{ in } (\tilde{G}^B)^m \\ &(2) \quad \text{or } A \perp\!\!\!\perp_G \{\gamma\} \mid (B \cup C) \text{ in } (\tilde{G}^B)^m. \end{aligned}$$

In the first case $\{\gamma\} \,\mathrm{IR}_\delta\, B | (A \cup C)$ and it follows from left contraction that

$$(A \cup \{\gamma\}) \,\mathrm{IR}_\delta\, B \mid C.$$

Application of left weak union and left decomposition yields $A \,\mathrm{IR}_\delta\, B | (C \cup \{\gamma\})$. With the induction hypothesis we therefore get

$$A \not\to B \mid (C \cup \{\gamma\}) \text{ and } \{\gamma\} \not\to B \mid (A \cup C).$$

Left intersection according to (1.3) and left decomposition for local independence yields $A \not\to B | C$.
The second case is the most complicated and the proof makes now use of right decomposition for local independence under the conditions given in Proposition 3.2.9. First, we have from (2) that $A \perp\!\!\!\perp_G \{\gamma\} | B \cup C$ in $(G_{\mathrm{An}(A \cup B \cup C)})^m$ since the additional edges starting in B can only yield additional paths between A and γ which are again intersected by B. Therefore, it holds

$$A \,\mathrm{IR}_\delta\, \{\gamma\} \mid (B \cup C).$$

With $A \,\mathrm{IR}_\delta\, B | C$, application of right contraction for δ–separation yields $A \,\mathrm{IR}_\delta\, (B \cup \{\gamma\}) | C$. Now, we can apply Lemma 1.2.12 to get $A \,\mathrm{IR}_\delta\, B | (C \cup \{\gamma\})$ from where it follows with the induction hypothesis that

$$A \not\to B \mid (C \cup \{\gamma\}) \text{ and } A \not\to \{\gamma\} \mid (B \cup C).$$

With right intersection for local independence we get $A \not\to (B \cup \{\gamma\})|C$. In addition, $\{\gamma\} \not\to A|(B \cup C)$ by the same arguments as given above for $A \not\to \{\gamma\}|(B \cup C)$. In order to apply Proposition 3.2.9 we still have to show that for all $k \in C$ either $A \operatorname{IR}_\delta \{k\}|(C \cup B \cup \{\gamma\})$ or $\{\gamma\} \operatorname{IR}_\delta \{k\}|(C \cup B \cup A)$ which by the induction hypothesis implies the corresponding local independencies. To see this, assume that there exists a vertice $k \in C$ for which neither holds. With the trail condition we then have that in $G_{\mathrm{An}(A \cup B \cup C)}$ there exists an allowed trail from A and γ, respectively, to k such that every vertex where edges do not meet head–to–head are not in $(C \cup B \cup \{\gamma\})\backslash\{k\}$ and $(C \cup B \cup A)\backslash\{k\}$, respectively, and every vertex where edges meet head–to–head or some of their descendants are in $(C \cup B \cup \{\gamma\})\backslash\{k\}$ respective $(C \cup B \cup A)\backslash\{k\}$. This would yield a path between A and γ which is not blocked by $C \cup B$ (note that k is a head–to–head node on this trail) in $G_{\mathrm{An}(A \cup B \cup C)}$. This in turn contradicts the separation of A and γ by $B \cup C$ in $G^B_{\mathrm{An}(A \cup B \cup C)}$ because the edges starting in B cannot contribute to this trail. Consequently we can apply right decomposition and get the desired result. □

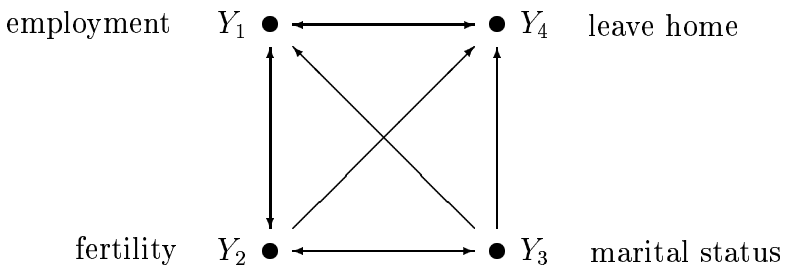
Note that the equivalence of the pairwise and local dynamic Markov properties immediately follows from left intersection assuming (3.5), left weak union and left decomposition. Thus, the above proof mainly aims at situations where A, B, C is no partition of V or $\mathrm{pa}(B) \not\subset C$.

The foregoing theorem also applies to multi–state processes where the marks correspond to the transitions. In this case the mark space possibly contains more elements than the number of vertices since the evolution of one component of the multi–state process may require several different transitions. This is for instance the case when considering a CFMP as in the following example.

Example: Let us again consider the example of a CFMP $\mathbf{Y} = (Y_1, \ldots, Y_4)$ where Y_1 describes the employment status of women, Y_2 the fertility, Y_3 the martial status, and Y_4 indicates whether the person still lives with her parents or not. The postulated pairwise local independencies $Y_4 \not\to Y_2|\mathbf{Y}_{\{1,3\}}$ and $Y_1 \not\to Y_3|\mathbf{Y}_{\{2,4\}}$ as well as $Y_4 \not\to Y_3|\mathbf{Y}_{\{1,2\}}$ result in the local independence graph shown in Figure 4.1. By the local dynamic Markov property we have that $\mathbf{Y}_{\{1,4\}} \not\to Y_3|Y_2$ since $\mathrm{pa}(3) = \{2\}$.

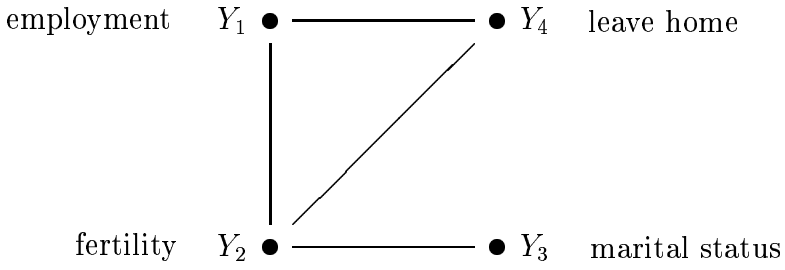
This can also be verified by the global dynamic Markov property.

Figure 4.1: Local independence graph for the CFMP $\mathbf{Y} \sim (Y_1, \ldots, Y_4)$ with $Y_1 \hat{=}$ 'employment status', $Y_2 \hat{=}$ 'fertility', $Y_3 \hat{=}$ 'marital status', $Y_4 \hat{=}$ 'leave home'.



The moral graph $(G^{\{3\}})^m$ is given in the following Figure 4.2. In addition, we get by the latter that $Y_1 \not\to Y_3|Y_2$ and $Y_4 \not\to Y_3|Y_2$, each of these statements involving only three components of the original process. //

Figure 4.2: Moral graph $(G^{\{3\}})^m$ for the above local independence graph.



The proposed local independence graphs are easily generalized to include time constant covariates $\mathbf{X} = (X_1, \ldots, X_R)$. In this case we distinguish between the vertices V_R representing these variables in the graph and the vertices V_P representing the counting processes. The local independence graph G is then given as *dynamic graph* (cf. Definition 1.1.5) such that the subgraph G_{V_R} is the conditional independence graph corresponding to the multivariate distribution of $\mathbf{X}$. This may be an

undirected graph, a DAG, or a chain graph, as appropriate. The global (dynamic) Markov property for this graph then reads as follows: Consider disjoint $A, B, C \subset V$ with $A, B \subset V_R$. Then, the graph separation

$$A \perp\!\!\!\perp_G B \mid C \text{ in } (G_{\mathrm{An}(A\cup B\cup C)})^m$$

should imply the conditional independence

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_{C\cap V_R}, H_\tau^{C\cap V_P},$$

where $H_\tau^{C\cap V_P}$ is the whole observed history of the processes contained in C. If $A \cap V_P \neq \emptyset$ or $B \cap V_P \neq \emptyset$, then δ–separation has to be applied in order to induce the local independencies. This proceeding is justified since the history generated by the whole system is now given by $\tilde{\mathcal{F}}_t = \mathcal{F}_t^{V_R} \vee \mathcal{F}_t^{V_P}$, where $\mathcal{F}_t^{V_R} = \sigma\{\mathbf{X}\}$, $\mathbf{X}$ being realized at time $t = 0$. Thus, Y_k being locally independent of X_j means that the $\tilde{\mathcal{F}}_t$–compensator Λ_k is $\mathcal{F}_t^{V_R\setminus\{j\}} \vee \mathcal{F}_t^{V_P}$–measurable. We restrain from going into further details, here.

4.2 Factorization of the likelihood

For classical conditional independence graphs, the property which is most important for simplifications of estimating and testing procedures is the factorization of the density according to the graph as formalized by (2.2). In this section, we therefore explore whether a similar result can be obtained for local independence graphs, where we restrict ourselves again to marked point processes.

Before deriving the likelihood given a specific local independence graph, we recall the likelihood for the general case. Based on the mark specific compensators $\Lambda_k(t)$ and intensity process $\lambda_k(t)$ the corresponding *crude* quantities are given by summation over all possible events, i.e.

$$\Lambda(t) = \sum_{k=1}^{K} \Lambda_k(t) \quad \text{and} \quad \lambda(t) = \sum_{k=1}^{K} \lambda_k(t).$$

It is easily checked that these are the compensator and intensity processes of the *cumulative counting process*

$$N(t) = \sum_{k=1}^{K} N_k(t).$$

In case that Λ is not absolutely continuous the continuous part is denoted by Λ^C and the jumps by $\Delta\Lambda$, i.e. $\Delta\Lambda(t) = \Lambda(t) - \Lambda(t^-)$ and $\Lambda(t) = \sum_{s \leq t} \Delta\Lambda(s) + \Lambda^C(t)$. The jumps $\Delta N_k(t)$ of the counting processes are defined analogously. Recall that the history process is defined as $H_t = \{(T_s, E_s) | s = 1, \ldots, S, T_s \leq t\}$. And with the subprocesses $H_t^A = \{(T_s, E_s) | s = 1, \ldots, S, T_s \leq t, \text{ and } \exists\, k \in A : E_s = e_k\}$, $A \subset V$, we have that $\mathcal{F}_t^A = \sigma\{H_t^A\}$. With these notations the likelihood process is given as follows.

Remark 4.2.1 *Likelihood for marked point processes (cf. Arjas, 1989)*
Let $\{(T_s, E_s) | s = 1, \ldots, S\}$ be a marked point process with mark space $\mathcal{E} = \{e_1, \ldots, e_K\}$ and associated counting processes $N_k(t)$, $k = 1, \ldots, K$, and history process H_t. Then, the *likelihood process* $L(t|H_t)$ is given as follows

$$L(t|H_t) = \prod_{T_s \leq t} \Lambda(dT_s; E_s) \cdot \prod_{s \leq t} (1 - \Delta\Lambda(s))^{1 - \Delta N(s)} \cdot \exp(-\Lambda^C(t)).$$

In the absolutely continuous case this simplifies to

$$L(t|H_t) = \prod_{T_s \leq t} \lambda(T_s; E_s) \cdot \exp\left(- \int_0^t \lambda(s)\, ds \right).$$

//

As noted by Arjas (1989), the above likelihood for marked point processes has already a product form over time, i.e. it is constructed as the product of the likelihood for a present occurrence or non–occurrence of an event given the past history. This is obviously very similar to the block recursive factorization for chain graph models (cf. part (1) of Corollary 2.1.7). In order to show how this can further be factorized according to a local independence graph let us first consider the simple situation of right censored survival data where the factorization that we are interested in is already well–known.

Definition 4.2.2 *Right censored survival data*
Let e_1 be the event of failure or death and e_2 the event of censoring, i.e. the mark space is given as $\mathcal{E} = \{e_1, e_2\}$. Both events are non–recurrent so that the marked point process is uniquely determined by the random times T_1 and T_2 of occurrences of both events, respectively. If censoring occurs before failure, i.e. $T_2 \leq T_1$, then

the failure time T_1 is not observable. The observable data for a single observational unit therefore consists of the time when it is last observed denoted by T^* and an indicator δ with $\delta = 1$ if the individual failed at T^* and $\delta = 0$ if it was censored. //

Let $\Lambda_k(t)$, $k = 1, 2$, be the mark specific $\mathcal{F}_t$–compensators, where $\mathcal{F}_t$ is the filtration generated by $\{T_1, T_2\}$ and assume that they are absolutely continuous. In case of right censored survival data the history process has a very simple structure. If $t < T^*$ then $H_t = \emptyset$ and otherwise $H_t = \{(T^*, E_{T^*})\}$, where $E_{T^*} = e_1$ if $\delta = 1$ and $E_{T^*} = e_2$ if $\delta = 0$. According to Remark 4.2.1 the likelihood $L(t|H_t)$ for a single observation and for $t \geq T^*$ is thus given as

$$L(t|H_t) = \Lambda_1(dT^*)^{\delta}\Lambda_2(dT^*)^{1-\delta} \cdot \exp(-\Lambda(t)). \tag{4.5}$$

If $t < T^*$ the first two terms in the above formula have to be removed. Using the definition of the crude compensator as sum over all mark specific compensators and writing $\Lambda_k(t) = \Lambda_k(t|H_{t-})$ to emphasize the dependence on the history, the above likelihood (4.5) can be reformulated as

$$L(t|H_t) = L_1(t|H_t)L_2(t \mid H_t) \tag{4.6}$$

with

$$\begin{aligned} L_1(t|H_t) &= \Lambda_1(dT^* \mid H_{T^*-})^{\delta} \cdot \exp(-\Lambda_1(t \mid H_{t-})), \\ L_2(t|H_t) &= \Lambda_2(dT^* \mid H_{T^*-})^{1-\delta} \cdot \exp(-\Lambda_2(t \mid H_{t-})). \end{aligned}$$

Thus, we have that the likelihood is a product over two factors $L_k(t|H_t)$, $k = 1, 2$, each depending only on the e_k specific compensator $\Lambda_k(t)$. But note that the latter is still a $\mathcal{F}_t$–compensator, i.e. it may depend on whether the other event has previously occurred or not. Let $\mathcal{F}_t^1$ be the subfiltration generated by T_1, only, and assume that $\Lambda_1(t)$ is $\mathcal{F}_t^1$–measurable for all $t \in \mathcal{T}$ so that $\Lambda_1(dt|H_{t-}) = \Lambda_1(dt|H_{t-}^1)$. Then, we get that

$$L_1(t|H_t) = L_1(t|H_t^1), \tag{4.7}$$

where

$$H_t^1 = \begin{cases} \emptyset, & t < T^* \vee \delta = 0 \\ (T^*, e_1), & \text{otherwise,} \end{cases}$$

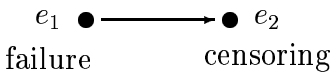
i.e. the first factor of the likelihood (4.6) only depends on the information about the possible prior occurrence of e_1. This assumption is known as *independent censoring* (cf. Andersen et al., 1993, p. 139) and is obviously equivalent to e_1 being locally independent of e_2. Heuristically, one could say that the censored individuals should not be more or less likely to fail than the others, i.e. the prior occurrence of the censoring event should not change the intensity for a subsequent failure. Since it is not possible to observe failure after censoring it is helpful to make sure that the censoring mechanism is independent because otherwise inference about $\Lambda_1(t)$ becomes quite difficult (e.g. Parner and Keiding, 1998). In addition, (4.7) is the usual likelihood expression for censored survival data when interested in inference about $\Lambda_1(t)$ omitting $L_2(t|H_t)$. If the latter depends on the parameter of interest, expression (4.7) is called a *partial likelihood* (Cox, 1975). Note that the information contributed to the (partial) likelihood (4.7) by a censored observation is $\exp(-\Lambda_1(t))$, so that the foregoing argumentation does not imply that censored observations are discarded.

The above motivation shows that given a local independence graph $G = (V, E)$ for two events with $V = \{1, 2\}$ and $E = \{(1, 2)\}$ (cf. Figure 4.3) we can find a mark specific factorization of the likelihood such that

$$L(t|H_t) = L_1(t|H_t^{\mathrm{cl}(1)})L_2(t|H_t^{\mathrm{cl}(2)}),$$

where $\mathrm{cl}(1) = \{1\}$ and $\mathrm{cl}(2) = \{1, 2\}$.

Figure 4.3: Local independence graph for independent censoring.



This is generalized to local independence graphs for marked point processes in the following theorem.

Theorem 4.2.3 *Factorization for local independence graphs*
Consider a marked point process $\{(T_s, E_s)|s = 1, \ldots, S\}$ with mark space $\mathcal{E} = \{e_1, \ldots, e_K\}$ and corresponding local independence graph $G = (V, E)$ with $V =$

$\{1, \ldots, K\}$. Assume that condition (3.5) holds. Then, the likelihood factorizes as follows:

$$L(t|H_t) = \prod_{k \in V} L_k(t|H_t^{\mathrm{cl}(k)}), \tag{4.8}$$

where

$$L_k(t|H_t^{\mathrm{cl}(k)}) = \prod_{T_{s(k)} \leq t} \Lambda_k(dT_{s(k)}) \cdot \prod_{s \leq t} (1 - \Delta\Lambda_k(s))^{1-\Delta N_k(s)} \cdot \exp(-\Lambda_k^C(t))$$

and $\{T_{s(k)} \mid s(k) \in \{1, \ldots, S\}, E_{s(k)} = e_k\}$ are the occurrence times of event e_k.

Proof:

First, we consider the absolutely continuous case. The argumentation is similar to the above example of right censored survival times. The likelihood of a general marked point process (cf. Remark 4.2.1) can be transformed as follows:

$$\begin{aligned}
L(t|H_t) &= \prod_{T_s \leq t} \lambda(T_s; E_s) \cdot \exp\left(-\int_0^t \lambda(s)\, ds\right) \\
&= \prod_{k=1}^{K} \prod_{T_s \leq t} \lambda_k(T_s)^{\mathbf{1}\{E_s = e_k\}} \cdot \exp\left(-\int_0^t \sum_{k=1}^{K} \lambda_k(s)\, ds\right) \\
&= \prod_{k=1}^{K} \left(\prod_{T_{s(k)} \leq t} \lambda_k(T_{s(k)}) \cdot \exp\left(-\int_0^t \lambda_k(s)\, ds\right) \right),
\end{aligned}$$

where $T_{s(k)}$ is defined as above. Now, by definition of a local independence graph and the equivalence of the pairwise and local dynamic Markov properties under condition (3.5) we have that $\lambda_k(s)$ is $\mathcal{F}_t^{\mathrm{cl}(k)}$–measurable. It follows that

$$\begin{aligned}
L_k(t|H_t) &= \prod_{T_{s(k)} \leq t} \lambda_k(T_{s(k)}) \cdot \exp\left(-\int_0^t \lambda_k(s)\, ds\right) \\
&= L_k(t|H_t^{\mathrm{cl}(k)}),
\end{aligned} \tag{4.9}$$

i.e. the (partial) likelihood L_k based on the whole past remains the same if the available information is restricted to how often those events that are parents of e_k in the graph and e_k itself have occurred in the past, symbolized by $H_t^{\mathrm{cl}(k)}$.

If the compensator is not absolutely continuous we additionally have to show that

$$\prod_{s\leq t}(1-\Delta\Lambda(s))^{1-\Delta N(s)} = \prod_{k=1}^{K}\prod_{s\leq t}(1-\Delta\Lambda_k(s))^{1-\Delta N_k(s)}.$$

To see this, recall that $\Delta\Lambda_k(t) = E(\Delta N_k(t)|\mathcal{F}_{t^-}) = P(\Delta N_k(t)=1|\mathcal{F}_{t^-})$ so that a discontinuity $\Delta\Lambda_k(t)\neq 0$ implies that there is a positive probability for event e_k to occur at time t. Since we make the assumption that no two events occur at the same time we have that for any $t\in\mathcal{T}$ where $\Delta\Lambda(t)\neq 0$ there exists an $e_k\in\mathcal{E}$ such that $\Delta\Lambda(t)=\Delta\Lambda_k(t)$. □

An interesting consequence of the above theorem regarding the relation of local and conditional independence has been noted by Schweder (1970, Theorems 3 and 4) for CFMPs and is formulated for the general case of marked point processes in the following proposition.

Corollary 4.2.4 *Conditional independence of histories*
In the situation of the above Theorem 4.2.3 we have for disjoint $A,B,C\subset V$, where C separates A and B, i.e. $A\perp\!\!\!\perp_G B|C$, in $(G_{\mathrm{An}(A\cup B\cup C)})^m$ that

$$\mathcal{F}_t^A \perp\!\!\!\perp \mathcal{F}_t^B \mid \mathcal{F}_t^C \quad \forall\, t\in\mathcal{T}.$$

Proof:
From (4.8) it follows that the marginal likelihood for the marked point process discarding events not in $\mathrm{An}(A\cup B\cup C)$, i.e. $\{(T_s,E_s)|s=1,\ldots,S';E_s\in\mathcal{E}_{\mathrm{An}(A\cup B\cup C)}\}$, is given by just integrating out the contributions of $e_k\notin\mathcal{E}_{\mathrm{An}(A\cup B\cup C)}$ yielding

$$L(t|H_t^{\mathrm{An}(A\cup B\cup C)}) = \prod_{k\in\mathrm{An}(A\cup B\cup C)} L_k(t|H_t^{\mathrm{cl}(k)}).$$

Further, the likelihood may be written as product over factors that only depend on $\mathrm{cl}(k)=\{k\}\cup\mathrm{pa}(k)$, $k\in\mathrm{An}(A\cup B\cup C)$. Let $\mathcal{C}=\{\mathrm{cl}(k)|k\in\mathrm{An}(A\cup B\cup C)\}$ be the set containing all such sets. Then, we have

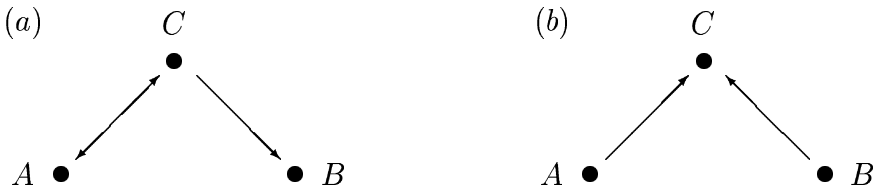
$$L(t|H_t^{\mathrm{An}(A\cup B\cup C)}) = \prod_{c\in\mathcal{C}} L_c(t|H_t^c).$$

The sets $\mathcal{C}$ are the cliques of the graph $(G_{\mathrm{An}(A\cup B\cup C)})^m$. Thus, the likelihood of $H_t^{\mathrm{An}(A\cup B\cup C)}$ factorizes in the way required by the factorization theorem (2.2). From

this follows the claimed conditional independence. □

Heuristically, the conditional independence $\mathcal{F}_t^A \perp\!\!\!\perp \mathcal{F}_t^B | \mathcal{F}_t^C$, $t \in \mathcal{T}$, means that at any time t the histories of the processes $\mathbf{Y}_A$ and $\mathbf{Y}_B$ are conditionally independent given the whole history of $\mathbf{Y}_C$ up to time t. Note that mutual local independence of $\mathbf{Y}_A$ and $\mathbf{Y}_B$ is a necessary but not sufficient prerequisite. In order to get a better insight into the above proposition, let us consider the two paradigmatic situations given in Figure 4.4.

Figure 4.4: Examples for Corollary 4.2.4.



In the situation of Figure 4.4 (a) we have that $\mathcal{F}_t^A \perp\!\!\!\perp \mathcal{F}_t^B | \mathcal{F}_t^C$ since the moral graph is given by just replacing the directed edges by undirected ones. This conditional independence is plausible since any effect of A on B is conveyed by C whereas B has no effect on A at all so that there is no backward causation. In contrast, Figure 4.4 (b) shows a situation where $\mathcal{F}_t^A \not\perp\!\!\!\perp \mathcal{F}_t^B | \mathcal{F}_t^C$. This is again plausible since C is a common 'consequence' of A as well as B so that conditioning on the past of $\mathbf{Y}_C$ might yield a dependence between the past of $\mathbf{Y}_A$ and $\mathbf{Y}_B$ although they are marginally independent. Thus, the interpretation of local independence graphs regarding conditional independence of the involved histories is similar as for DAGs despite the cycles in the graph (cf. Proposition 2.1.4).

Note that in the above corollary we implicitly assumed that the processes are unrelated at time $t = 0$. This might not be the case if there are time constant covariates realized at $t = 0$ simultaneously affecting different counting processes. These should then be included in in the condition, i.e. $\mathcal{F}_t^C$ has to be enlarged by a suitable $\mathcal{F}_0$.

The heuristic interpretation of local independence already mentioned just after its

definition, namely that $A \not\to B|C$ means that the presence of $\mathbf{N}_B$ is independent of the past of $\mathbf{N}_A$ given its own past and the one of $\mathbf{N}_C$, can now be shown to be true.

Remark 4.2.5 *Local independence as conditional independence*
In particular, it follows from the factorization (4.8) of the likelihood that $A \not\to B|C$ implies

$$N_B(t) \perp\!\!\!\perp \mathcal{F}_{t^-}^{V\setminus \mathrm{cl}(B)} \mid \mathcal{F}_{t^-}^{\mathrm{cl}(B)}$$

as already conjectured in Chapter 3. This, again, holds by the general factorization property (2.2). //

Further implications of the factorization of the likelihood are discussed in Chapter 5 in the context of estimation and statistical tests for local independence graphs.

4.3 Collapsibility

Collapsibility of graphical models means that the independence structure of a subset of the variables is properly portrayed by the corresponding subgraph, i.e. discarding the other variables does not induce any additional dependencies (cf. Frydenberg, 1990a). Obviously, this property is central to any reduction of complexity and heavily based on the separation Theorem 4.1.4. Therefore, we explore in this section how the previous findings can be used to formulate conditions for an analogous concept regarding local independence graphs. Let us first define collapsibility for local independence graphs.

Definition 4.3.1 *Collapsibility*
Let $G = (V, E)$ be the local independence graph of a multivariate process $\mathbf{Y}_V = (Y_1, \ldots, Y_K)$. For $A \subset V$, we say that $\mathbf{Y}_V$ is *(weakly) collapsible onto* A (or *over* $B = V\setminus A$) if the subgraph G_A is the local independence graph of the subprocess $\mathbf{Y}_A = \{Y_k | k \in A\}$. Further, if it holds that the $\mathcal{F}_t$–compensator $\Lambda_A(t)$ of $\mathbf{Y}_A$ is $\mathcal{F}_t^A$–measurable we say that $\mathbf{Y}_V$ is *strongly collapsible onto* A. //

In the following, we consider conditions for collapsibility of general marked point processes. Conditions for strong collapsibility of course depend on the chosen model and are therefore only treated for very special situations, below.

4.3.1 Weak collapsibility

According to the definition of a local independence graph, collapsibility onto $A \subset V$ implies for any $j, k \in A$ with $(j,k) \notin E$ that $\{j\} \not\to \{k\} | A \backslash \{j,k\}$, i.e. that marginalizing w.r.t $B = V \backslash A$ induces no additional association among the vertices in A. This holds by the global dynamic Markov property if $A \backslash \{j,k\}$ δ–separates $\{j\}$ from $\{k\}$ in the whole local independence graph G. In order to formulate graphical conditions to verify this separation we need some more terminology.

Definition 4.3.2 *Complete subgraph / connected component*
Let $G = (V, E)$ be a directed (not necessarily acyclic) graph where $(k,j), (j,k) \in E$ is allowed. For $A \subset V$ the subgraph G_A (or simply A) is called

(1) *complete* if $(j,k) \in E$ for all $j, k \in A$, $j \neq k$, i.e. there are bidirected edges between any two vertices in A.

(2) a *connected component* if there exists a trail between any two vertices $j, k \in A$ but no trail between any $j \in A$ and $k \in V \backslash A$. //

Note that with the above definition the connected components of a (sub)graph always constitute a partition of this (sub)graph. We can now prove the following result.

Theorem 4.3.3 *Weak collapsibility for local independence graphs*
Let $G = (V, E)$ be the local independence graph of a multivariate counting process $\mathbf{N}_V = (N_1, \ldots, N_K)$ with all assumptions of Theorem 4.1.4 fulfilled. For $A \subset V$ and $B = V \backslash A$, let $B_1, \ldots, B_L$ be the connected components of B and $B'_l = B_l \cap \text{an}(A)$. The process $\mathbf{N}$ is collapsible onto A if the following conditions hold: For every B'_l, $l = 1, \ldots, L$,

(1) $\text{ch}(B'_l) \cap A$ is complete and

(2) for every $k \in \text{ch}_{G_{\text{An}(A)}}(B'_l)$ and every $j \in A$ with $j \notin \text{pa}(k)$:

$$\text{ch}_{G_{\text{An}(A)}}(j) \cap (\text{ch}_{G_{\text{An}(A)}}(B'_l) \cup B'_l) = \emptyset.$$

Proof:

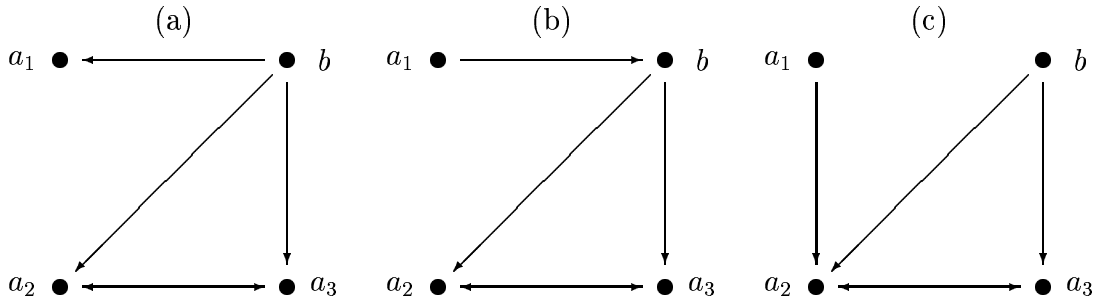
Without loss of generality we can suppose that $L = 1$ and that $B \subset \mathrm{an}(A)$. The latter is justified since $k \notin \mathrm{an}(A)$ does not affect the local independence structure of A, anyway.

Assume that the above conditions hold and let $j, k \in A$ with $(j, k) \notin E$. We have to show that $A \backslash \{j, k\}$ δ–separates $\{j\}$ from $\{k\}$. Since $\mathrm{ch}(B)$ is complete it follows that either (1) $k \notin \mathrm{ch}(B)$ or (2) $j \notin \mathrm{ch}(B)$. In case of (1) we have $B \not\to \{k\} | A \backslash \{j, k\}$. In addition, $\{j\} \not\to \{k\} | V \backslash \{j, k\}$ is equivalent to $\{j\} \not\to \{k\} \,| (A \backslash \{j, k\}) \cup B$ yielding with left intersection according to (1.3) and left decomposition that $\{j\} \not\to \{k\} | A \backslash \{j, k\}$. If, in contrast, $k \in \mathrm{ch}(B)$ then it must hold that $j \notin \mathrm{ch}(B)$. Further, the second condition of the above theorem ensures that any allowed trail from j to k is blocked by $A \backslash \{j, k\}$. This can be seen as follows. In addition to $j \notin \mathrm{ch}(B)$ we have that $\mathrm{ch}(j) \cap B = \emptyset$. Thus, there is no edge in either direction between B and $\{j\}$ in G. Finally, $\{j\}$ and B have no common children, so that marrying parents does not induce an edge between $\{j\}$ and B in $(G^k_{\mathrm{An}(A)})^m$. Therefore, no path between k and j in $(G^k_{\mathrm{An}(A)})^m$ contains vertices in B. □

Unfortunately, the conditions given in the above theorem are not easy to verify even in small local independence graphs. To illustrate this, let us consider some examples.

Examples: Let $G = (V, E)$ with $V = \{a_1, a_2, a_3, b\}$ and $A = \{a_1, a_2, a_3\}$, $B = \{b\}$. Figure 4.5 shows three local independence graphs that are not collapsible onto A (note that for all graphs we have $\mathrm{An}(A) = V$).

Figure 4.5: Examples for local independence graphs which are not collapsible onto $A = \{a_1, a_2, a_3\}$.



In situation (a), $\mathrm{ch}(B)$ is not complete. Marginalizing w.r.t. B would therefore possibly induce additional associations among A. In situation (b), we have that $a_1 \notin \mathrm{pa}(a_2, a_3)$, where $\mathrm{ch}(B) = \{a_2, a_3\}$, but $\mathrm{ch}(a_1) \cap B \neq \emptyset$. This implies that marginalization of B possibly yields an influence of a_1 on a_2 and a_3. Finally, in situation (c), $\mathrm{ch}(a_1) \cap \mathrm{ch}(B) \neq \emptyset$ so that due to the marrying parents effect an association between a_1 and a_3 could result from marginalizing w.r.t. B.
The following Figure 4.6 shows two local independence graphs where collapsibility onto $A = \{a_1, a_2, a_3\}$ holds. In situation (a), it is easily checked that for the whole graph $a_1 \not\to a_2 | a_3$ and $a_3 \not\to a_1 | a_2$ without involving b so that the process is collapsible onto A.

Figure 4.6: Examples for local independence graphs which are collapsible onto $A = \{a_1, a_2, a_3\}$.

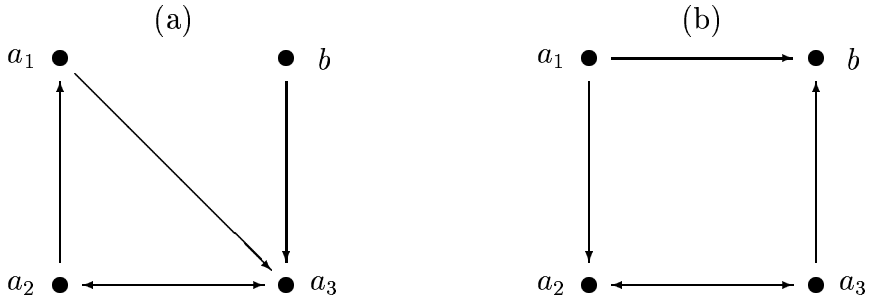


Figure 4.6 (b) shows a situation where $B = \{b\}$ is no element of $\mathrm{an}(A)$. Thus, marginalization does not affect the local independence structure of the remaining components. Let now in situation (b) $A = \{a_1, a_2\}$ and $B = \{b, a_3\}$. Then, we have again that $b \notin \mathrm{an}(A)$ so that marginalizing w.r.t. b has no effect on the other processes even though it is a common child of a_1 and a_3. Further, $a_2 \not\to a_1$ so that a_3 can also be discarded. //

The above theorem gives conditions for weak collapsibility. It would, however, be desirable to have a graphical characterization which is *equivalent* to collapsibility. This is not possible without assuming a more specific model class and would require to define collapsibility for model classes. For instance, if $\mathcal{M}(G)$ denotes the class of all composable Markov processes on G then collapsibility of $\mathcal{M}(G)$ onto A would

imply that the class $\mathcal{M}(G)_A$ obtained by marginalizing $\mathcal{M}(G)$ w.r.t. $V \backslash A$ is the same as the class $\mathcal{M}(G_A)$ of all composable Markov processes on G_A. However, a more detailed treatment of this topic is beyond the scope of this thesis.

4.3.2 Strong collapsibility

The condition for strong collapsibility given in the following theorem is straightforward. It generalizes Theorem 2 of Schweder (1970) to the situation of general marked point processes which are not necessarily associated to composable Markov processes.

Theorem 4.3.4 *Strong collapsibility*
Let $G = (V, E)$ be the local independence graph of a multivariate counting process $\mathbf{N}_V = (N_1, \ldots, N_K)$ with all assumptions of Theorem 4.1.4 fulfilled. Consider a subset $A \subset V$ with $\mathrm{pa}(A) = \emptyset$, i.e. A is an ancestral set. The local independence graph of $\mathbf{N}_A$ is then given as the subgraph $G_A = (A, E_A)$ of G and the $\mathcal{F}_t$–compensator $\Lambda_A(t)$ is $\mathcal{F}_t^A$–measurable.

Proof:
If $\mathrm{pa}(A) = \emptyset$, the $\mathcal{F}_t$–compensators $\Lambda_t^A(t)$ are due to the definition of local independence and with property (3.5) $\mathcal{F}_t^A$–measurable. □

The above theorem provides an easily checked graphical condition for strong collapsibility since an ancestral set can readily be identified by verifying that there are no edges from outside pointing at this set.

A consequence is that for a CFMP $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ with local independence graph G we have: If $\mathrm{pa}(A) = \emptyset$ then the subprocess Y_A is again a Markov process for $A \subset V$. Thus, with the notation mentioned above $\mathcal{M}(G)_A \subset \mathcal{M}(G_A)$ for any ancestral set $A \subset V$. As demonstrated by Schweder (1970) the converse is not necessarily true. Assume for instance that $\mathbf{Y} \sim (Y_1, Y_2)$ and $\mathcal{S}_1 = \{0, 1\}$ where the state $y_1 = 1$ is absorbing. Then, we have that Y_1 is also a Markov process even if it is not locally independent of Y_2. To see this, let $t_1 < t_2 < \ldots < t_n$. Since $Y(t_j) = 0$ implies $Y(t_i) = 0$ for all $i = 0, 1, \ldots, j$, we have that

$$P(Y_1(t_i) = y \mid Y_1(t_{i-1}) = 0, \ldots, Y_1(t_1) = 0) = P(Y_1(t_i) = y \mid Y_1(t_{i-1}) = 0)$$

for $y = 0, 1$. In addition,

$$\begin{aligned} P(Y_1(t_i) = y \mid Y_1(t_{i-1}) = 1, Y_1(t_{i-2}) = y_{i-2}, \ldots, Y_1(t_1) = y_1) &= \\ P(Y_1(t_i) = y \mid Y_1(t_{i-1}) = 1) &= y, \quad y = 0, 1. \end{aligned}$$

Consequently, Y_1 is in any case a Markov process. This mainly seems to be due to the absorbing state which makes it easy to determine the past or future knowing a present value $Y_1(t) = 0$ or $Y_1(t) = 1$, respectively, so that conditioning on the past becomes redundant. Survival analysis being one of the main applications for event history analysis, we have to keep the foregoing example in mind since failure or death always constitutes such a transition into an absorbing state. The following section addresses this issue in more detail.

4.4 Local independence graphs for stopped processes

When the mark space of the considered marked point process contains an event that constitutes a transition into an absorbing state such as death or failure then the remaining components typically depend on this specific one. Consider for instance the situation where various events concerning the health status are measured such as onset of different side effects of a medicamentation or of another intervention and the time of death. Then, it is obvious that no other events will take place after death implying that the intensities of the former become zero when death occurs. Since the corresponding local independence graph shows these local dependencies by arrows the graph might become quite complex. Besides, the information that death affects the intensities for all remaining events is not very interesting. A way to cope with this situation is to consider *stopped processes.* Applied to our example this means that the marked point process is restricted to the time before death including the time when it occurs. Thus, only those local dependencies remain valid for the stopped process which operate before this absorbing event. Note, that this is not the same as conditioning on the person being alive up to a certain time since the time of death remains unknown and random.

In order to treat the properties of local independence graphs for stopped processes more formally we introduce some notation.

Definition 4.4.1 *Stopping time / stopped process*
Let $\mathcal{F}_t$ be a filtration on $(\Omega, \mathcal{F}, P)$. A nonnegative random variable T on $(\Omega, \mathcal{F}, P)$is called $\mathcal{F}_t$*–stopping time* if $\{T \leq t\} \in \mathcal{F}_t$ for all $t \in \mathcal{T}$. Let further Y be a $\mathcal{F}_t$–adapted process and T a $\mathcal{F}_t$–stopping time. Then, the process $Y^T = \{Y(t \wedge T) | t \in \mathcal{T}\}$ is called the process *stopped at* T.
In case that $\mathcal{F}_t$ is generated by a CFMP $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ we define the time when component Y_k is in state y_k for the first time as $T(y_k) = \inf\{t | N(t; (\cdot, y_k)) = 1\}$, where $N(t; (\cdot, y_k)) = \sum_{\mathbf{y}} N_k(t; (\mathbf{y}, y_k))$.
Similarly, if $\mathcal{F}_t$ is generated by a marked point process $\{(T_s, E_s)\}$, $\mathcal{E} = \{e_1, \ldots, e_K\}$, the random variables $T(k) = \inf\{t | N_k(t) = 1\}$ are the times of the first occurrence of e_k, $k = 1, \ldots, K$. //

Obviously, $T(y_k)$ and $T(k)$ are $\mathcal{F}_t^k$– as well as $\mathcal{F}_t$– stopping times.
For our purposes it is important to note that the compensators and intensity processes of stopped processes are simply given as the stopped compensators and intensity processes of the original process.

Remark 4.4.2 *Compensator and intensity of stopped process*
If a process Y has compensator Λ then the stopped process Y^T has the compensator Λ^T. Thus, in the absolutely continuous case the intensity process of Y^T is given as $\lambda(t)\mathbf{1}\{t \leq T\}$ (Andersen et al., 1993, p. 81). //

For a stopped CFMP $\mathbf{Y}^{T(y_k)}$ it is obvious that each component locally depends on Y_k since the intensity process requires the information on the occurrence of y_k. Thus, it makes no sense to retain the original definition of local independence for stopped processes. Instead, we propose to speak of local independence when the required information on the previous occurrence of y_k just factors out through an indicator function.

Definition 4.4.3 *Local independence for stopped process*
Let $\mathbf{Y} = (Y_1, \ldots, Y_K)$ be a multivariate stochastic process and T a $\mathcal{F}_t^k$–stopping time. For disjoint $A, B, C \subset V$ such that $k \notin A \cup C$ we say that the local independence

$A \not\to B|C$ holds for the stopped process $\mathbf{Y}^T$ if there exists a $\mathcal{F}_t^{B\cup C}$–measurable predictable process $\tilde{\Lambda}_B$ such that the compensator of $\mathbf{Y}_B^T$ is given as product $\Lambda_B^T(t) = \tilde{\Lambda}_B(t)\mathbf{1}\{t \leq T\}$. //

Heuristically, the above definition applies to situations, where the original compensator $\Lambda_B(t)$ can be specified as a function of $H_t^{B\cup C}$ as long as $t \leq T$. Then, the process $\tilde{\Lambda}_B(t)$ can just be chosen as this function. Note that local independence for stopped processes holds the same asymmetric graphoid properties as the original local independence as far as $k \notin A \cup C$, i.e. as far as the component w.r.t. which the stopping time is defined is neither an element of the separating set nor of the set which is claimed to be irrelevant. When interpreting these stopped processes we have to take into account that all involved intensities are conditional on $t \leq T$. Thus, $A \not\to B|C$ for a stopped process means that A is irrelevant for B given C as long as $t \leq T$. The local independence graph for stopped processes is readily given by deleting the directed edges starting in k as described for the special case of a CFMP in the next proposition.

Proposition 4.4.4 *Local independence graph for stopped CFMP*
Let $\mathbf{Y} \sim (Y_1, \ldots, Y_K)$ be a CFMP with local independence graph $G = (V, E)$. Assume that for a state y_k^* of Y_k with $P(Y_k(0) = y_k^*) = 0$ it holds that the transition intensities regarding changes in component $Y_j(t)$ are constant for different values of $Y_k(t^-)$ as long as $Y_k(t^-) \neq y_k^*$. Formally, this means that for all $j \in \mathrm{ch}(k)$, $\mathbf{y}_{\mathrm{cl}(j)\setminus k} = \tilde{\mathbf{y}}_{\mathrm{cl}(j)\setminus k} \in \mathcal{S}_{\mathrm{cl}(j)\setminus k}$, and for all $y_j' \in \mathcal{S}_j$, $y_j' \neq y_j$:

$$\alpha_j(t; (\mathbf{y}_{\mathrm{cl}(j)}, y_j')) = \alpha_j(t; (\tilde{\mathbf{y}}_{\mathrm{cl}(j)}, y_j')) \quad \forall\, y_k, \tilde{y}_k \in \mathcal{S}_k;\ \tilde{y}_k \neq y_k;\ y_k, \tilde{y}_k \neq y_k^*. \quad (4.10)$$

Then, the stopped process $\mathbf{Y}^{T(y_k^*)}$ has the modified local independence graph $G^{T(y_k^*)} = (V, E^{T(y_k^*)})$ with $E^{T(y_k^*)} = \{(i,j) \in E | i \neq k\}$.

Proof:
In the following $\mathrm{cl}(\cdot)$ and $\mathrm{ch}(\cdot)$ refer to the original local independence graph G. Due to the equalities (4.10) we may choose $\tilde{\alpha}_j(t; (\mathbf{y}_{\mathrm{cl}(j)\setminus k}, y_j'))$ as the value $\alpha_j(t; (\mathbf{y}_{\mathrm{cl}(j)}, y_j'))$ which is constant in the k-th component as long as $y_k \neq y_k^*$, $j \in \mathrm{ch}(k)$. Then, the intensities for the stopped process are given as

$$\lambda_j(t; (\mathbf{y}_{\mathrm{cl}(j)}, y_j')) = \alpha_j(t; (\mathbf{y}_{\mathrm{cl}(j)}, y_j'))\mathbf{1}\{\mathbf{Y}_{\mathrm{cl}(j)}(t^-) = \mathbf{y}_{\mathrm{cl}(j)}\}\mathbf{1}\{t \leq T(y_k^*)\}, \quad j \notin \mathrm{ch}(k),$$

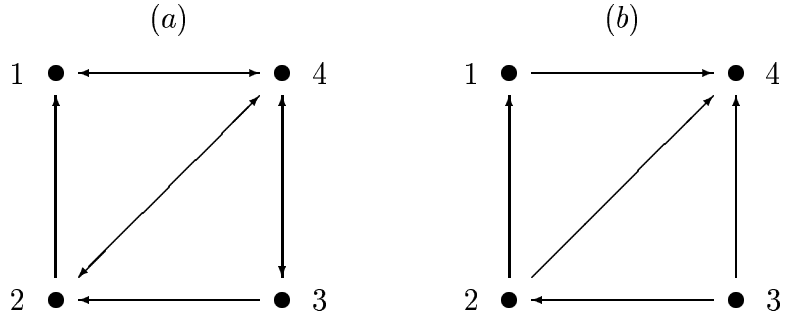
and for $j \in \mathrm{ch}(k)$:

$$\lambda_j(t;(\mathbf{y}_{\mathrm{cl}(j)\setminus k}, y_j')) = \tilde{\alpha}_j(t;(\mathbf{y}_{\mathrm{cl}(j)\setminus k}, y_j'))\mathbf{1}\{\mathbf{Y}_{\mathrm{cl}(j)\setminus k}(t^-) = \mathbf{y}_{\mathrm{cl}(j)\setminus k}\}\mathbf{1}\{t \leq T(y_k^*)\}.$$

Thus, the latter intensities can be written as required by the above definition with $\tilde{\lambda}_j(t;(\mathbf{y}_{\mathrm{cl}(j)\setminus k}, y_j') = \tilde{\alpha}_j(t;(\mathbf{y}_{\mathrm{cl}(j)\setminus k}, y_j')\mathbf{1}\{\mathbf{Y}_{\mathrm{cl}(j)\setminus k}(t^-) = \mathbf{y}_{\mathrm{cl}(j)\setminus k}\}$ which is $\mathcal{F}_t^{\mathrm{cl}(j)\setminus k}$–measurable, $j \in \mathrm{ch}(k)$. The former intensities for $j \notin \mathrm{pa}(k)$ do not depend on Y_k anyway. This yields the desired result. □

Example: Let us consider the situation of $\mathbf{Y} \sim (Y_1, \ldots, Y_4)$ with $\mathcal{S}_k = \{0, 1\}$, $k = 1, \ldots, 4$, where each component indicates whether an event e_k has occurred, i.e. $\mathbf{Y}$ is itself a multivariate counting process.

Figure 4.7: Local independence graph (a) for the whole CFMP $\mathbf{Y} \sim (Y_1, \ldots, Y_4)$ and (b) for the stopped process $\mathbf{Y}^{T(4)}$.



For one component, say Y_4, we assume that the state $y_4 = 1$ is such that $(y_1, y_2, y_3, 1)$, $y_1, y_2, y_3 \in \{0, 1\}$, are absorbing states. It might for instance indicate death of the patient whereas Y_1, Y_2, Y_3 indicate the occurrence of different side effects or diseases. We therefore have that all three side effects locally depend on death. Assume further that $Y_3 \not\to Y_1|\mathbf{Y}_{\{2,4\}}$, $Y_1 \not\to Y_2|\mathbf{Y}_{\{3,4\}}$, and $\mathbf{Y}_{\{1,2\}} \not\to Y_3|Y_4$. The local independence graph of $\mathbf{Y}$ is given in Figure 4.7 (a). The corresponding graph for the stopped process (Figure 4.7 (b)) shows that for instance $\mathbf{Y}_{\{1,2\}} \not\to Y_3$ as long as the person is alive, i.e. the side effect no. 3 remains unaffected by the occurrences of the other side effects or diseases. Analogously, the graph for the original process shows that $\mathbf{Y}_{\{1,2\}} \not\to Y_3|Y_4$. The graph for the stopped process thus reveals no new independencies but makes explicit that we always condition on the person being alive.

Further, we have for example by Corollary 4.2.4 that $\mathcal{F}_t^1 \perp\!\!\!\perp \mathcal{F}_t^3 | \mathcal{F}_t^2$ for $t \leq T(4)$. Note that this is not the same as $\mathcal{F}_t^1 \perp\!\!\!\perp \mathcal{F}_t^3 | \mathcal{F}_t^{\{2,4\}}$ since conditioning on a specific history $\mathcal{F}_t^4$ causes $\mathcal{F}_t^1$ and $\mathcal{F}_t^3$ to be dependent as can be seen from the likelihood. For instance, if both diseases Y_1 and Y_3 increase the intensity for death and we know that death has not occurred up to a certain time t then the information that disease no. 1 has occurred before t makes it less probable that disease no. 3 has also occurred before t implying that the histories $\mathcal{F}_t^1$ and $\mathcal{F}_t^3$ are dependent. This is also the reason why we choose, in contrast to Schweder (1970), to include in the graph the component w.r.t. which the stopping time is defined, here Y_4. //

The concept of local independence graphs for stopped processes is of course also applicable to more general situations than Markov processes as long as the intensity processes can be written as claimed by Definition 4.4.3. This might eventually be difficult if the dependence on the past includes a duration dependence but we desist from going into details, here.

Although local independence graphs for stopped processes reveal no really new independencies they have the advantage of a clearer graphical representation without loosing information. In addition, they meet the intuition that death or failure only affect the remaining processes by preventing further transitions but without any substantial meaning. Therefore, it might be reasonable to replace the original graph by a graph for the stopped processes whenever there are such terminal events.

4.5 Causal reasoning in local independence graphs

Local independence graphs, as introduced so far, are a priori no causal models since they only describe certain forms of dependencies on the past. It is well known that 'correlation is not causation' and this holds of course for processes, too. Nevertheless, the foregoing definitions and results may be used to formalize and clarify causal reasoning based on marked point processes. In this section, we propose an approach to causality focusing on the distinction between conditioning by *intervention* and conditioning by *observation.* The aim is to find conditions for the identifyability of the effect of an intervention from observational data. A 'complete' treatment of

the subject, however, is not possible due to the different facets of causality. Thus, the following has to be regarded as tentative and is restricted to an outline of the basic ideas. It is based on the concepts developed by Pearl (1993, 1995) who studied causal reasoning in connection with conditional independence graphs, where some of the results have previously been found by Robins (1986) for the special case of longitudinal studies. A concise and unifying overview is given by Lauritzen (2000) and for a more detailed treatment we refer to Pearl (2000). Other approaches to causal reasoning for the special situation of marked point process have been proposed by Eerola (1994), Arjas and Eerola (1992), Parner (1999), and Parner and Arjas (1999) and are briefly addressed in the last subsection.

4.5.1 Causal graphs

In order to distinguish between 'causation and correlation' we consider two different ways of conditioning on the history of a (sub)process: The ordinary conditioning by *observation* and the conditioning by *action* or *intervention*. The latter means that the conditioning value results from an intervention coming from 'outside' of the observable system (Pearl, 1993; Pearl, 2000, pp. 22; Pearl and Robins, 1995; Lauritzen, 2000). As pointed out by Dawid (2000), all assumptions concerning the intervention should be made explicit and enter the graphical representation as far as they affect the conditional or local independence structure. Conditioning by observation of a specific history H_t is further on denoted by $f(\cdot|H_t)$, where f is some probabilistic function such as e.g. a probability distribution or an intensity process. Conditioning by intervention is symbolized by $f(\cdot|H_t^{V\setminus A}||H_t^A)$, $A \subset V = \{1,\ldots,K\}$, where H_t^A is the subhistory which is set to a specific value. One could, for instance, think of $\mathbf{N}_A$ being a counting process indicating when certain drugs are taken. Setting H_t^A to a specific value then means that instead of leaving the medication to the choice of the physician or the patient, a fixed treatment is prescribed.

The preceding outline is of course no precise description of what is meant by an intervention and we restrain from going into details, here, since an appropriate discussion heavily depends on the actual data situation. In the following, all that is required for a local independence graph to reflect the causal structure of a system of processes, is the condition formulated in next definition.

Definition 4.5.1 *Causal local independence graph*
Consider a marked point process on $(\Omega, \mathcal{F}, P)$ with local independence graph $G = (V, E)$. The graph $G = (V, E)$ is called *causal local independence graph* for the distribution P with respect to a subset $A \subset V$, if the likelihood $L(t|H_t^{V\setminus A}||\hat{H}_t^A)$ induced by setting the pre–t history of the subprocess $\mathbf{N}_A$ to $\hat{H}_t^A$ is given by

$$L(t|H_t^{V\setminus A}||\hat{H}_t^A) = \begin{cases} \prod_{k\in V\setminus A} L_k(t|H_t^{\mathrm{cl}(k)}) & \text{on } H_t^A = \hat{H}_t^A, \\ 0, & \text{otherwise,} \end{cases} \qquad t \in \mathcal{T}. \tag{4.11}$$

//

The modification of the likelihood according to (4.11) corresponds to deleting the contribution of H_t^A to the original likelihood and setting it to $\hat{H}_t^A$ where it is relevant for the remaining components, i.e. for all mark specific factors L_k, $k \in V\setminus A$ with $\mathrm{pa}(k) \cap A \neq \emptyset$. This principle is known as *intervention formula* and appears for instance in Pearl (1993) and Spirtes et al. (1993). From (4.11) it can be seen that the local independence graph for the remaining 'free' components $V\setminus A$ is simply given as the subgraph $G_{V\setminus A}$ without the 'marrying parents effect'. Thus, the causal assumption is that the conditional specifications remain the same for all processes which are not used for intervention. This implies that the compensators Λ_k for $k \in V\setminus A$ also stay the same regardless of whether $\hat{H}_t^A$ has been *observed* or *fixed* at its value.
Further, it follows with Corollary 4.2.4 that given the intervention in A we still have for all disjoint subsets $B, C, D \subset V\setminus A$

$$\mathcal{F}_t^B \perp\!\!\!\perp \mathcal{F}_t^C \mid \mathcal{F}_t^D \quad \forall\, t \in \mathcal{T},$$

whenever D separates B and C in the moralized subgraph $(G_{\mathrm{An}(B\cup C\cup D)\setminus A})^m$. When $\mathrm{pa}(A) = \emptyset$, the same conditional independence statements as given above hold for $B, C, D \subset V\setminus A$ when observing $\hat{H}_t^A$ instead of intervening. This can be seen by noting that moralizing the original local independence graph to $(G_{\mathrm{An}(B\cup C\cup D)})^m$ and considering then the subgraph $(G_{\mathrm{An}(B\cup C\cup D)})^m_{\mathrm{An}(B\cup C\cup D)\setminus A}$ yields the same as $(G_{\mathrm{An}(B\cup C\cup D)\setminus A})^m$ if $\mathrm{pa}(A) = \emptyset$. An intervention may therefore also be interpreted as changing the distribution of the whole processes by deleting the directed edges pointing towards the set used for the intervention. In addition, it follows from the foregoing statement that if $\mathrm{pa}(A) = \emptyset$, the effect of an intervention in A is the same

as when conditioning on A. This can also be seen by noting that another way of expressing (4.11) is

$$L(t|H_t^{V\setminus A}||\hat{H}_t^A) = \left.\frac{L(t|H_t)}{\prod_{k\in A} L_k(t|H_t^{\mathrm{cl}(k)})}\right|_{H_t^A=\hat{H}_t^A}. \tag{4.12}$$

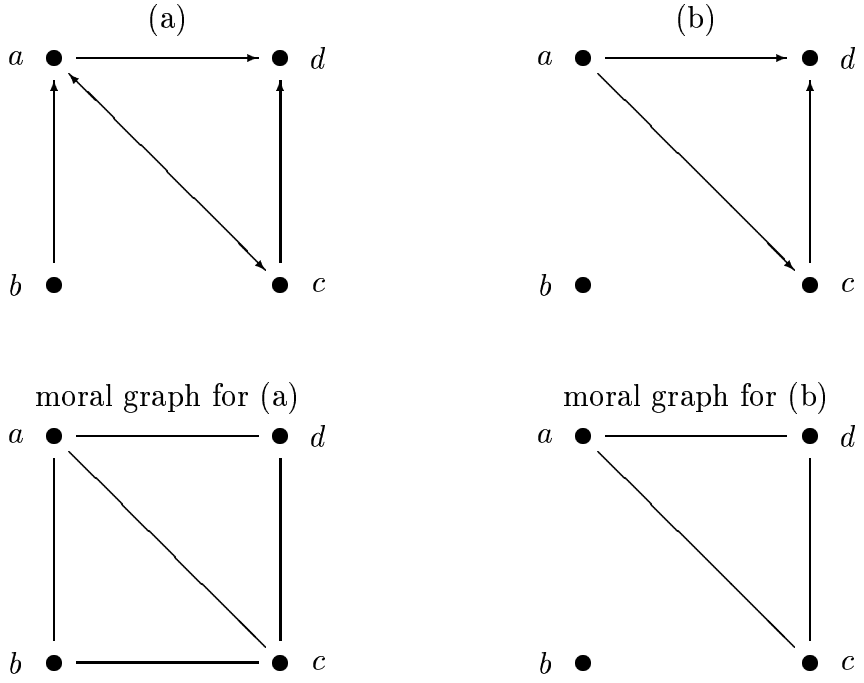
For $\mathrm{pa}(A) = \emptyset$, the denominator is identical to the marginal likelihood of $\hat{H}_t^A$.

Example: Consider the example of a marked point process with mark space $\mathcal{E} = \{a, b, c, d\}$ and local independence graph $G = (V, E)$ with $V = \mathcal{E}$ and $E = \{(a, c), (a, d), (b, a), (c, a), (c, d)\}$ as given in Figure 4.8 (a). Then, we get from (4.11) that when setting H_t^a to a specific history $\hat{H}_t^a$ the likelihood of the remaining components reads as

$$L(t|H^{bcd}||\hat{H}_t^a) = L_b(t|H_t^b)L_c(t|\hat{H}_t^a, H_t^c)L_d(t|\hat{H}_t^a, H_t^{cd}).$$

Fixing the specific history $\hat{H}_t^a$ might for example be done by preventing event a to occur or by forcing it to occur at specific points in time.

Figure 4.8: Local independence graph in the example (a) without intervention in a and (b) with intervention in a, and the corresponding moral graphs.



In this example, $\mathrm{pa}(a) \neq \emptyset$, implying on the one hand that for instance $\mathcal{F}_t^b \perp\!\!\!\perp \mathcal{F}_t^d | \mathcal{F}_t^a$ does not hold in general since there is a path from b to d via c in the moral graph G^m. But on the other hand, given an intervention in a this path is prevented because knowing that H_t^a has be *set* to $\hat{H}_t^a$ induces no dependence between H_t^b and H_t^c so that the claimed conditional independence does hold given the intervention as shown in Figure 4.8 (b). //

The next lemma follows from the definition of a causal local independence graph as can be seen from the foregoing argumentation.

Lemma 4.5.2 *Intervention in ancestral set*
In the situation of Definition 4.5.1 consider an intervention in $A \subset V$ with $\mathrm{pa}(A) = \emptyset$, i.e. fix the history of the subprocess $\mathbf{N}_A$ at $\hat{H}^A$. Then, the compensator and the probability distribution regarding the remaining components are given as if $\hat{H}_t^A$ had been observed. This is symbolized as follows:

$$\begin{aligned} \Lambda_k(t||\hat{H}_{t^-}^A) &= \Lambda_k(t)|_{H_{t^-}^A = \hat{H}_{t^-}^A}, \quad k \in V \backslash A, \\ \pi_t(\tilde{B}||\hat{H}_t^A) &= \pi_t(\tilde{B})\Big|_{H_t^A = \hat{H}_t^A}, \quad \tilde{B} \in \mathcal{H}^{V \backslash A}, \end{aligned}$$

where $\pi_t(\tilde{B}) = P(\tilde{B}|\tilde{\mathcal{F}}_t)$ is the conditional probability of B given a possibly reduced history $\tilde{\mathcal{F}}_t \subset \mathcal{F}_t$. //

The foregoing lemma is crucial for causal inference from observational data, where the effect of an intervention has to be estimable from conditional observations, only. This is addressed in the following subsection.

4.5.2 Intervention graphs and identifiability

In general, the main question regarding causal inference is how to calculate a causal effect from observational data and to formulate conditions that ensure the correctness of these calculations. It is of course not obvious that a quantity which is defined as the result of an intervention should be computable from observational data, where no intervention has actually been performed. Nevertheless, recent developments in the field of causal reasoning show that this is possible and can be formalized using conditional independence models and suitably augmented directed acyclic graphs,

so–called *intervention graphs* (Pearl, 1993; Spirtes et al., 1993; Dawid, 2000). These findings suggest the following approach to causal reasoning for local independence graphs: Assume that the local independence graph for a multivariate stochastic process is given and the interest lies in the effect of an intervention in a specific subset of the components. A graphical representation is then given by augmenting the graph such that the intervention process is symbolized by an additional vertex F and the distribution is appropriately modified to include the intervention.

Definition 4.5.3 *Intervention graph*
Let $G = (V, E)$ be the local independence graph for a marked point process. The *intervention graph* for an intervention in $a \in V$ is given as $G' = (V \cup \{F\}, E \cup \{(F, a)\})$, where F is a random variable realized at time 0 and taking either the value *idle,* symbolized by ϕ, or some value $\hat{H}^a \in I\!H^a$ in the set of histories of the component a. The distributional assumption regarding the effect of an intervention is modified as follows:

$$\Lambda_a(dt \mid H_{t^-}^{\mathrm{cl}(a)}, F) = \begin{cases} \Lambda_a(dt \mid H_{t^-}^{\mathrm{cl}(a)}), & F = \phi, \\ 1, & F = \hat{H}^a \text{ and } (t,a) \in \hat{H}_t^a, \\ 0, & F = \hat{H}^a \text{ and } (t,a) \notin \hat{H}_t^a, \end{cases}$$

so that an intervention forces an event to occur at a specific time, recalling the heuristic relation $\Lambda_a(dt \mid H_{t^-}^{\mathrm{cl}(a)}, F) = P(N_a(dt) | H_{t^-}^{\mathrm{cl}(a)}, F)$. //

In the above definition F is constructed such that it has no parents in the graph. It follows with Lemma 4.5.2 that all compensators and probabilities regarding the original components in $V \backslash \{a\}$ remain the same regardless of whether F is considered as observed or as fixed at a specific value.
In particular, we have that a subset $C \subset V$ of covariate processes may be regarded as identifying the effect of an intervention on the component y if it fulfills the following assumption.

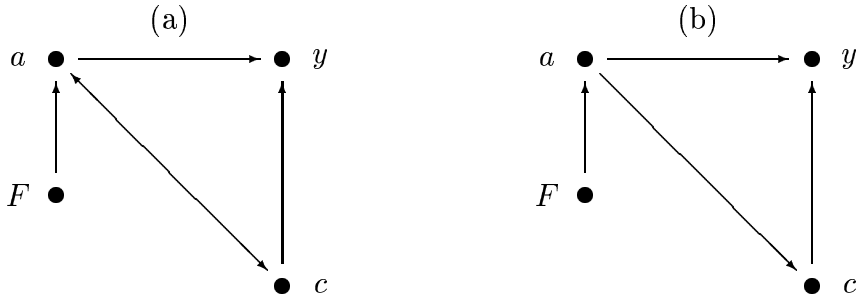
Corollary 4.5.4 *Intervention equals observation*
Consider a local independence graph $G = (V, E)$ of a marked point process with the assumptions of Theorem 4.1.4. Let G' be the intervention graph w.r.t. an intervention in $a \in V$. For $C \subset V \backslash \{a, y\}$ such that $C \cup \{a\}$ δ–separates F from y in G', we have that the evolution of the component y given an intervention $\hat{H}^a$ in a is correctly described by its $\mathcal{F}_t^{C \cup \{a,y\}}$–compensator restricted to $\hat{H}^a$.

Proof:
This follows immeditately from Lemma 4.5.2, Theorem 4.1.4 and the definition of local independence: Intervening in F has the same effect as observing F and, due to the separation, the $\mathcal{F}_t^{C\cup\{a,y,F\}}$–compensator is the same as the $\mathcal{F}_t^{C\cup\{a,y\}}$–compensator. □

It follows from the above corollary that the effect of an intervention $\hat{H}^a$ on any function of the $\mathcal{F}_t^{C\cup\{a,y\}}$–compensator $\tilde{\Lambda}_y(t)$ can be calculated as if $\hat{H}^a$ had been observed. However, some important functions such as prediction w.r.t. $N_y(t+h)$ are no function of the $\mathcal{F}_t^{C\cup\{a,y\}}$–compensator alone, as addressed in the following subsection.

Example: In the graph depicted in Figure 4.9 (a), we have that $\{a, c\}$ δ–separate F from y, whereas this does not hold for a alone. If c is not included in the separating set, it takes the role of an unobserved confounder (cf. next subsection). In situation (b), a alone δ–separates F from y so that c can be discarded. Here, c is a pure mediating process or intermediate on the pathway from a to y. //

Figure 4.9: Intervention graphs for an intervention in the history H_t^a of component N_a of a marked point process.



4.5.3 Open questions

The foregoing results might be taken as basis for deducing further conditions regarding the identifyability of causal effects on specific functions of some outcome variable. For instance, one might be interested in the *instantaneous causal effect* of setting the history of N_a to $\hat{H}_{t-}^a$ on N_y. This could be defined as the compensator

$\tilde{\Lambda}_y(t|H^y_{t^-}||\hat{H}^a_{t^-})$ of $N_y(t)$ given the own past and the intervention but discarding further covariate processes. In contrast to Corollary 4.5.4, confounding processes, like N_c in Figure 4.9 (a), would have to be corrected for because $\tilde{\Lambda}_y(t|H^y_{t^-}||\hat{H}^a_{t^-})$ is no function of the $\mathcal{F}_t^{C\cup\{a,y\}}$–compensator alone. To see this, recall that the compensator based on a reduced history is given as the expectation w.r.t. the distribution of the discarded covariate processes (cf. Theorem 3.1.9). It can be supposed that the correction is only possible if c is δ–separated from F by a suitable subset of the considered processes ensuring that the causal effect on c can be identified.

A more interesting question deals with the identifyability of the predictive causal effect. Let Y be a random variable in a space $(W, \mathcal{V})$ such that Y is also defined on $(\Omega, \mathcal{F}, P)$. The *prediction process* $\{\mu_t\}_{t\geq 0}$ is the conditional probability given the pre–t history of the marked point process (Arjas and Eerola, 1993; Eerola, 1994, p. 34), i.e.

$$\mu_t(\tilde{B}) = P(Y \in \tilde{B} \mid \mathcal{F}_t), \quad \tilde{B} \in \mathcal{V},$$

where $\mathcal{F}_t$ is the history generated by $\mathbf{N}$ as before. Usually, we are interested in some $\tilde{B} \in \mathcal{F}_{t+h}$, $h \geq 0$, since for $\tilde{B} \in \mathcal{F}_t$ the prediction process is either 0 or 1 depending on whether $\tilde{B}$ has occurred or not. It can be shown that there exist regular versions of transition probabilities $\mu_t^*(\cdot|\cdot)$ such that we can use the representation $\mu_t(B) = \mu_t^*(B|H_t)$, where H_t is the history process (Norros, 1985). The *predictive causal effect* of the subprocess $\{\mathbf{N}_A(s)|0 \leq s \leq t\}$, $A \subset V$, on Y could then be defined as the probability of $Y \in B$ given the intervention $F = \hat{H}^A_t$, denoted by $\mu_t^*(B||\hat{H}^A_t)$. Note that Corollary 4.5.4 does not apply as mentioned above. In the situation of confounding processes $\mathbf{N}_C$, the so called *G–formula* developed by Robins (1986, 1989, 1998) might be adapted to the present framework yielding

$$\mu_t^*(B||\hat{H}^A_t) = \int_{\mathbb{H}^C} \mu_t^*(B \mid H^C_t, \hat{H}^A_t)\pi_t(dH^C \mid \hat{H}^A_t), \tag{4.13}$$

implying that the confounding process has to be observed in order to identify the predictive causal effect. Different conditions on the dependence structure of the involved processes can be found in the literature to ensure that the above equality holds. Robins (1986, 1989, 1998) considers mainly the situation of discrete time and formulates the *assumption of no unmeasured confounders* by considering the counterfactual outcome $y_{\hat{H}^A_t}$ that would have been observed, had the treatment been

administered according to $\hat{H}_t^A$. Pearl and Robins (1995) show how this condition can be reformulated so that it may be read off a suitable conditional independence graph. A precise definition of no unmeasured confounders which also applies to the continuous time situations is given by Parner and Arjas (1999) (see also Parner, 1999, p. 60) but without recurrence to the above G–formula or the concept of intervention. With the present notation, it can be seen that (4.13) holds if

$$\mu_t^*(B \mid H_t^C || \hat{H}_t^A) = \mu_t^*(B \mid H_t^C, \hat{H}_t^A)$$

and

$$\pi_t(dH^C || \hat{H}_t^A) = \pi_t(dH^C \mid \hat{H}_t^A).$$

In terms of conditional and local independence, we get that the first condition is ensured by $Y \perp\!\!\!\perp F | (H_t^C, H_t^A)$, where F denotes the intervention variable. The second condition holds if N_C is locally independent of F given N_A. However, the implications of the former condition for the local independence structure are not clear so that there is no obvious way to formulate a suitable graphical condition which could be read off the local independence graph. It would in particular be desirable to have similar criteria as the so–called *back–door* and *front–door criteria* (Pearl, 1995; Pearl, 2000, pp. 78) which can be formulated in a way that they can be verified on the original graph instead of the intervention graph.

Much of the findings regarding causal inference have been motivated by situations of counfounding and mediating variables in epidemiological studies. In the special situation of longitudinal studies, both types of variables may occur so that neither conditioning on the covariates nor ignoring the covariates yield the correct result. Instead, the G–formula should apply. This is illustrated by the following concluding example.

Example: The example is adapted from Robins (1989; see also Lauritzen, 2000). Consider a study made in a population of AIDS patients. The study involves three processes: (1) The treatment process, where the initial treatment with AZT is randomized whereas subsequent treatment depends on the progression of the disease; (2) the counting process indicating whether a patient develops pneumonia which can be regarded as time dependent covariate; and (3) the survival process indicating

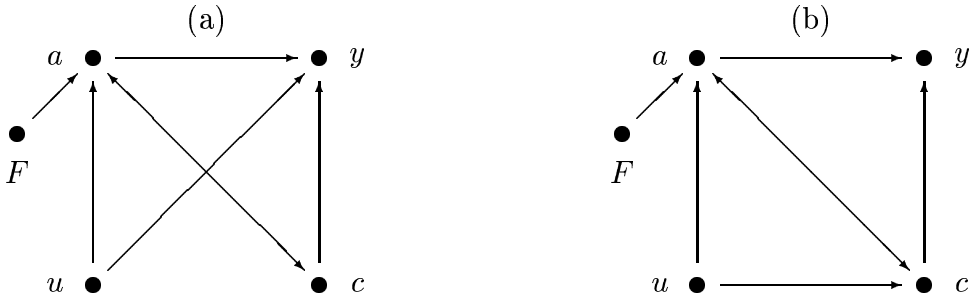
whether a patient survived up to a specific time. With regard to the dependence structure we may assume that the development of pneumonia depends on the treatment process and vice versa since a patient who has developed pneumonia will subsequently receive a treatment with antibiotics whereas this is again randomized for the remaining patients. Both processes are supposed to affect survival. Note that in this situation it makes sense to consider only the stopped process since only the local dependence structure for patients that are alive is of interest. The main causal question in this study is whether the treatment with AZT and / or antibiotics has an effect on the survival. This can also be formulated as counterfactual question: Would the patients who received placebo have survived longer if they had received AZT? The intervention F could therefore be regarded as setting the treatment history to 'the patient receives all treatments in any case' or 'the patient receives no treatment in any case'. The intervention graph is given as in Figure 4.9, where a stands for the treatment process, y for the survival process and c for the development of pneumonia. To formalize the effect of treatment on survival, we could for instance consider the probability that survival lasts at least until τ^* given a specific treatment history $\hat{H}_t^a$ up to $t < \tau^*$, i.e.

$$\mu^*(\tilde{B}||\hat{H}_t^a) = P(N_y(\tau^*) = 0||\hat{H}_t^a),$$

i.e. $\tilde{B} = \{N_y(\tau^*) = 0\}$. The G–formula (4.13) then states that $\mu^*(\tilde{B}||\hat{H}_t^a)$ is given as the expectation of $\mu^*(\tilde{B}|H_t^c, \hat{H}_t^a)$ w.r.t. the conditional distribution of H_t^c given $\hat{H}_t^a$. In other words, we have to integrate over all possible histories regarding the development of pneumonia, i.e. no pneumonia during the study and development of pneumonia at $t' \in (0, t]$, weighted with the corresponding conditional probability given the treatment history. The conditions for this to be true read in this situation as follows: First, we have to make sure that $N_y(\tau^*) \perp\!\!\!\perp F|\mathcal{F}_t^c, \mathcal{F}_t^a$. This would for instance be violated if there is an unobserved process affecting the treatment as well as the survival process (cf. Figure 4.10 (a)). Such a process could be some information on the progression of the health status of the patient which is used to determine the treatment but which is not contained in the process indicating pneumonia. Note that the condition also implies that the intervention only fixes the pre–t history and not H_{t+h}^a, $h > 0$, since this would affect the prediction of $N_y(\tau^*)$ and would prevent the estimability from observed data. The definition of an intervention graph should

thus slightly be modified, but we desist from going into details, here.

Figure 4.10: Intervention graphs with an unobserved process u which inhibits the identifyability of the causal effect.



Further, it has to be ensured that the covariate process indicating pneumonia is locally independent of the intervention F given the treatment process (and that the patient is still alive, of course). This could be violated if some undocumented information on the health status of the patient related to his susceptibility to pneumonia is used to determine the treatment process (cf. Figure 4.10 (b)). Note that both conditions are indeed fulfilled if the treatment is randomized as described above. //

Chapter 5

Statistical inference for local independence graphs

In this final chapter, we consider inference based on local independence graphs for event history data. In the first section we discuss the implication of the factorization of the likelihood shown in the foregoing chapter. This aims at simplifications for statistical inference procedures similar to those known for conditional independence graphs.

In the subsequent sections we consider estimation under local independence restrictions and tests for local independence exploiting the factorization results. The latter may be taken as basis for selection procedures. It is shown that it is not necessary to invent new methods but that inference can rely on the well–known non– and semi-parametric as well as likelihood based estimation and testing procedures. Thus, in Sections 5.2 and 5.3 we mainly restate the corresponding standard results from counting process theory applied to the framework of local independence graphs. However, since local independence graphs apply to a very general model class we restrict our considerations to the special situation of multiplicative intensity models which include composable Markov processes. Further, we subdivide the considered models into those where the individuals constitute homogeneous groups such that inference can be based on aggregated counting processes, and those where for instance the time of the occurrence of previous events has to be taken into account as can be achieved by suitable regression models.

5.1 Preliminaries

Throughout this chapter we consider a marked point process $\{(T_s, E_s)|s = 1, \ldots, S\}$ with mark space $\mathcal{E} = \{e_1, \ldots, e_K\}$ as described in Definition 3.1.1. The local independence graph $G = (V, E)$ with $V = \{1, \ldots, K\}$ represents the local independence structure of the associated multivariate counting process $\mathbf{N} = (N_1, \ldots, N_K)$, where $N_k(t) = N(t; e_k)$ is the mark specific counting process, $k \in V$. Further, we consider filtrations $\{\mathcal{F}_t^A\}_{t\geq 0}$ for subsets $A \subset V$ with $\mathcal{F}_t^A = \sigma\{N_a(s)|a \in A, 0 \leq s \leq t\}$.

As shown in Section 4.2 the likelihood factorizes as

$$L(t|H_t) = \prod_{k\in V} L_k(t|H_t^{\mathrm{cl}(k)}), \tag{5.1}$$

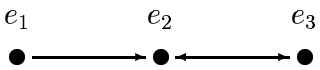
where

$$L_k(t|H_t^{\mathrm{cl}(k)}) = \prod_{T_{s(k)}\leq t} \Lambda_k(dT_{s(k)}) \cdot \prod_{s\leq t}(1 - \Delta\Lambda_k(s))^{1-\Delta N_k(s)} \cdot \exp(-\Lambda_k^C(t))$$

and $\{T_{s(k)}|s(k) \in \{1, \ldots, S\}, E_{s(k)} = e_k\}$ are the occurrence times of event e_k. Whether the above factors of the likelihood should or may be treated separately depends on the actual model specification and parametrization. As pointed out by Arjas (1989) this is a question of the information on the parameters of interest contained in an innovation. Applied to the present situation this means that given a parametrization $\Lambda(t|\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)^\top$, any observed marked point (t, e_k) should be a non–innovation for $\Lambda_j(t|\boldsymbol{\theta})$, $j \neq k$, i.e. $\Lambda_k(t|\boldsymbol{\theta})$ should be independent of those parts of $\boldsymbol{\theta}$ which Λ_j depends on. This implies that $\boldsymbol{\theta}$ can be partitioned into K subvectors $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$, such that we may write $\Lambda_k(t|\boldsymbol{\theta}) = \Lambda_k(t|\boldsymbol{\theta}_k)$. The underlying principle has been termed *partial model specification* since it has been developed for situations where it is not possible to specify the intensity process for every type of event or marked point. In the context of local independence graphs it is, however, necessary to specify if not the whole system then at least those subprocesses for which the local independence structure is not known a priori by subject matter knowledge. Further, for specific models it may still be appropriate to treat the factors of (5.1) separately even if the parameters are not distinct. This has then to be justified using the results for partial likelihoods (Cox, 1975; Gill, 1984).

Example: To illustrate the factorization of the likelihood let us consider an example similar to the one introduced in Section 3.2.2 with $\mathcal{E} = \{e_1, e_2, e_3\}$ and local independence graph $G = (V, E)$ where $V = \{1, 2, 3\}$ and $E = \{(1, 2), (2, 3), (3, 2)\}$ as depicted in Figure 5.1.

Figure 5.1: Assumed local independence graph in the example.



Assume that the events can occur in any order but each only once with absolutely continuous compensators. Let T_1, T_2, and T_3 be the random times of their occurrences. Expressing the intensity processes via appropriate hazard rates $\alpha(t)$ we get that

$$\lambda_1(t) = \mathbf{1}\{t \leq T_1\}\alpha_{1|\cdot}(t)$$

$$\lambda_2(t) = \begin{cases} 0, & t > T_2 \\ \alpha_{2|0}(t), & t \leq \min(T_1, T_2, T_3) \\ \alpha_{2|1}(t|T_1), & T_1 < t \leq \min(T_2, T_3) \\ \alpha_{2|3}(t|T_3), & T_3 < t \leq \min(T_1, T_2) \\ \alpha_{2|13}(t|T_1, T_3), & \max(T_1, T_3) < t \leq T_2, \end{cases}$$

$$\lambda_3(t) = \begin{cases} 0, & t > T_3 \\ \alpha_{3|0}(t), & t \leq \min(T_2, T_3) \\ \alpha_{3|2}(t|T_2), & T_2 < t \leq T_3, \end{cases}$$

where $\alpha_{1|\cdot}$ is the hazard for the occurrence of e_1 regardless of past events, whereas $\alpha_{2|0}$ denotes the hazard for the occurrence of e_2 given that no other event has previously occurred. Further, $\alpha_{2|1}(t|T_1)$ is the hazard for e_2 after the occurrence of e_1 at an earlier point T_1 in time but before e_3 and the other hazard rates are defined analogously. The (partial) likelihoods are given as follows. For event e_1 we have that $L_1(t|H_t) = L_1(t|H_t^1)$ since $\mathrm{cl}(1) = \{1\}$, i.e. regardless of whether one of the other events occurred previously it holds that

$$L_1(t|H_t^1) = \alpha_{1|\cdot}(T_1)^{N_1(t)} \exp\left(-\int_0^{T_1 \wedge t} \alpha_{1|\cdot}(s)\, ds\right).$$

In contrast, $L_2(t|H_t)$ depends on the past of the whole system since $\mathrm{cl}(2) = \{1,2,3\}$ and we have for instance if $T_1 < T_3 < T_2$ (and $t > T_3$) that

$$\begin{aligned} L_2(t|H_t) &= \alpha_{2|13}(T_2)^{N_2(t)} \cdot \\ &\quad \exp\left(-\int_0^{T_1} \alpha_{2|0}(s)\,ds - \int_{T_1}^{T_3} \alpha_{2|1}(s|T_1)\,ds - \int_{T_3}^{T_2\wedge t} \alpha_{2|13}(s|T_1,T_3)\,ds\right). \end{aligned}$$

Finally, with $\mathrm{cl}(3) = \{2,3\}$ we get that $L_3(t|H_t) = L_3(t|H_t^{\{2,3\}})$ is independent of the past as far as the occurrence of e_1 is concerned. If for instance $T_1 < T_2 < T_3$ (and $t > T_2$) we have

$$L_3(t|H_t^{\{2,3\}}) = \alpha_{3|2}(T_3)^{N_3(t)} \exp\left(-\int_0^{T_2} \alpha_{3|0}(s)\,ds - \int_{T_2}^{T_3\wedge t} \alpha_{3|2}(s|T_2)\,ds\right).$$

As addressed in Section 3.2.2, we can alternatively consider the multi–state process $\mathbf{Y}(t)$ with state space $\mathcal{S} = \{0,1\}^3$ indicating whether an event has occurred before t or not. With the above notation the intensity for instance of $\tilde{N}_2(t;((1,0,0),(1,1,0)))$ is given as

$$\tilde{\lambda}_2(t;((1,0,0),(1,1,0))) = \alpha_{2|1}(t|T_1)\mathbf{1}\{T_1 < t \leq \min(T_2,T_3)\}.$$

The likelihood based on the multi–state formulation with $\tilde{N}_k(t;(\mathbf{y},\mathbf{y}'))$ is of course the same. Note that no conditional independence statement for the involved filtrations can be established, i.e. Corollary 4.2.4 does not apply, here, since conditioning on $\mathcal{F}_t^2$ induces a dependence between $\mathcal{F}_t^1$ and $\mathcal{F}_t^3$. In contrast, it would apply if for instance $E = \{(2,1),(2,3),(3,2)\}$. //

The foregoing example is picked up again in the following sections in order to illustrate the different models and methods. Most of the time we assume that all three event times T_1, T_2, and T_3 are observable, i.e. that there is no censoring. However, the presented methods in general remain valid under independent censoring and left truncation (Andersen et al., 1993, pp. 135).

When considering the estimating and testing problem for local independence graphs it is most natural to specify the statistical model in terms of regression models due to the possible dependence on the internal history of the marked point process (and possibly on additional fixed covariates). Consider for instance the above example and

assume that n independent individuals have been observed where $N^i = (N_1^i, N_2^i, N_3^i)$, $i = 1, \ldots, n$, are the individual counting processes, each with the local independence graph as given in Figure 5.1. Then, regression models have to be specified w.r.t. $\alpha_{2|1}(t|t_1)$, $\alpha_{2|3}(t|t_3)$, $\alpha_{2|13}(t|t_1, t_3)$, and $\alpha_{3|2}(t|t_2)$ such that for instance if $T_1^i < \min(T_2^i, T_3^i)$ we have $\lambda_2^i(t) = \alpha_{2|1}(t|T_1^i)\mathbf{1}\{T_1^i < t \leq \min(T_2^i, T_3^i)\}$ etc. Thus, the individuals cannot be regarded as forming a homogeneous group. In contrast, assuming that the hazard rates do not depend on the time of the occurrences of previous events, i.e. $\alpha_{2|1}(t|t_1) = \alpha_{2|1}(t)$ etc., we have for instance

$$\lambda_2^i(t) = \alpha_{2|1}(t)\mathbf{1}\{T_1^i < t \leq \min(T_2^i, T_3^i)\} \quad \text{if} \quad T_1^i < \min(T_2^i, T_3^i),$$

which is a special case of the multiplicative intensity model (Aalen, 1978) since the first factor is deterministic and does not depend on the individual. This allows to aggregate the individual counting process simplifying the statistical model and inference. We therefore treat the latter case and the situation where regression models are required separately in the following sections.

5.2 Models for aggregated counting processes

In this subsection, we consider the situation of n independent individual multivariate counting processes $\mathbf{N}^i = (N_1^i, \ldots, N_K^i)$, $i = 1, \ldots, n$, assuming that the individual intensities satisfy

$$\lambda_k^i(t|\boldsymbol{\theta}) = \alpha_k(t|\boldsymbol{\theta})Z_k^i(t), \quad k = 1, \ldots, K;\ i = 1, \ldots, n, \tag{5.2}$$

where $\alpha_k(t|\boldsymbol{\theta})$ is deterministic, independent of i, and possibly depending on a finite dimensional unknown parameter vector $\boldsymbol{\theta}$, and $Z_k^i(t)$ is a predictable process independent of $\boldsymbol{\theta}$. The model (5.2) is a special case of the multiplicative intensity model where α is allowed to depend on i. In most situations α_k is a hazard rate, relative hazard, or type k transition intensity, and $Z_k^i(t)$ is an indicator for the i-th individual to be at risk for event e_k. Since $\alpha_k(t|\boldsymbol{\theta})$ is deterministic the dependence on the history is fully determined by $Z_k^i(t)$. Examples for models with multiplicative intensity are for instance finite state Markov processes with constant or piecewise constant transition intensities or with transition intensities whose dependence on the time parameter t is governed by $\boldsymbol{\theta}$. The multiplicative intensity model is retained

under independent censoring and left truncation so that the following results also hold for such censored data. Note that $\alpha_k(t|\boldsymbol{\theta})$ as well as $Z_k^i(t)$ may be vector valued.

Example (continued): In the foregoing example we have a multiplicative intensity model if the different hazard rates do not depend on the time of a previous event. Then, the intensity process $\lambda_2(t|\boldsymbol{\theta})$, for instance, is given as

$$\lambda_2(t|\boldsymbol{\theta}) = (\alpha_{2|0}(t|\boldsymbol{\theta}), \alpha_{2|1}(t|\boldsymbol{\theta}), \alpha_{2|3}(t|\boldsymbol{\theta}), \alpha_{2|13}(t|\boldsymbol{\theta})) \begin{pmatrix} \mathbf{1}\{t \leq \min(T_1^i, T_2^i, T_3^i)\} \\ \mathbf{1}\{T_1^i < t \leq \min(T_2^i, T_3^i)\} \\ \mathbf{1}\{T_3^i < t \leq \min(T_1^i, T_2^i)\} \\ \mathbf{1}\{\max(T_1^i, T_3^i) < t \leq T_2^i\} \end{pmatrix}.$$

Obviously it is easier to consider transition specific intensity processes $\lambda_{2|0}$, $\lambda_{2|1}$ etc. corresponding to the different states of the multivariate counting process. As, for instance,

$$\lambda_{2|1} = \alpha_{2|1}(t|\boldsymbol{\theta})\mathbf{1}\{T_1^i < t \leq \min(T_2^i, T_3^i)\}$$

the multiplicative intensity model then still holds. //

Under the assumption (5.2) we have that the likelihood based on $\mathbf{N}^i = (N_1^i, \ldots, N_K^i)$, $i = 1, \ldots, n$, is proportional to the one based on $\mathbf{N} = (N_1, \ldots, N_K)$ with $N_k = \sum_i N_k^i$ which has multiplicative intensities $\alpha_k(t|\boldsymbol{\theta})Z_k(t)$, where $Z_k(t) = \sum_i Z_k^i(t)$ is in most situations given as the number of individuals in the risk set (Andersen et al., 1993, p. 177). Further, we have that the following implication obviously holds.

Lemma 5.2.1 *Individual local independence graph*
Let $G = (V, E)$ be a directed graph. Under the model assumption (5.2) we have that G is the local independence graph of the individual counting process $\mathbf{N}^i = (N_1^i, \ldots, N_K^i)$ if and only if it is the local independence graph of the aggregated counting process $\mathbf{N} = (N_1, \ldots, N_K)$ with $N_k = \sum_{i=1}^n N_k^i$. //

Based on the presented framework of a multiplicative intensity model for aggregated counting processes we now consider maximum likelihood estimation and nonparametric inference under the specific restrictions given by local independence graphs. Since the latter are typically not known a priori we also describe for both frameworks statistical tests of pairwise local independence which might build the basis for selection procedures.

5.2.1 Maximum likelihood estimation

Assume that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)^\top \in \Theta$ is a Q–dimensional parameter in an open subset of the Q–dimensional Euclidean space Θ. Under regularity conditions (cf. Andersen et al., 1993, p. 420) the ML estimator $\hat{\boldsymbol{\theta}}$ resulting from maximizing the likelihood (4.8) has similar properties as in the case of i.i.d. random variables, i.e. there exists with probability tending to one a consistent solution of the estimating equation which is asymptotically normally distributed around the true parameter. This even holds when using a partial likelihood as described above.

Due to the factorization (4.8) and assuming (5.2) the vector $\mathbf{U}(\tau|\boldsymbol{\theta})$ of score statistics has components

$$\begin{aligned} U_q(\tau|\boldsymbol{\theta}) &= \sum_{k=1}^{K} \left[\sum_{T_{s(k)} \le \tau} \frac{\partial}{\partial \theta_q} \log \left(\alpha_k(T_{s(k)}|\boldsymbol{\theta}) Z_k(T_{s(k)}|H_t^{\mathrm{cl}(k)}) \right) \right. \\ &\quad \left. - \int_0^\tau \frac{\partial}{\partial \theta_q} (\alpha_k(t|\boldsymbol{\theta}) Z_k(t|H_t^{\mathrm{cl}(k)}))\, dt \right], \end{aligned}$$

$q = 1, \ldots, Q$, where τ is the upper bound of the observed time interval $\mathcal{T} = [0, \tau)$. For the special case of distinct parameters for each mark specific intensity we therefore get the following result.

Corollary 5.2.2 *ML estimator*
Consider a marked point process with $\mathcal{E} = \{e_1, \ldots, e_K\}$ and the corresponding counting process $\mathbf{N} = (N_1, \ldots, N_K)$ which is based on aggregated independent individual counting processes $\mathbf{N}^i = (N_1^i, \ldots, N_K^i)$, $i = 1, \ldots, n$. Assume that the multiplicative intensity model (5.2) holds and that every $\mathbf{N}^i$, $i = 1, \ldots, n$, has the local independence graph G. Further, assume that $\boldsymbol{\theta}$ can be partitioned into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_K^\top)^\top$ such that the intensity $\lambda_k(t|\boldsymbol{\theta})$ only depends on $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kJ_k})^\top$, i.e. we may write $\alpha_k(t|\boldsymbol{\theta}) = \alpha_k(t|\boldsymbol{\theta}_k)$. Then, the ML estimator $\hat{\boldsymbol{\theta}}$ is given as solution of the equations

$$\sum_{T_{s(k)} \le \tau} \frac{\partial}{\partial \theta_{kj}} \log \left(\alpha_k(T_{s(k)}|\boldsymbol{\theta}_k) Z_k(T_{s(k)}|H_t^{\mathrm{cl}(k)}) \right) = \int_0^\tau \frac{\partial}{\partial \theta_{kj}} (\alpha_k(t|\boldsymbol{\theta}_k) Z_k(t|H_t^{\mathrm{cl}(k)}))\, dt,$$

$j = 1, \ldots, J_k$, $k = 1, \ldots, K$. Under the regularity conditions given in Andersen et al. (1993, p. 420) the above equation has a solution $\hat{\boldsymbol{\theta}}$ with probability tending to

one and $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}^0$ as $n \to \infty$, where $\boldsymbol{\theta}^0$ is the true parameter value. Further, the components $\hat{\boldsymbol{\theta}}_k$, $k = 1, \ldots, K$, are asymptotically independent with the following asymptotic distributions:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^0) \xrightarrow{D} \mathcal{N}(0, \Sigma_k).$$

The inverse asymptotic covariance matrix Σ_k^{-1} may be estimated consistently by $\mathcal{I}_k(\tau|\hat{\boldsymbol{\theta}})/n$ where the information matrix $\mathcal{I}_k(\tau|\boldsymbol{\theta})$ has the entries

$$\int_0^\tau \frac{\partial^2}{\partial\theta_{ki}\partial\theta_{kj}} \left(\alpha_k(t|\boldsymbol{\theta}_k) Z_k(t|H_t^{\mathrm{cl}(k)})\right) dt - \sum_{T_{s(k)} \leq \tau} \frac{\partial^2}{\partial\theta_{ki}\partial\theta_{kj}} \log\left(\alpha_k(T_{s(k)}|\boldsymbol{\theta}_k) Z_k(T_{s(k)}|H_t^{\mathrm{cl}(k)})\right)$$

for $i, j = 1, \ldots, J_k$, $k = 1, \ldots, K$. //

The above result differs from the general result only in that it makes explicit on which part of the internal history the components of the ML estimator and its asymptotic distribution depend.

Based on the asymptotic distribution of the (partial) ML estimator which can be used to derive a Wald or a score test, we can also formulate a likelihood ratio test. Let $\hat{\boldsymbol{\theta}}^0$ be the ML estimator under the null hypothesis with P dimensions and $\hat{\boldsymbol{\theta}}$ the estimator under a larger model with Q dimensions, $Q > P$, then

$$2(C_\tau(\hat{\boldsymbol{\theta}}) - C_\tau(\hat{\boldsymbol{\theta}}^0)) \sim \chi^2_{Q-P},$$

where $C_\tau(\boldsymbol{\theta})$ is the loglikelihood. This obviously becomes much simpler when the null hypothesis only restricts the parameter vector $\boldsymbol{\theta}_k$ governing the intensity for an event e_k. This would for instance be the case if a local independence of e_k from some e_j, $j \neq k$, is tested. In the next section, we consider this special testing problem in more detail for composable Markov processes.

5.2.2 LR tests for composable Markov Processes

We now apply the above results to the special case of composable finite state Markov processes (CFMP) as defined in Section 3.2.3. However, here we only treat the situation of constant transition intensities. Consider n individual CFMPs $\mathbf{Y}^i \sim (Y_1^i, \ldots, Y_K^i)$ with local independence graph $G = (V, E)$ and constant transition

intensities $\alpha_k(\mathbf{y}_{\mathrm{cl}(k)}, y'_k)$, $\mathbf{y}_{\mathrm{cl}(k)} \in \mathcal{S}_{\mathrm{cl}(k)}$, $y'_k \in \mathcal{S}_k$, $y_k \neq y'_k$, $k = 1, \ldots, K$. Thus, the parameter vector $\boldsymbol{\theta}$ of this model is given as these constant transition intensities and the partition into $\boldsymbol{\theta}_k$ by all $\alpha_k(\cdot, \cdot)$. The intensity process for an individual transition specific counting process $N^i(t; (\mathbf{y}_{\mathrm{cl}(k)}, y'_k))$ is given as

$$\lambda^i(t; (\mathbf{y}_{\mathrm{cl}(k)}, y'_k)) = \alpha_k(\mathbf{y}_{\mathrm{cl}(k)}, y'_k)\mathbf{1}\{\mathbf{Y}^i_{\mathrm{cl}(k)}(t^-) = \mathbf{y}_{\mathrm{cl}(k)}\}$$

and thus constitutes a multiplicative intensity model as demanded by (5.2). The aggregated counting processes $N(t; (\mathbf{y}_{\mathrm{cl}(k)}, y'_k)) = \sum_{i=1}^n N^i(t; (\mathbf{y}_{\mathrm{cl}(k)}, y'_k))$ therefore have the intensity processes $\alpha_k(\mathbf{y}_{\mathrm{cl}(k)}, y'_k) Z(t; \mathbf{y}_{\mathrm{cl}(k)})$, where

$$Z(t; \mathbf{y}_{\mathrm{cl}(k)}) = \sum_{i=1}^{n} \mathbf{1}\{\mathbf{Y}^i_{\mathrm{cl}(k)}(t^-) = \mathbf{y}_{\mathrm{cl}(k)}\}$$

is the size of the risk set.

ML estimator

According to the above score equations the ML estimator of the transition intensities is now given as

$$\hat{\alpha}_k(\mathbf{y}_{\mathrm{cl}(k)}, y'_k) = \frac{N_k(\tau; (\mathbf{y}_{\mathrm{cl}(k)}, y'_k))}{\int_0^\tau Z(t; \mathbf{y}_{\mathrm{cl}(k)})\, dt},$$

where the nominator is the size of the risk set at time t integrated over the observed time interval, i.e. it can be regarded as the 'exposure time' or 'total time on test'.

LR test on local independence

When aiming at inference about the local independence structure the basic test is the one on pairwise local independence. This might for instance be used within a selection procedure so that we now assume that nothing is known about the local independence structure, i.e. we start with a complete graph where $\mathrm{cl}(k) = V$ for all $k \in V$. The corresponding testing problem is then given as $H_0 : \{j\} \not\to \{k\} | V \backslash \{j, k\}$ vs. $H_1 : \{j\} \longrightarrow \{k\} | V \backslash \{j, k\}$. In the present situation, the null hypothesis is equivalent to requiring that for all $\mathbf{y}, \tilde{\mathbf{y}} \in \mathcal{S}$ where $\mathbf{y}$ and $\tilde{\mathbf{y}}$ only differ in the j-th component and for all $y'_k \neq y_k$ it holds that

$$\alpha_k(\mathbf{y}, y'_k) = \alpha_k(\tilde{\mathbf{y}}, y'_k).$$

Thus, the components of the estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^0$ are given as

$$\hat{\alpha}_k(\mathbf{y}, y'_k) = \frac{N_k(\tau; (\mathbf{y}, y'_k))}{\int_0^\tau Z(t; \mathbf{y})\, dt} \quad \text{and} \quad \hat{\alpha}^0_k(\mathbf{y}, y'_k) = \frac{N_k(\tau; (\mathbf{y}_{-j}, y'_k))}{\int_0^\tau Z(t; \mathbf{y}_{-j})\, dt},$$

where $\mathbf{y}_{-j} = \mathbf{y}_{V\setminus\{j\}}$. If there are no further model restrictions on the transitions such as absorbing or transient states or other local independencies then the number of parameters under H_0 is $|\mathcal{S}_j| \cdot |\mathcal{S}_{-j}| \cdot (|\mathcal{S}_k| - 1)$ less than in the full model. Thus, the test statistic given as $2(C^k_\tau(\hat{\boldsymbol{\theta}}_k) - C^k_\tau(\hat{\boldsymbol{\theta}}_k{}^0))$ is χ^2–distributed with this number of degrees of freedom where

$$C^k_\tau(\hat{\boldsymbol{\theta}}_k) = \sum_{\mathbf{y}\in\mathcal{S}} \sum_{y'_k \neq y_k} \left[\sum_{T_{s(\mathbf{y},y'_k)}} \log\left(\hat{\alpha}_k(\mathbf{y}, y'_k) Z(T_{s(\mathbf{y},y'_k)}; \mathbf{y})\right) - \hat{\alpha}_k(\mathbf{y}, y'_k) \int_0^\tau Z(t; \mathbf{y})\, dt \right].$$

Again, $T_{s(\mathbf{y},y'_k)}$ denotes the observed times of a transition from $\mathbf{y}$ to y'_k and $\boldsymbol{\theta}_k = (\alpha_k(\mathbf{y}, y'_k) | \mathbf{y} \in \mathcal{S},\, y'_k \in \mathcal{S}_k \text{ with } y'_k \neq y_k)$, i.e. the parameters are the constant transition intensities themselves.

In a similar manner we may construct LR tests for testing a graph $G^0 = (V, E^0)$ versus a graph $G^1 = (V, E^1)$, $E^0 \subset E^1$. In this case, the LR test is based on comparing the loglikelihoods for those $k \in V$ where $j \neq k$ exists with $(j, k) \in E^1$ and $(j, k) \notin E^0$. The number of degrees of freedom again depend on possible other model restrictions and on the state spaces $\mathcal{S}_k$, $k \in V$.

5.2.3 Nonparametric inference

For the case that no parametric assumption is made in the multiplicative intensity model (5.2) we now turn to nonparametric methods. Although one is in general interested in the transition intensities or hazards α_k, $k \in V$, we restrict ourselves to the estimation of the cumulative hazards

$$A_k(t) = \int_0^t \alpha_k(s)\, ds, \quad k = 1, \ldots, K;\ t \in \mathcal{T},$$

which are assumed to be finite. A very general nonparametric estimator for $A_k(t)$ is given by the Nelson–Aalen estimator (Aalen, 1978) and corresponding estimators of α_k may for instance be derived by kernel function smoothing (cf. Andersen et al.,

1993, pp. 229). The Nelson–Aalen estimator $\hat{A}_k$ is given as

$$\hat{A}_k(t) = \int_0^t Z_k(s)^{-1} dN_k(s) = \sum_{T_{s(k)} \leq t} Z_k(T_{s(k)})^{-1}, \tag{5.3}$$

i.e. as an increasing right continuous step–function with increments $1/Z_k(T_{s(k)})$ at the jump times $T_{s(k)}$ of N_k. Note that this estimator is already mark specific so that the factorization of the likelihood has no particular effect. Further, it requires that the model is specified such that $Z_k(t)$ is one–dimensional. It can be shown that under appropriate regularity conditions the Nelson–Aalen estimator is uniformly consistent on compact intervals (Andersen et al., 1993, p. 190). With regard to the asymptotic distribution we have for the vector $\hat{\mathbf{A}} = (\hat{A}_1, \ldots, \hat{A}_K)^\top$ that

$$\sqrt{n}(\hat{\mathbf{A}} - \mathbf{A}) \xrightarrow{D} \mathbf{U} = (U_1, \ldots, U_K)^\top,$$

where the processes $U_1, \ldots, U_K$ are independent Gaussian martingales with $U_k(0) = 0$ and $Cov(U_k(s), U_k(t)) = \sigma_k^2(s \wedge t)$. The covariance may be estimated as

$$\hat{\sigma}_k^2(t) = \sum_{T_{s(k)} \leq t} \frac{\mathbf{1}\{Z_k(T_{s(k)}) > 0\}}{Z_k(T_{s(k)})^2}.$$

Another proposition for the estimation of the variance which is especially suited for aggregated counting processes and able to cope with ties (cf. Andersen et al., 1993, p. 181) is given as

$$\tilde{\sigma}_k^2(t) = \sum_{T_{s(k)} \leq t} \frac{\Delta N_k(T_{s(k)})(Z_k(T_{s(k)}) - \Delta N_k(T_{s(k)}))}{Z_k(T_{s(k)})^3}.$$

Applying the above results to a composable finite state Markov process with a given local independence structure we immediately get the following corollary by exploiting the knowledge that $\alpha_k(t; (\mathbf{y}, y_k'))$ does not depend on $\mathbf{y}_{V \setminus \mathrm{cl}(k)}$.

Corollary 5.2.3 *Nelson–Aalen estimator for CFMP*
Consider a sample $\mathbf{Y}^i$, $i = 1, \ldots, n$, of independent composable finite Markov processes with local independence graph G and transition intensities $\alpha_k(t; (\mathbf{y}, y_k'))$, $\mathbf{y} \in \mathcal{S}$, $y_k' \neq y_k$. The Nelson–Aalen estimator of the transition specific cumulative intensities is given as

$$\hat{A}_k(t; (\mathbf{y}, y_k')) = \sum_{T_{s(\mathbf{y}_{\mathrm{cl}(k)}, y_k')} \leq t} Z(T_{s(\mathbf{y}_{\mathrm{cl}(k)}, y_k')}; \mathbf{y}_{\mathrm{cl}(k)})^{-1},$$

where $Z(t; \mathbf{y}_{\mathrm{cl}(k)}) = \sum_{i=1}^{n} \mathbf{1}\{\mathbf{Y}^i_{\mathrm{cl}(k)}(t^-) = \mathbf{y}_{\mathrm{cl}(k)}\}$ and $T_{s(\mathbf{y}_{\mathrm{cl}(k)}, y'_k)}$ are the times of any transition from a state $\tilde{\mathbf{y}}$ with $\tilde{\mathbf{y}}_{\mathrm{cl}(k)} = \mathbf{y}_{\mathrm{cl}(k)}$ to the state $\tilde{\mathbf{y}}'$ where only the k-th component has changed to y'_k. //

This corollary confirms that instead of estimating the $|\mathcal{S}| \cdot \sum_{k=1}^{K}(|\mathcal{S}_k| - 1)$ dimensional vector $(\hat{A}_k(t; (\mathbf{y}, y'_k))|\ \mathbf{y} \in \mathcal{S},\ y'_k \in \mathcal{S}_k,\ y'_k \neq y_k,\ k = 1, \ldots, K)$ the procedure may be reduced to the $\sum_{k=1}^{K} |\mathcal{S}_{\mathrm{cl}(k)}|(|\mathcal{S}_k| - 1)$ dimensional vector $(\hat{A}_k(t; (\mathbf{y}_{\mathrm{cl}(k)}, y'_k))|\ \mathbf{y}_{\mathrm{cl}(k)} \in \mathcal{S}_{\mathrm{cl}(k)},\ y'_k \in \mathcal{S}_k,\ y'_k \neq y_k,\ k = 1, \ldots, K)$.

Example (continued): Consider again the previous example. As stated above, this can be reformulated as multi–state process with separate counting process $\tilde{N}_k(t; (\mathbf{y}, y'_k))$ for each possible transition, with e.g. the transition from $(1, 0, 0)$ to $(1, 1, 0)$ indicating that e_2 occurs after e_1 but before e_3. The corresponding cumulative hazard may be estimated as

$$\hat{A}_{2|1}(t) = \sum_{T_{s(2|1)} \leq t} \frac{1}{Z_{2|1}(T_{s(2|1)})},$$

where $T_{s(2|1)}$ denotes the times of transitions from $(1, 0, 0)$ to $(1, 1, 0)$ and

$$Z_{2|1}(t) = \sum_{i=1}^{n} \mathbf{1}\{T^i_1 < t \leq \min(T^i_2, T^i_3)\}$$

is the number of individuals for whom event e_1 has already occurred before t but neither e_2 nor e_3. In addition, we know by the local independence structure depicted in Figure 5.1 that the estimation of $A_1(t)$, for instance, can discard the observed order in which the three events occur. This yields

$$\hat{A}_{1|\cdot}(t) = \sum_{T_{s(1)} \leq t} \frac{1}{Z_{1|\cdot}(T_{s(1)})},$$

where $T_{s(1)}$ denote the times of occurrences of e_1 regardless of past events and

$$Z_{1|\cdot}(t) = \sum_{i=1}^{n} \mathbf{1}\{t \leq T^i_1\}.$$

In other words, it is not necessary to separately estimate $A_{1|0}(t)$, $A_{1|2}(t)$, $A_{1|3}(t)$, and $A_{1|23}(t)$. //

Confidence bands for $A_k(t)$ may be derived by exploiting the above asymptotic distribution. Since this does not directly contribute to inference about the local independence structure we restrict in the following our attention to the testing problem.

Nonparametric test on local independence

Consider again a CFMP, but the transition intensities are now allowed to depend on time. The hypothesis of pairwise local independence $\{j\} \not\to \{k\}|V\backslash\{j,k\}$ may then be written as follows: For all $\mathbf{y}, \tilde{\mathbf{y}} \in \mathcal{S}$ where $\mathbf{y}$ and $\tilde{\mathbf{y}}$ only differ in the j-th component and for all $y'_k \neq y_k$ it holds that

$$\alpha_k(t;(\mathbf{y},y'_k)) = \alpha_k(t;(\tilde{\mathbf{y}},y'_k)) \quad \forall\, t \in \mathcal{T}, \tag{5.4}$$

i.e. we hypothesize the equality of a number $|\mathcal{S}_j|$ of hazard rates for each $\mathbf{y}_{-j} \in \mathcal{S}_{-j}$ and each $y'_k \neq y_k$. In order to make clear that the hypothesized equality is not between different values of k but between different $y_j \in \mathcal{S}_j$ we index the following quantities instead with y_j and consider $\mathbf{y}_{-j}$ and y'_k as fixed. An appropriate test is then based on comparing the Nelson–Aalen estimators $\hat{A}_{y_j}(t;(\mathbf{y},y'_k))$ for each transition with those for the hypothesized common values $\int \alpha_k(s;(\mathbf{y}_{-j},y'_k))\,ds$ which are given in a slightly modified way by

$$\tilde{A}_{y_j}(t;(\mathbf{y},y'_k)) = \sum_{T_{s(\mathbf{y}_{-j},y'_k)} \leq t} \frac{\mathbf{1}\{Z_{y_j}(T_{s(\mathbf{y}_{-j},y'_k)};\mathbf{y}) > 0\}}{Z(T_{s(\mathbf{y}_{-j},y'_k)};\mathbf{y}_{-j})},$$

where $Z(t;\mathbf{y}_{-j}) = \mathbf{1}\{\mathbf{Y}_{-j}(t^-) = \mathbf{y}_{-j}\} = \sum_{y_j} Z_{y_j}(t;\mathbf{y})$. Under the null hypothesis the above estimator and $\hat{A}_{y_j}(t;(\mathbf{y},y'_k))$ are equal except for random variation, their difference forming a local square integrable martingale. It is thus reasonable to base the test statistic on their difference. Generalizing this idea yields the test statistic

$$D_{y_j}(t;(\mathbf{y},y'_k)) = \int_0^t W_{y_j}(t;(\mathbf{y},y'_k))\, d[\hat{A}_{y_j}(s;(\mathbf{y},y'_k)) - \tilde{A}_{y_j}(s;(\mathbf{y},y'_k))],$$

where W_{y_j} are nonnegative locally bounded predictable weight processes. Under the very general assumption that $W_{y_j}(t;(\mathbf{y},y'_k)) = W(t)Z_{y_j}(t;\mathbf{y})$, which covers most of the practical applications, we have that the test statistic equals

$$\sum_{T_{s(\mathbf{y},y'_k)} \leq t} W(T_{s(\mathbf{y},y'_k)}) - \sum_{T_{s(\mathbf{y}_{-j},y'_k)} \leq t} W(T_{s(\mathbf{y}_{-j},y'_k)}) \frac{Z_{y_j}(T_{s(\mathbf{y}_{-j},y'_k)};\mathbf{y})}{Z(T_{s(\mathbf{y}_{-j},y'_k)};\mathbf{y}_{-j})}.$$

Unbiased estimates $\hat{\sigma}_{y_j,\tilde{y}_j}(t)$ of the variances and covariances of $D_{y_j}(t;(\mathbf{y},y'_k))$, $y_j \in \mathcal{S}_j$, are given by (Andersen et al., 1982)

$$\sum_{T_{s(\mathbf{y}_{-j},y'_k)}\leq t} W^2(T_{s(\mathbf{y}_{-j},y'_k)}) \frac{Z_{y_j}(T_{s(\mathbf{y}_{-j},y'_k)};\mathbf{y})}{Z(T_{s(\mathbf{y}_{-j},y'_k)};\mathbf{y}_{-j})} \left(\delta_{y_j,\tilde{y}_j} - \frac{Z_{\tilde{y}_j}(T_{s(\mathbf{y}_{-j},y'_k)};\tilde{\mathbf{y}})}{Z(T_{s(\mathbf{y}_{-j},y'_k)};\mathbf{y}_{-j})}\right),$$

where $\delta_{y_j,\tilde{y}_j}$ is the Kronecker symbol. The $|\mathcal{S}_j| \times |\mathcal{S}_j|$ matrix with the above elements is denoted by $\hat{\Sigma}(t;(\mathbf{y}_{-j},y'_k))$. Let $\mathbf{D}(t;(\mathbf{y}_{-j},y'_k)) = (D_{y_j}(t;(\mathbf{y},y'_k))|y_j \in \mathcal{S}_j)^\top$. Then, it can be shown (Andersen et al., 1982) that the test statistic

$$\mathbf{D}(t;(\mathbf{y}_{-j},y'_k))^\top \hat{\Sigma}(t;(\mathbf{y}_{-j},y'_k))^- \mathbf{D}(t;(\mathbf{y}_{-j},y'_k))$$

is asymptotically χ^2–distributed with $|\mathcal{S}_j| - 1$ degrees of freedom, where $\hat{\Sigma}^-$ is the generalized inverse. Testing all the equalities in the hypothesis (5.4) corresponds to a stratified test where the strata are given by the combinations of $\mathbf{y}_{-j} \in \mathcal{S}_{-j}$ and $y'_k \in \mathcal{S}_k$ with $y'_k \neq y_k$. The above results can be generalized to this situation in the following way. Let L be the number of strata, i.e. $L = |\{(\mathbf{y}_{-j},y'_k)|\mathbf{y}_{-j} \in \mathcal{S}_{-j}$ and $y'_k \in \mathcal{S}_k$ with $y'_k \neq y_k\}|$ and index the above quantities by $l = 1,\ldots,L$ replacing $(\mathbf{y}_{-j},y'_k)$. Then, we have under the hypothesis of local independence that

$$\left(\sum_{l=1}^{L} \mathbf{D}_l(t)\right)^\top \left(\sum_{l=1}^{L} \hat{\Sigma}_l(t)\right)^- \left(\sum_{l=1}^{L} \mathbf{D}_l(t)\right) \tag{5.5}$$

is again asymptotically χ^2–distributed with $|\mathcal{S}_j| - 1$ degrees of freedom which result from the rank of the covariance matrix. The maximal rank of $\hat{\Sigma}$ is $|\mathcal{S}_j| - 1$ so that in the two–sample case, i.e. if $|\mathcal{S}_j| = 2$, the test may equivalently be based on only one of $D_{y_j}(t;(\mathbf{y}_{-j},y'_k))$, $y_j \in \mathcal{S}_j$. The modified test statistic

$$\left(\sum_{l=1}^{L} D_{y_j;l}(t)\right) \left(\sum_{l=1}^{L} \hat{\sigma}^2_{y_j;l}(t)\right)^{-1/2} \tag{5.6}$$

is asymptotically standard normally distributed under the null hypothesis. The above tests are readily generalized to hypotheses of the kind $H_0 : A \not\to \{k\}|V\backslash A$ which can be reformulated as equality of $|\mathcal{S}_A|$ hazards for each $\mathbf{y}_{V\backslash A}$ and $y'_k \neq y_k$.

Example (continued): Testing in our example whether $\{1\} \not\to \{2\}|\{3\}$ is equivalent to testing

$$H_0 : \ \alpha_{2|1}(t) = \alpha_{2|0}(t) \quad \text{and} \quad \alpha_{2|13}(t) = \alpha_{2|3}(t).$$

The first part of the hypothesis is concerned with the situation where e_2 occurs before e_3 and implies that the intensity for e_2 is then unaffected by the order in which e_1 and e_2 occur. The second part consists of the same statement but for the situation where e_3 occurred before e_2. In the following, we only consider the first part of the hypothesis. Assuming that the three failure times T_1^i, T_2^i, and T_3^i are observable for all individuals we reduce the sample to those individuals where e_2 occurred before e_3, the sample size being now denoted by n^{23}. Choosing $W(t) = Z_{2|\cdot}(t) = \sum_{i=1}^{n} \mathbf{1}\{t \leq T_2^i < T_3^i\}$ the total number of individuals where event e_2 has not occurred before t regardless of whether e_1 has occurred before we get

$$D_{2|1}(t) = \sum_{T_{s(2|1)} \leq t} Z_{2|\cdot}(T_{s(2|1)}) - \sum_{T_{s(2|\cdot)} \leq t} Z_{2|1}(T_{s(2|\cdot)})$$

and

$$D_{2|0}(t) = \sum_{T_{s(2|0)} \leq t} Z_{2|\cdot}(T_{s(2|0)}) - \sum_{T_{s(2|\cdot)} \leq t} Z_{2|0}(T_{s(2|\cdot)}),$$

where $T_{s(2|\cdot)}$ denotes the points in time where event e_2 occurs for any of the individuals again regardless of whether e_1 has occurred before. Note that $Z_{2|\cdot}(T_{s(2|1)})$ is a transformation of the range of $T_{s(2|1)}$ w.r.t. all points in time where an event e_2 occurs without any other event before or with e_1 having occurred before, i.e. $R(T_{s(2|1)}) = n^{23} - Z_{2|\cdot}(T_{s(2|1)}) + 1$ (assuming that there are no censored observations). The resulting test may therefore be regarded as a generalization of the *Wilcoxon* or *Kruskal–Wallis* test to right censored data. The estimated variances and covariances in this situation are given by

$$\begin{aligned}
\hat{\sigma}_{00}(t) &= \sum_{T_{s(2|\cdot)} \leq t} \left(Z_{2|\cdot}(T_{s(2|\cdot)}) Z_{2|0}(T_{s(2|\cdot)}) - Z_{2|0}(T_{s(2|\cdot)})^2 \right), \\
\hat{\sigma}_{11}(t) &= \sum_{T_{s(2|\cdot)} \leq t} \left(Z_{2|\cdot}(T_{s(2|\cdot)}) Z_{2|1}(T_{s(2|\cdot)}) - Z_{2|1}(T_{s(2|\cdot)})^2 \right), \\
\hat{\sigma}_{01}(t) &= \sum_{T_{s(2|\cdot)} \leq t} -Z_{2|0}(T_{s(2|\cdot)}) Z_{2|1}(T_{s(2|\cdot)}).
\end{aligned}$$

Since we are considering a two–sample situation the test may equivalently be based only on the asymptotic standard normal distribution of $D_{2|0}(t)/\sqrt{\hat{\sigma}_{00}(t)}$.
Another test results when choosing $W(t) = \mathbf{1}\{Z_{2|\cdot}(t) > 0\}$. This yields

$$D_{2|1}(t) = N_{2|1}(t) - \sum_{T_{s(2|\cdot)} \leq t} \frac{Z_{2|1}(T_{s(2|\cdot)})}{Z_{2|\cdot}(T_{s(2|\cdot)})}$$

and

$$D_{2|0}(t) = N_{2|0}(t) - \sum_{T_{s(2|\cdot)} \le t} \frac{Z_{2|0}(T_{s(2|\cdot)})}{Z_{2|\cdot}(T_{s(2|\cdot)})},$$

where the last terms may be regarded as the number of events $\{T_1 < T_2 < T_3\}$ and $\{T_2 < T_1 < T_3\}$ expected under the null hypothesis, respectively. The corresponding test is known as *log–rank test* for survival data analysis. The estimated variances and covariances in this case read as

$$\begin{aligned}
\hat{\sigma}_{00}(t) &= \sum_{T_{s(2|\cdot)} \le t} \frac{Z_{2|0}(T_{s(2|\cdot)})}{Z_{2|\cdot}(T_{s(2|\cdot)})} \left(1 - \frac{Z_{2|0}(T_{s(2|\cdot)})}{Z_{2|\cdot}(T_{s(2|\cdot)})}\right), \\
\hat{\sigma}_{11}(t) &= \sum_{T_{s(2|\cdot)} \le t} \frac{Z_{2|1}(T_{s(2|\cdot)})}{Z_{2|\cdot}(T_{s(2|\cdot)})} \left(1 - \frac{Z_{2|1}(T_{s(2|\cdot)})}{Z_{2|\cdot}(T_{s(2|\cdot)})}\right), \\
\hat{\sigma}_{01}(t) &= \sum_{T_{s(2|\cdot)} \le t} -\frac{Z_{2|0}(T_{s(2|\cdot)}) Z_{2|1}(T_{s(2|\cdot)})}{Z_{2|\cdot}(T_{s(2|\cdot)})^2}.
\end{aligned}$$

The analogous quantities for the second part of the hypothesis can be calculated in a similar manner by considering the subsample of individuals where $T_3 < T_2$. Inserting all these quantities for $t = \tau$ in the test statistic (5.5) or (5.6) and comparing this with the corresponding quantiles we get an asymptotic level α test for the above hypothesis. //

Product–limit estimator for composable Markov processes

In the special case of Markov processes one might additionally be interested in estimating the transition probabilities

$$P_{\mathbf{y},\mathbf{y}'}(s,t) = P(\mathbf{Y}(t) = \mathbf{y}' | \mathbf{Y}(s) = \mathbf{y}), \quad s < t;\ \mathbf{y}, \mathbf{y}' \in \mathcal{S}.$$

The product–limit estimator (Aalen and Johannsen, 1978) is based on the above Nelson–Aalen estimator since the matrix $\mathbf{P}(s,t)$ of all transition probabilities can be expressed as function of $A(t;(\mathbf{y},\mathbf{y}'))$, $\mathbf{y},\mathbf{y}' \in \mathcal{S}$, defined above. The matrix version of the product–limit estimator is given as

$$\hat{\mathbf{P}}(s,t) = \prod_{s < T_r \le t} (\mathbf{I} + \Delta \hat{\mathbf{A}}(T_r)), \tag{5.7}$$

where T_r are the observed transition times and $\mathbf{I}$ is the $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix. Thus, the product–limit estimator is a finite product of matrices. If one or more

transitions are observed at time u, then the contribution to (5.7) is a matrix with entry $(\mathbf{y}, \mathbf{y}')$ equal to $\Delta N(u; (\mathbf{y}, \mathbf{y}'))/Z(u; \mathbf{y})$ for $\mathbf{y} \neq \mathbf{y}'$ and entry $(\mathbf{y}, \mathbf{y})$ equal to $1 - \Delta N(u; \mathbf{y})/Z(u; \mathbf{y})$ with $N(u; \mathbf{y}) = \sum_{\mathbf{y}' \neq \mathbf{y}} N(u; (\mathbf{y}, \mathbf{y}'))$. Note that under the model assumption there is only one transition at a time with probability one so that in general $N(u; \mathbf{y})/Z(u; \mathbf{y}) = Z(u; \mathbf{y})^{-1}$ and only one of the $(\mathbf{y}, \mathbf{y}')$ entries differs from zero.

In case that $\mathbf{Y}$ is assumed to be a CFMP with local independence graph $G = (V, E)$ the non–diagonal elements of the matrix of the estimated transition probabilities reduce to

$$\hat{p}(s, t; (\mathbf{y}, \mathbf{y}')) = \frac{\Delta N(u; (\mathbf{y}_{\mathrm{cl}(k)}, y'_k))}{Z(u; \mathbf{y}_{\mathrm{cl}(k)})}$$

for $\mathbf{y}, \mathbf{y}'$ with $y_k \neq y'_k$.

5.3 Regression models

So far we considered models where the intensity process might depend on the question whether another event has previously occurred but not on the time of this occurrence. When the latter kind of dependence is to be included one has to turn to regression models, where the intensities are specified conditional on a suitably chosen covariate process. In the following, we still consider the absolutely continuous case, but where the mark specific individual intensity processes for n independent individuals are given as

$$\lambda_k^i(t|\boldsymbol{\theta}) = \alpha_k^i(t|\boldsymbol{\theta}, \mathbf{X}^i(t)) Z_k^i(t), \quad k = 1, \ldots, K;\ i = 1, \ldots, n;\ t \in \mathcal{T}, \tag{5.8}$$

where $\mathbf{X}^i$ is the covariate process of the i-th individual. This process should be $\mathcal{F}_t$–predictable and locally bounded. For our purposes, the covariates $\mathbf{X}^i(t)$ are assumed to be functions of the individual strict pre–t history H_{t-}^i, whereas $Z_k^i(t)$ usually indicates whether the individual is at risk for event e_k at time t. However, $\mathbf{X}^i$ may in general include further possibly fixed covariates but we will not take this explicitly into account. The only difference between the above model (5.8) and the previous multiplicative intensity model (5.2) is that α_k depends through the covariates on i and thus prevents the aggregation of the mark specific counting processes.

A special class of models for $\alpha_k^i(t|\boldsymbol{\theta}, \mathbf{X}^i(t))$ is now focussed on. This is the class of *multiplicative hazards* or *relative risk* regression models, where

$$\alpha_k^i(t|\boldsymbol{\theta}, \mathbf{X}^i(t)) = \alpha_k^0(t|\boldsymbol{\gamma}) r(\boldsymbol{\beta}^\top \mathbf{X}_k^i(t)), \quad k = 1, \ldots, K;\ i = 1, \ldots, n, \tag{5.9}$$

for a nonnegative real function $r(\cdot)$ so that $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$. In the above model formula, $\alpha_k^0(t|\boldsymbol{\gamma})$ are the baseline hazard functions which are often left unspecified, and $\mathbf{X}_k^i(t)$ are those covariates which affect the type k intensity. These have to be computable from $\mathbf{X}^i(t)$ and $\mathcal{F}_{t-}$. The model is called relative risk regression model because the regression parameter vector $\boldsymbol{\beta}$ specifies how the ratio between the hazards of different individuals depends on the covariates. In particular, if $r(\cdot) = \exp(\cdot)$ we have that the ratio of the hazards of two individuals with covariates $\mathbf{X}^i(t)$ and $\mathbf{X}^j(t)$ is simply given as

$$\exp\left(\boldsymbol{\beta}^\top (\mathbf{X}^i(t) - \mathbf{X}^j(t))\right).$$

Note that in the above notation we have that the vector of regression parameters $\boldsymbol{\beta}$ is the same for all type k intensities. This can be obtained by a suitable augmentation of the mark specific parameters and covariate vectors and it allows for the possibility that there are parameters common to all transitions.

Example (continued): Consider again the marked point process with three non–recurrent events $\mathcal{E} = \{e_1, e_2, e_3\}$ but without any local independence assumptions. In the following any covariate process is based on the history process $H_t^i = \{(T_s, E_s)|s = 1, \ldots, S;\ T_s \leq t,\ E_s \in \mathcal{E}\}$. Let us consider different plausible models and discuss their implications.

One of the most simple models is to assume that the baseline hazards for an event are the same regardless of what has happened in the past and to let the covariates just indicate which events have previously occurred, i.e. $\mathbf{X}^i(t) = (\mathbf{1}\{T_1^i < t\}, \mathbf{1}\{T_2^i < t\}, \mathbf{1}\{T_3^i < t\})^\top$. Assuming that the occurrence of an event affects the intensities for the other two events in the same way we may specify $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$ so that $\mathbf{X}_k^i$ essentially equals $\mathbf{X}^i$ but contains a zero in the k-th components. The intensities are then given as

$$\lambda_k^i(t) = \alpha_k^0(t) r(\beta_j \mathbf{1}\{T_j^i < t\} + \beta_l \mathbf{1}\{T_l^i < t\}) \mathbf{1}\{t \leq T_k^i\}, \quad k = 1, 2, 3;\ j, l \neq k,$$

implying that the occurrence of an event e_j changes the hazard for e_k by a factor $\exp(\beta_j)$. In this model, local dependencies are solely governed by the regression parameters. Since these are the same for all intensities it is not possible to model different local independence structures for the different events. It makes therefore more sense to consider $\boldsymbol{\beta} = (\beta_{1|2}, \beta_{1|3}, \beta_{2|1}, \beta_{2|3}, \beta_{3|1}, \beta_{3|2})^\top$ and $\mathbf{X}_1^i = (\mathbf{1}\{T_2^i < t\}, \mathbf{1}\{T_3^i < t\}, 0, 0, 0, 0)^\top$, $\mathbf{X}_2^i = (0, 0, \mathbf{1}\{T_1^i < t\}, \mathbf{1}\{T_3^i < t\}, 0, 0)^\top$ etc. This results in the following intensity processes:

$$\lambda_k^i(t) = \alpha_k^0(t) r(\beta_{k|j}\mathbf{1}\{T_j^i < t\} + \beta_{k|l}\mathbf{1}\{T_l^i < t\})\mathbf{1}\{t \leq T_k^i\}, \quad k = 1, 2, 3. \tag{5.10}$$

Now, each local independence is equivalent to a vanishing regression parameter, for instance $\beta_{1|2} = 0$ implies $\{2\} \not\to \{1\}|\{3\}$. Note that the above model does not allow for interactions. This would require additional regression parameters and covariates $\beta_{k|jl}$ and $\mathbf{1}\{\max(T_j, T_l) < t\}$, respectively.

Still, the foregoing models include no dependence on the time of previous events. This could for instance be achieved by choosing covariates with components based on $T_k^i\mathbf{1}\{T_k^i < t\}$ or on the durations $(t - T_k^i)\mathbf{1}\{T_k^i < t\}$, $k = 1, \ldots, K$, resulting in intensities of the form

$$\lambda_{k|jl}^i(t) \;=\; \alpha_k^0(t|\boldsymbol{\gamma}) r(\beta_{k|j}T_j^i + \beta_{k|l}T_l^i)\mathbf{1}\{\max(T_j^i, T_l^i) < t \leq T_k^i\}$$

or

$$\lambda_{k|jl}^i(t) \;=\; \alpha_k^0(t|\boldsymbol{\gamma}) r(\beta_{k|j}(t - T_j^i) + \beta_{k|l}(t - T_l^i))\mathbf{1}\{\max(T_j^i, T_l^i) < t \leq T_k^i\}.$$

Here, the regression parameters $\boldsymbol{\beta} = (\beta_{1|2}, \ldots, \beta_{3|2})^\top$ measure the influence of the occurrence times of or the duration since previous events, again assuming no interactions.

Since in the foregoing models the baseline hazard is still the same for an event e_k regardless of the past, local independence restrictions only affect the regression parameters. A last generalization therefore consists in allowing for transition specific baseline hazards, i.e. $\alpha_{k|0}^0$, $\alpha_{k|j}^0$, $\alpha_{k|l}^0$, and $\alpha_{k|jl}^0$. The transition specific intensities $\lambda_{k|jl}^i$ e.g. for the duration model are then given as

$$\begin{aligned}
\lambda_{k|0}^i(t) &= \alpha_{k|0}^0(t|\boldsymbol{\gamma}) r(0)\mathbf{1}\{t \leq \min(T_1^i, T_2^i, T_3^i)\}, \\
\lambda_{k|j}^i(t) &= \alpha_{k|j}^0(t|\boldsymbol{\gamma}) r(\beta_{k|j}(t - T_j^i))\mathbf{1}\{T_j^i < t \leq \min(T_k^i, T_l^i)\}, \\
\lambda_{k|jl}^i(t) &= \alpha_{k|jl}^0(t|\boldsymbol{\gamma}) r(\beta_{k|j}(t - T_j^i) + \beta_{k|l}(t - T_l^i))\mathbf{1}\{\max(T_j^i, T_l^i) < t \leq T_k^i\}.
\end{aligned}$$

Again, one could for instance include multiplicative interactions by a suitable augmentation of the covariate vector and additional interaction parameters $\beta_{k|jl}$. //

In the next section, we present maximum likelihood estimation and semiparametric estimation for the special case of multiplicative hazards models.

5.3.1 Maximum likelihood estimation

Assume that in the multiplicative intensity model (5.8) $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)^\top$ is a Q–dimensional parameter with similar assumptions as in Section 5.2.1. The main example is again the relative risk regression model above with $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$ where a variety of well–known models is already covered when choosing $r(\cdot)$ as the exponential function. This includes by a suitable specification of $\alpha_k^0(t|\boldsymbol{\gamma})$ for instance the simple exponential regression, the Weibull regression, and the Gompertz–Makeham regression.

The likelihood is similar to the non–regression situation but with an additional product over the individuals, i.e. $L(t|\boldsymbol{\theta}) = \prod_{k \in V} L_k(t|\boldsymbol{\theta})$ with the factors equal to

$$\prod_{i=1}^{n} \prod_{T^i_{s(k)} \le t} \alpha_k^i(t|\boldsymbol{\theta}, \mathbf{X}^i(T^i_{s(k)})) \cdot \exp\left(- \int_0^t \sum_{j=1}^{n} \alpha_k^j(s|\boldsymbol{\theta}, \mathbf{X}^j(s)) Z_k^j(s)\, ds \right), \qquad (5.11)$$

where $T^i_{s(k)}$ are the points in time where an event e_k occurs for the i-th individual. Let $C(\tau|\boldsymbol{\theta})$ and $\mathbf{U}(\tau|\boldsymbol{\theta})$ be the loglikelihood and the score vector at the maximal observation time τ. In the multiplicative hazard model (5.9) with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_R)^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^\top$, the score functions are for instance given by

$$\frac{\partial}{\partial \gamma_r} C(\tau|\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \left[\sum_{T^i_{s(k)} \le \tau} \frac{\partial}{\partial \gamma_r} \log(\alpha_k^0(\tau|\boldsymbol{\gamma})) - \int_0^\tau r\left(\boldsymbol{\beta}^\top \mathbf{X}_k^i(t)\right) \frac{\partial}{\partial \gamma_r} \alpha_k^0(t|\boldsymbol{\gamma})\, dt \right],$$

$r = 1, \ldots, R$, and

$$\frac{\partial}{\partial \beta_p} C(\tau|\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \left[\sum_{T^i_{s(k)} \le \tau} \frac{\partial}{\partial \beta_p} \log\left(r(\boldsymbol{\beta}^\top \mathbf{X}_k^i(T^i_{s(k)})) \right) - \int_0^\tau \alpha_k^0(t|\boldsymbol{\gamma}) \frac{\partial}{\partial \beta_p} r\left(\boldsymbol{\beta}^\top \mathbf{X}_k^i(t)\right) dt \right],$$

$p = 1, \ldots, P$. Under similar regularity conditions the maximum (partial) likelihood estimator defined as solution of $\mathbf{U}(\tau|\boldsymbol{\theta}) = 0$ has the same properties as given in

Section 5.2.1. This also holds for the corresponding likelihood ratio test.

With regard to inference about local independence structures a simplification is again mainly given if the parameter vector may be partitioned into mark specific subvectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, allowing for a separate maximization of the factors $L_k(\tau|\boldsymbol{\theta}_k)$. Assumptions about local independence in regression models of the above kind (5.8) and especially (5.9) usually affect the regression parameters which specify the dependence on earlier events. The baseline hazard is only affected by a local independence restriction if it is allowed to be different depending on past events. A general formulation for multi–state processes associated to general marked point processes would be notationally quite complicated so that we restrict ourselves in the following to the same example as before.

Example (continued): For the marked point process with the three non–recurrent events, we now assume a simple exponential regression model, where the baseline hazards are transition specific constants and $r(\cdot)$ is the exponential function. In addition, we also assume transition specific regression parameters as described above without interactions. In case that nothing is known about the local independence structure the intensity processes are given as

$$\begin{aligned}
\lambda^i_{k|0}(t) &= \gamma_{k|0}\, \mathbf{1}\{t \leq \min(T^i_1, T^i_2, T^i_3)\}, \\
\lambda^i_{k|j}(t) &= \gamma_{k|j}\, \exp(\beta_{k|j} T^i_j)\, \mathbf{1}\{T^i_j < t \leq \min(T^i_k, T^i_l)\}, \\
\lambda_{k|jl}(t) &= \gamma_{k|jl}\, \exp(\beta_{k|j} T^i_j + \beta_{k|l} T^i_l)\, \mathbf{1}\{\max(T^i_j, T^i_l) < t \leq T^i_k\}.
\end{aligned}$$

Here, the parameter vector $\boldsymbol{\theta}$ may obviously be partitioned into mark specific subvectors $\boldsymbol{\theta}_k = (\gamma_{k|0}, \gamma_{k|j}, \gamma_{k|l}, \gamma_{k|jl}, \beta_{k|j}, \beta_{k|l})^\top$ and the likelihood can be maximized by separately maximizing its mark specific factors (5.11). To derive the score functions let $\lambda_{k|\cdot} = \lambda_{k|0} + \lambda_{k|j} + \lambda_{k|l} + \lambda_{k|jl}$ and partition the sample into those individuals $I_{1|0}$ where $\{T^i_1 < \min(T^i_2, T^i_3)\}$, $I_{1|2}$ where $\{T^i_2 < T^i_1 < T^i_3\}$, and $I_{1|3}$, $I_{1|23}$ analogously. Still assuming that no censoring takes place, this yields for the foregoing model and for $k = 1$ the following score functions w.r.t. $\boldsymbol{\gamma}$:

$$\frac{\partial}{\partial \gamma_{1|0}} C_1(\tau|\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{i \in I_{1|0}} \frac{\partial}{\partial \gamma_{1|0}} \log(\gamma_{1|0}) - \int_0^\tau \frac{\partial}{\partial \gamma_{1|0}} \sum_{i=1}^n \lambda^i_{1|\cdot}(t)\, dt$$

$$= \frac{|I_{1|0}|}{\gamma_{1|0}} - \sum_{i=1}^{n} \min(T_1^i, T_2^i, T_3^i),$$

$$\begin{aligned}\frac{\partial}{\partial \gamma_{1|2}} C_1(\tau|\boldsymbol{\gamma}, \boldsymbol{\beta}) &= \sum_{i \in I_{1|2}} \frac{\partial}{\partial \gamma_{1|2}} \log(\gamma_{1|2}) - \int_0^\tau \frac{\partial}{\partial \gamma_{1|2}} \sum_{i=1}^{n} \lambda_{1|\cdot}^i(t)\, dt \\ &= \frac{|I_{1|2}|}{\gamma_{1|2}} - \sum_{i \in I_{1|2}} \exp(\beta_{1|2} T_2^i)(T_1^i - T_2^i) \\ &\quad - \sum_{i \in I_{1|23}} \exp(\beta_{1|2} T_2^i)(T_3^i - T_2^i) \mathbf{1}\{T_2^i < T_3^i\}.\end{aligned}$$

Analogously to the latter we have $\frac{\partial}{\partial \gamma_{1|3}} C_1(\tau|\boldsymbol{\gamma}, \boldsymbol{\beta})$ and similarly

$$\frac{\partial}{\partial \gamma_{1|23}} C_1(\tau|\boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{|I_{1|23}|}{\gamma_{1|23}} - \sum_{i \in I_{1|23}} \exp(\beta_{1|2} T_2^i + \beta_{1|3} T_3^i)(T_1^i - \max(T_2^i, T_3^i)).$$

From the above formulae we can see that the baseline hazard corresponds to the ratio between observed transitions and (weighted) exposure times. This reveals that individuals for whom the corresponding transitions could not be observed also contribute through their exposure time. To derive the scores resulting from differentiation w.r.t. $\boldsymbol{\beta}$ we have to consider the subsample $I_{1|2} \cup I_{1|23}$ of those individuals for whom the event e_2 occurred before e_1. Then, we get

$$\begin{aligned}\frac{\partial}{\partial \beta_{1|2}} C_1(\tau|\boldsymbol{\gamma}, \boldsymbol{\beta}) &= \sum_{i \in I_{1|2}} \left(\frac{\partial}{\partial \beta_{1|2}} (\beta_{1|2} T_2^i) - \int_{T_2^i}^{T_1^i} \frac{\partial}{\partial \beta_{1|2}} \gamma_{1|2} \exp(\beta_{1|2} T_2^i)\, dt \right) \\ &\quad + \sum_{i \in I_{1|23}} \left(\frac{\partial}{\partial \beta_{1|2}} (\beta_{1|2} T_2^i + \beta_{1|3} T_3^i) \right. \\ &\qquad - \int_{T_2^i}^{T_3^i} \frac{\partial}{\partial \beta_{1|2}} \gamma_{1|2} \exp(\beta_{1|2} T_2^i) \mathbf{1}\{T_2^i < T_3^i\}\, dt \\ &\qquad \left. - \int_{T_2^i \vee T_3^i}^{T_1^i} \frac{\partial}{\partial \beta_{1|2}} \gamma_{1|23} \exp(\beta_{1|2} T_2^i + \beta_{1|3} T_3^i)\, dt \right) \\ &= \sum_{i \in I_{1|2} \cup I_{1|23}} T_2^i - \sum_{i \in I_{1|2}} \gamma_{1|2} T_2^i \exp(\beta_{1|2} T_2^i)(T_1^i - T_2^i) \\ &\quad - \sum_{i \in I_{1|23}} \left(\gamma_{1|2} T_2^i \exp(\beta_{1|2} T_2^i)(T_3^i - T_2^i) \mathbf{1}\{T_2^i < T_3^i\} \right. \\ &\qquad \left. + \gamma_{1|23} T_2^i \exp(\beta_{1|2} T_2^i + \beta_{1|3} T_3^i)(T_1^i - \max(T_2^i, T_3^i)) \right),\end{aligned}$$

where $s \vee t = \max(s,t)$. A similar expression can be obtained for $\frac{\partial}{\partial \beta_{1|3}} C_1(\tau|\boldsymbol{\gamma},\boldsymbol{\beta})$. The scores for $k = 2,3$ can also be computed analogously.

In the present class of regression models, assumptions concerning local independencies affect both, the baseline hazard as well as the regression parameters. E.g. the local independence $\{2\} \not\to \{1\}|\{3\}$ in the present exponential model is equivalent to the parameter restrictions $\gamma_{1|2} = \gamma_{1|0}$, $\gamma_{1|23} = \gamma_{1|3}$, and $\beta_{1|2} = 0$. The score function w.r.t. $\boldsymbol{\gamma}$ in such a restricted model is then simplified to

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{1|0}} \tilde{C}_1(\tau|\boldsymbol{\gamma},\boldsymbol{\beta}) &= \frac{|I_{1|0} \cup I_{1|2}|}{\gamma_{1|0}} - \sum_{i=1}^{n} \min(T_1^i, T_3^i), \\
\frac{\partial}{\partial \gamma_{1|3}} \tilde{C}_1(\tau|\boldsymbol{\gamma},\boldsymbol{\beta}) &= \frac{|I_{1|3} \cup I_{1|23}|}{\gamma_{1|3}} - \sum_{i \in I_{1|3} \cup I_{1|23}} \exp(\beta_{1|3} T_3^i)(T_1^i - T_3^i),
\end{aligned}$$

and differentiating w.r.t. $\beta_{1|3}$ yields

$$\frac{\partial}{\partial \beta_{1|3}} \tilde{C}_1(\tau|\boldsymbol{\gamma},\boldsymbol{\beta}) = \sum_{i \in I_{1|3} \cup I_{1|23}} \left(T_3^i - \gamma_{1|3} T_3^i \exp(\beta_{1|3} T_3^i)(T_1^i - T_3^i)\right).$$

As noted earlier, another sensible model specification might be that all dependencies on earlier events are governed only by the regression parameters. This implies that we set a priori $\gamma_{k|0} = \gamma_{k|j} = \gamma_{k|l} = \gamma_{k|jl} = \gamma_k$. Then, the restrictions induced by local independencies affect only the regression parameters. If this kind of model assumption seems reasonable inference concerning local independence structures might as well be based on the semiparametric procedures where α_k^0 is left unspecified. Since the implications of local independence structures are in both situations almost the same as far as the estimation of the regression parameters is concerned we refer to the next section where such semiparametric models are treated. //

Although in the above example we assumed that for all individuals all three failure times T_1^i, T_2^i, T_3^i are observable, recall that the properties of the estimators remain the same under independent right censoring and left truncation.

Finally, let us mention that likelihood ratio tests of local independence might be constructed in a similar manner as in Section 5.2.1 but we refrain from going into details, here.

5.3.2 Semiparametric estimation for multiplicative hazards

In this section, we still consider the multiplicative hazards model (5.9) but with the mark specific baseline hazards $\alpha_k^0(t)$ left unspecified. They are only required to be nonnegative and to satisfy

$$A_k^0(t) = \int_0^t \alpha_k^0(s)\, ds < \infty, \quad k = 1, \ldots, K.$$

Note that the baseline hazards are now assumed to be the same for any event e_k without further subdivision into transitions between different states corresponding to past events. Thus, the dependence on previous events is only modeled via the regression parameters and the covariates but not through the baseline hazards anymore. For local independence graphs, the covariates $\mathbf{X}^i$ contain at least the information on the strict pre–t history H_{t-}^i, possibly transformed to the duration since the previous events. The likelihood $L(\tau|\mathbf{A}^0, \boldsymbol{\beta})$ now takes the following form:

$$\prod_{i=1}^n \prod_{k \in V} \prod_{T_{s(k)}^i \le \tau} [dA_k^0(T_{s(k)}^i) r(\boldsymbol{\beta}^\top \mathbf{X}_k^i(T_{s(k)}^i))] \cdot \exp\left(- \sum_{k \in V} \int_0^\tau \sum_{i=1}^n r(\boldsymbol{\beta}^\top X_k^i(s))\, dA_k^0(s) \right).$$

The estimation of the regression parameters makes use of the fact that for a fixed value of $\boldsymbol{\beta}$ the Nelson–Aalen estimator for the unspecified baseline hazard is given as

$$\hat{A}_k^0(t|\boldsymbol{\beta}) = \int_0^t \frac{\mathbf{1}\{Z_k(s) > 1\}}{\sum_{j=1}^n r(\boldsymbol{\beta}^\top \mathbf{X}_k^j(s))}\, dN_k(s) = \sum_{i=1}^n \sum_{T_{s(k)}^i \le t} \frac{1}{\sum_{j=1}^n r(\boldsymbol{\beta}^\top \mathbf{X}_k^j(T_{s(k)}^i))}. \quad (5.12)$$

Inserting

$$d\hat{A}_k^0(t|\boldsymbol{\beta}) = \frac{\Delta N_k(t)}{\sum_{j=1}^n r(\boldsymbol{\beta}^\top \mathbf{X}_k^j(t))}$$

in the above likelihood $L(\tau|\mathbf{A}^0, \boldsymbol{\beta})$ and leaving out the factors that do not depend on $\boldsymbol{\beta}$ yields the partially maximized likelihood (Gill, 1984)

$$\tilde{L}(\tau|\boldsymbol{\beta}) = \prod_{i=1}^n \prod_{k \in V} \prod_{T_{s(k)}^i \le \tau} \frac{r(\boldsymbol{\beta}^\top \mathbf{X}_k^i(T_{s(k)}^i))}{\sum_{j=1}^n r(\boldsymbol{\beta}^\top \mathbf{X}_k^j(T_{s(k)}^i))} \quad (5.13)$$

which is the Cox partial likelihood (Cox, 1972, 1975) for the situation of censored survival data and fixed covariates, as is well–known for $r(\cdot) = \exp(\cdot)$. Heuristically,

one can say that the factors of the above likelihood (5.13) correspond to the conditional probabilities for an event e_k to occur to the i-th individual at the specific time $T^i_{s(k)}$ given that the event e_k occurs to any individual at that time. It can be shown under regularity conditions (cf. Andersen et al., 1993, pp. 496) that the estimator resulting from maximizing the above (partial) likelihood (5.13) is asymptotically normally distributed.

Again, if there exist distinct parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ such that $r(\boldsymbol{\beta}^\top \mathbf{X}^i_k(t)) = r(\boldsymbol{\beta}_k^\top \mathbf{X}^i(t))$, $k = 1, \ldots, K$, we can maximize the above likelihood by separately maximizing the mark specific (partial) likelihoods $\tilde{L}_k(\tau|\boldsymbol{\beta}_k)$. In case that $r(\cdot) = \exp(\cdot)$ we get from (5.13) the corresponding (partial) loglikelihoods

$$\tilde{C}_k(\tau|\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{T^i_{s(k)} \leq \tau} \left[\boldsymbol{\beta}_k^\top \mathbf{X}^i(T^i_{s(k)}) - \log \left(\sum_{j=1}^{n} \exp\left(\boldsymbol{\beta}_k^\top \mathbf{X}^j(T^i_{s(k)})\right) Z^j_k(T^i_{s(k)}) \right) \right].$$

The separate maximization together with the implications of local independence restrictions is the objective of the following corollary.

Corollary 5.3.1 *Semiparametric estimator for exponential regression*
Consider n independent marked point processes $\{(T^i_s, E^i_s)|E^i_s \in \mathcal{E}, T^i_s \in \mathcal{T}\}$, $\mathcal{E} = \{e_1, \ldots, e_K\}$, and their corresponding individual counting processes N^i_k, $k \in V = \{1, \ldots, K\}$, with intensity processes that hold the multiplicative hazard model (5.9) and $r(\cdot) = \exp(\cdot)$. Let further the covariate process $\mathbf{X}^i$ be a function of the individual histories H^i_{t-}. Assume that $G = (V, E)$ is the local independence graph of $(N^i_1, \ldots, N^i_K)$, $i = 1, \ldots, n$, and that there exist distinct parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ such that $\exp(\boldsymbol{\beta}^\top \mathbf{X}^i_k(t)) = \exp(\boldsymbol{\beta}_k^\top \mathbf{X}^i(t))$, $k \in V$. Then, it holds:

(1) The above model is equivalent to a multiplicative hazard model given by

$$\lambda^i_k(t) = \alpha^0_k(t) \exp\left(\boldsymbol{\beta}^\top_{k|\mathrm{cl}(k)} \mathbf{X}^i_{\mathrm{cl}(k)}(t)\right) Z^i_k(t), \quad k = 1, \ldots, K;\ i = 1, \ldots, n,$$

where $\mathbf{X}^i_{\mathrm{cl}(k)}(t)$ is a function of the reduced history $H^{\mathrm{cl}(k);i}_{t-} = \sigma\{(T^i_s, E^i_s)|E^i_s \in \{e_k|k \in \mathrm{cl}(k)\}, 0 < T^i_s < t\}$ and $\boldsymbol{\beta}_{k|\mathrm{cl}(k)}$ contains only those components of $\boldsymbol{\beta}_k$ which govern the dependence on this history.

(2) The semiparametric estimators $\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)}$, $k \in V$, of the regression parameters given as solution of the score equations

$$\mathbf{U}(\tau|\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{T^i_{s(k)} \leq \tau} \left(\mathbf{X}^i_{\mathrm{cl}(k)}(T^i_{s(k)}) - \frac{S^{(1)}_k(T^i_{s(k)})}{S^{(0)}_k(T^i_{s(k)})} \right) = 0, \quad k = 1, \ldots, K,$$

with

$$S^{(0)}_k(t) = \sum_{j=1}^{n} \exp\left(\boldsymbol{\beta}^{\top}_{k|\mathrm{cl}(k)} \mathbf{X}^j_{\mathrm{cl}(k)}(t)\right) Z^j_k(t)$$

and

$$S^{(1)}_k(t) = \sum_{j=1}^{n} \mathbf{X}^j_{\mathrm{cl}(k)}(t) \exp\left(\boldsymbol{\beta}^{\top}_{k|\mathrm{cl}(k)} \mathbf{X}^j_{\mathrm{cl}(k)}(t)\right) Z^j_k(t),$$

have the following properties: Under the regularity conditions given in Andersen et al. (1993, pp. 496) there exists a unique solution with probability tending to one and $\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)} \xrightarrow{P} \boldsymbol{\beta}^0_{k|\mathrm{cl}(k)}$, where the latter is the true parameter. In addition, we have the asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)} - \boldsymbol{\beta}^0_{k|\mathrm{cl}(k)}) \xrightarrow{D} \mathcal{N}(0, \Sigma_k(\tau)).$$

A consistent estimator of the inverse covariance matrix $\Sigma_k(\tau)^{-1}$ is given by $\mathcal{I}_k(\tau|\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)})/n$, i.e. by inserting the estimated regression parameters in the information matrix

$$\mathcal{I}_k(\tau|\boldsymbol{\beta}_{k|\mathrm{cl}(k)}) = \sum_{i=1}^{n} \sum_{T^i_{s(k)} \leq \tau} \frac{S^{(2)}_k(T^i_{s(k)}) - S^{(1)}_k(T^i_{s(k)})}{S^{(0)}_k(T^i_{s(k)})},$$

where

$$S^{(2)}_k(t) = \sum_{j=1}^{n} (\mathbf{X}^j_{\mathrm{cl}(k)} \mathbf{X}^{j\top}_{\mathrm{cl}(k)}(t) \exp\left(\boldsymbol{\beta}^{\top}_{k|\mathrm{cl}(k)} \mathbf{X}^j_{\mathrm{cl}(k)}(t)\right) Z^j_k(t).$$

(3) For the semiparametric estimators of the baseline hazards

$$\hat{A}^0_k(t|\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)}) = \sum_{i=1}^{n} \sum_{T^i_{s(k)} \leq t} \frac{1}{\sum_{j=1}^{n} \exp\left(\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)} \mathbf{X}^j_{\mathrm{cl}(k)}(T^i_{s(k)})\right)},$$

resulting from inserting $\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)}$ in (5.12), it holds that

$$(\sqrt{n}(\hat{A}^0_k(\cdot|\hat{\boldsymbol{\beta}}_{k|\mathrm{cl}(k)}) - A^0_k(\cdot))|k = 1, \ldots, K)$$

converges weakly to a K–variate Gaussian process with mean zero. //

The above results remain valid for more general risk functions $r(\cdot)$ under some additional conditions for the existence of a solution to the score equations. Further, all results have been stated under the assumption that there are no ties, i.e. no two events at the same time. Although ties have probability zero they may in praxis occur due to limited measurement precision. Then, the estimators can be adjusted by allowing for several jumps at a time.

It is straightforward to construct tests of local independence from the above asymptotic distribution of $\hat{\boldsymbol{\beta}}$ following the likelihood ratio principle. The hypothesis $H_0 : \{j\} \not\to \{k\}|V\backslash\{j,k\}$ is equivalent to $\boldsymbol{\beta}_{k|j} = 0$, where $\boldsymbol{\beta}_{k|j}$ is the vector of parameters that govern the effect of the history $H^j_{t^-}$ on $\lambda_k(t)$. The degrees of freedom depend of course on the dimension of this parameter.

Example (continued): Considering again a marked point process with $\mathcal{E}$ consisting of three non–recurrent events and local independence structure given in Figure 5.1, we may specify the mark specific intensities as follows:

$$\begin{aligned}
\lambda^i_1(t) &= \alpha^0_1(t)\mathbf{1}\{t \le T^i_1\},\\
\lambda^i_2(t) &= \alpha^0_1(t)\exp(\boldsymbol{\beta}^\top_{2|13}\mathbf{X}^i_{\{1,3\}}(t))\mathbf{1}\{t \le T^i_2\},\\
\lambda^i_3(t) &= \alpha^0_1(t)\exp(\boldsymbol{\beta}^\top_{3|2}\mathbf{X}^i_2(t))\mathbf{1}\{t \le T^i_3\},
\end{aligned}$$

where $\mathbf{X}^i_{\{1,3\}}(t)$ and $\mathbf{X}^i_2(t)$ are functions of $H^{\{1,3\};i}_{t^-}$ and $H^{2;i}_{t^-}$, respectively. As in an earlier example (5.10), the covariate processes might for instance just indicate whether the other events have occurred or not, their occurrence changing the baseline hazard multiplicatively. Thus, $\mathbf{X}^i_{\{1,3\}}(t) = (\mathbf{1}\{T^i_1 < t\}, \mathbf{1}\{T^i_3 < t\})$ and $\mathbf{X}^i_2(t) = \mathbf{1}\{T^i_2 < t\}$ which together with $\boldsymbol{\beta}_{2|13} = (\beta_{2|1}, \beta_{2|3})$ and $\beta_{3|2}$ yields the score equations

$$\begin{aligned}
|I_{2|1\cdot}| &= \sum_{i=1}^n \frac{|I^i_{2|1}|\exp(\beta_{2|1}) + |I^i_{2|13}|\exp(\beta_{2|1}+\beta_{2|3})}{|I^i_{2|0}| + |I^i_{2|1}|\exp(\beta_{2|1}) + |I^i_{2|3}|\exp(\beta_{2|3}) + |I^i_{2|13}|\exp(\beta_{2|1}+\beta_{2|3})},\\
|I_{2|3\cdot}| &= \sum_{i=1}^n \frac{|I^i_{2|3}|\exp(\beta_{2|3}) + |I^i_{2|13}|\exp(\beta_{2|1}+\beta_{2|3})}{|I^i_{2|0}| + |I^i_{2|1}|\exp(\beta_{2|1}) + |I^i_{2|3}|\exp(\beta_{2|3}) + |I^i_{2|13}|\exp(\beta_{2|1}+\beta_{2|3})},\\
|I_{3|2\cdot}| &= \sum_{i=1}^n \frac{|I^i_{3|2\cdot}|\exp(\beta_{3|2})}{|I^i_{3|0} \cup I^i_{3|1}| + |I^i_{3|2\cdot}|\exp(\beta_{3|2})},
\end{aligned}$$

where $I_{2|1.} = I_{2|1} + I_{2|13}$ and $I^i_{2|1} = \{j | T^j_1 < T^i_2 \leq \min(T^j_2, T^j_3); j = 1, \ldots, n\}$ and the other quantities are defined analogously. Thus, the estimators $\hat{\beta}_{2|1}$, $\hat{\beta}_{2|3}$, and $\hat{\beta}_{3|2}$ are given by equating the number of observed transitions with the expected ones.
If we want to test the hypothesis $H_0 : \{3\} \not\to \{2\}|\{1\}$ which is equivalent to $\beta_{2|3} = 0$ then the estimator $\tilde{\beta}_{2|1}$ of $\beta_{2|1}$ under the null hypothesis is given as solution of

$$|I_{2|1.}| = \sum_{i=1}^{n} \frac{|I^i_{2|1.}| \exp(\beta_{2|1})}{|I^i_{2|0} \cup I^i_{2|3}| + |I^i_{2|1.}| \exp(\beta_{2|1})}.$$

It follows that the 'likelihood ratio' test statistic for the above hypothesis reads as

$$2 \sum_{i=1}^{n} \Big(\hat{\beta}_{2|1} \mathbf{1}\{T^i_1 < T^i_2\} + \hat{\beta}_{2|3} \mathbf{1}\{T^i_3 < T^i_2\} - \tilde{\beta}_{2|1} \mathbf{1}\{T^i_1 < T^i_2\} -$$
$$(|I^i_{2|1}| \exp(\hat{\beta}_{2|1}) + |I^i_{2|3}| \exp(\hat{\beta}_{2|3}) + |I^i_{2|13}| \exp(\hat{\beta}_{2|1} + \hat{\beta}_{2|3}) - |I^i_{2|3}| - |I^i_{2|1.}| \exp(\tilde{\beta}_{2|1})) \Big).$$

This is approximately χ^2–distributed with one degree of freedom as implied by Corollary 5.3.1. //

Note that in the preceding example we did not consider dependencies on the times of previous events although this was used to motivate this section. However, the above considerations may easily be generalized to this kind of dependencies requiring even more notational effort so that we desist from doing so.

Discussion

Local independence is an intuitive association concept for events histories. It relates the presence of a process to the past of another process against the background of a specific dynamic system described by a multivariate stochastic process. In this thesis, it has been shown that a straightforward and meaningful graphical representation of local independence structures is possible, comprising more information than only pairwise local independencies. The central result, in this respect, is given by the equivalence of the dynamic Markov properties in Chapter 4. In particular, the graphical δ–separation provides a simple method to assess whether local independence structures remain valid even though part of the information on past events is ignored. In the following, we want to stress the main implications of this result and address some open questions that might be subject to future research.

Let us illustrate the importance of the separation theorem for statistical modeling and inference by some of its implications. Basically, δ–separation facilitates the identification of subsets of events, so–called collapsible subsets, which can be analyzed separately without distortion of the independence structure. This property of the model clearly contributes to reduce the complexity of any statistical procedure. The collapsibility property may also be used to decompose a graph into sensible subgraphs in order to simplify the interpretation.
Further, it is well–known that in observational studies a problem arises when important and possibly confounding variables or processes remain unobserved. It is, of course, impossible to eliminate this danger completely. Yet, it might be instructive to explicate the conditions required for accurate inference from observational data so as to justify at least their plausibility in a given data situation. It seems desirable to follow–up this topic of causal inference from local independence graphs beyond

the results of the present thesis.
Another aspect concerns possible simplifications of statistical inference procedures implied by the separation theorem as described in Chapter 5. The power of a statistical test is usually increased when the model can be reduced to a submodel that contains only the minimal information relevant for the considered hypotheses. The same effect can be obtained for the precision of estimation procedures. A nonparametric statistical test for local independence without the Markov assumption for the underlying process has not been derived yet.

The topic of estimation and testing is closely related to model selection procedures which is obviously of major importance for the practical use of local independence graphs. Simple stepwise procedures, such as forward and backward selection, might be immediately applicable to local independence graphs based on the statistical tests presented in Chapter 5. Such standard procedures, however, usually do not exploit the graphical structure and the collapsibility properties. These could enter the analysis as follows. A statistical test of local independence, i.e. a test on whether a specific directed edge should be included in the graph or not, is always conditional on some history. Due to the different dynamic Markov properties the conditioning set of histories may be chosen in different ways and the conditions for collapsibility permit to reduce this set to a minimal relevant, not necessarily unique, subset. Thus, an edge–deletion or –inclusion test can be performed in several ways. On the one hand, this can be used to double–check the test decision. On the other hand, it would be desirable to derive criteria and algorithms for an 'optimal' choice of the conditioning set, as for instance maximal power of the corresponding test as mentioned above. Further, it is not satisfactory that the standard selection procedures issue only one final model although the data usually support more than one model. A method yielding several plausible models could be more informative. It therefore seems reasonable to explore alternative strategies, as for instance the Edwards–Havránek procedure (Edwards and Havránek, 1985) or Bayesian model selection, with regard to their applicability to local independence graphs.

Besides the separation theorem and its implications for statistical inference, a remarkable property of local independence should be emphasized. It is an *asymmetric*

irrelevance relation. Although several examples for asymmetric irrelevance relations may be found in the literature on probabilistic modeling, no general framework has been developed yet. The findings presented in Chapter 3 regarding the axioms of asymmetric graphoids as well as the derivation of the asymmetric δ–separation for the representation of these relations (cf. Chapter 1) may constitute a first step in this direction. An interesting aspect for future research in this context concerns the completeness of δ–separation. The question, whether we can obtain an exhaustive characterization of local independence structures by the graphical representation proposed in this thesis, remains open. Similarly, further investigations are needed to find out whether specific classes of point processes may be regarded as 'Markov–perfect' with respect to a local independence graph, i.e. they should comprise no more than those independencies that can be read off the graph.

Additional generalizations seem possible with regard to the underlying processes. Although the presentation of local independence graphs is mostly restricted to the framework of marked point processes and therefore mainly applies to event history data, recall that the original definition of local independence as given in Chapter 3 refers to general multivariate processes which allow for a Doob–Meyer decomposition and could cover the continuous as well as the discrete time situation. Moreover, the heuristic interpretation of local independence as relation between the presence of one process and the past of another one is obviously not restricted to marked point processes. These two aspects suggest that a substantial generalization is possible. However, since local independence is formally defined as a property of the compensator it is not possible to capture all forms of dependencies through this concept in more general processes. In addition, the orthogonality of the martingales does not ensure independence of the increments but only uncorrelation and it is furthermore an unrealistic assumption for the discrete time situation. An application of the main ideas presented in the preceding chapters to other frameworks than marked point processes would therefore require considerably more investigation regarding the properties and interpretation of the resulting graphs. Nevertheless, it has become clear that local independence graphs constitute a promising approach to graphical modeling of event history data and an extension to more general dynamic systems is by all means desirable.

Appendix A

Conditional expectation and conditional independence

In this chapter, we develop the general notions of conditional expectation and conditional independence. We adopt a rather technical approach in order to clarify the relation among these two concepts. In addition, this approach is chosen with view to the theory of martingales presented in the next appendix. For a thorough treatment of conditional expectation we refer to Bauer (1991), or Gänssler and Stute (1977), and for conditional independence to Dawid (1979; 1980). Overviews regarding these topics may for instance be found in Shao (1999) and Lauritzen (1996), respectively.

A.1 Conditional expectation

The notion of conditional expectation is central to conditional independence as well as to the martingale theory used in this thesis. We therefore give a brief overview over the definition as well as some basic results. We restrict ourselves to real–valued random variables, but the results can be generalized e.g. to set–valued random variables such as the history processes considered in Chapter 3.

Definition A.1.1 *Conditional expectation*
Let X be a random variable on $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G}$ be a sub–σ–algebra of $\mathcal{F}$. The *conditional expectation of X given $\mathcal{G}$* is denoted by $E(X|\mathcal{G})$ and given by any random variable Y that satisfies

(1) Y is $\mathcal{G}$–measurable and

(2) $\int_B Y\,dP = \int_B X\,dP$ for all $B \in \mathcal{G}$. //

The conditional expectation $E(X|\mathcal{G})$ can be viewed as the information on X contained in $\mathcal{G}$. If $E(|X|) < \infty$ then the conditional expectation exists. It is in general not unique, but if Y and W are two $\mathcal{G}$–measurable random variables that satisfy the above conditions we have $P(\{Y \neq W\}) = 0$, i.e. they are equivalent. Different conditional expectations of X given $\mathcal{G}$ are called *versions* of $E(X|\mathcal{G})$. In the following, $E(X|\mathcal{G})$ is treated as if it was unique and all statements concerning conditional expectations are only P–almost sure. Some important properties are given in the next proposition.

Proposition A.1.2 *Properties of* $E(X|\mathcal{G})$
Let X be a random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ a sub–σ–algebra of $\mathcal{F}$. Assuming that the following conditional expectations always exist, we have:

(1) $\mathcal{G} = \{\emptyset, \Omega\}$ implies $E(X|\mathcal{G}) = E(X)$.

(2) $E(E(X|\mathcal{G})) = E(X)$.

(3) If X is $\mathcal{G}$–measurable, then $E(X|\mathcal{G}) = X$.

(4) For sub–σ–fields $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$, we have $E(E(X|\mathcal{G}_1)|\mathcal{G}_2) = E(E(X|\mathcal{G}_2)|\mathcal{G}_1) = E(X|\mathcal{G}_1)$.

(5) If Y is $\mathcal{G}$–measurable, then $E(XY|\mathcal{G}) = YE(X|\mathcal{G})$.

(6) $E(aX + bY|\mathcal{G}) = aE(X|\mathcal{G}) + bE(Y|\mathcal{G})$ for any constant real values a and b.

(7) For any convex, real–valued function g we have $E(g(X)|\mathcal{G}) \geq g(E(X|\mathcal{G}))$ (Jensen inequality). //

Proof: Bauer (1991, pp. 121) □

Let us now focus on the case, where the conditioning σ–algebra is generated by a random variable. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{B}$ the borel σ–field on $\mathbb{R}$. The *σ–field generated by* X is defined as $\sigma(X) = \{X^{-1}(B)|B \in \mathcal{B}\}$.

Definition A.1.3 *Conditional expectation for random variables*
Let $(\Omega, \mathcal{F}, P)$ be a probability space and X and Y two random variables on this space. Then, the *conditional expectation of X given Y,* denoted by $E(X|Y)$, is defined to be the conditional expecation of X given $\sigma(Y)$, i.e. $E(X|\sigma(Y))$. //

The conditional expectation can be calculated by exploiting the following results.

Lemma A.1.4 *Factorization*
Let $(\Omega, \mathcal{F}, P)$ be a probability space and X and Y two random variables on this space. Then, there exists a Borel function g on the space of Y such that

$$E(X|Y)(\omega) = g(Y(\omega)).$$

This is called the *factorization property* for conditional expectations.

Proof: Gänssler and Stute (1977, p. 190). □

We may therefore define $E(X|Y = y) = g(y)$ which is no random variable anymore but a real number. Further, we define P_X as the probability distribution induced by a random variable X, i.e. $P_X : \mathcal{B} \to [0,1]$ with $P_X(B) = P(X^{-1}(B))$, $B \in \mathcal{B}$. Similarly, the joint distribution $P_{X,Y}$ of X and Y is given as $P_{X,Y}(B,C) = P(X^{-1}(B) \cap Y^{-1}(C))$, $B, C \in \mathcal{B}$. Consider now the absolutely continuous case, where this joint distribution has a joint density $f : \mathbb{R}^2 \to \mathbb{R}^+$. Then, we have the following corollary.

Corollary A.1.5 *Calculation of conditional expectation*
Under the foregoing assumptions and provided that

$$f(y) = \int f(x,y)\, dx > 0,$$

the conditional expectation is given as

$$E(X|Y = y) = \frac{\int x f(x,y)\, dx}{f(y)} \quad \text{for } P_Y\text{–almost all } y \in \mathbb{R}.$$

Proof: Bauer (1991, p. 130). □

A similar result can be obtained for discrete probability distributions. Due to the preceding corollary, it makes sense to define the *conditional density* of X given $Y = y$ as

$$f(x|y) = \frac{f(x,y)}{f(y)},$$

for $f(y) > 0$.

All the foregoing results can be applied to conditional probabilites by noting that $P(A|\mathcal{G}) = E(\mathbf{1}\{A\}|\mathcal{G})$, $A \in \mathcal{F}$. More precisely we have:

Definition A.1.6 *Conditional probability*
Let $(\Omega, \mathcal{F}, P)$ be a probability space and X and Y two random variables on this space. The *regular conditional probability of X given $Y = y$* denoted by $P_{X|Y=y}$ is defined as function from $\mathcal{B} \times \mathbb{R} \to [0,1]$ with

(1) $A \to P_{X|Y=y}(A|y)$ is a probability measure on $\mathcal{B}$ for all $y \in \mathbb{R}$, and

(2) $y \to P_{X|Y=y}(A|y)$ is a version of $P(X^{-1}(A)|Y = y)$ for all $A \in \mathcal{B}$.

Instead of $P_{X|Y=y}(\cdot|y)$ we also write $P_{X|Y}(\cdot|y)$. //

Conditions for the existence and uniqueness of regular conditional probabilities may be found in Bauer (1991, pp. 396) or Gänssler and Stute (1977, pp. 196). With this notion of conditional probability it can be shown for a Borel function g with $E(|g(X,Y)|) < \infty$ that

$$E(g(X,Y)|Y = y) = \int_{\mathbb{R}} g(x,y) dP_{X|Y}(x|y).$$

In particular, it follows for the case where a density exists that

$$P_{X|Y}(A|Y = y) = \int_A f(x|y)\, dx$$

for all $y \in \{f_Y > 0\}$ with $f_Y(y) = \int f(x,y)\, dx$.

A.2 Conditional independence

Before defining conditional independence we consider marginal independence.

Definition A.2.1 *Independence*
Let $(\Omega, \mathcal{F}, P)$ be a probability space. A family $(A_i)_{i\in I}$, $A_i \in \mathcal{F}$, *is independent* if for any finite subset $J \subset I$

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j). \tag{A.1}$$

//

Independence of random variables is simply defined as independence of the corresponding induced σ–algebras.

Definition A.2.2 *Independence of random variables*
A family $(X_i)_{i\in I}$ of random variables on $(\Omega, \mathcal{F}, P)$ is *independent* if the family of σ–fields $(\sigma(X_i))_{i\in I}$ is independent, i.e. if (A.1) holds for any finite subset $J \subset I$ and any $(A_j)_{j\in J}$ with $A_j \in \sigma(X_j)$. //

In particular, we have the following implications of independence for conditional expectations.

Corollary A.2.3 *Independence for conditional expectations*
Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$. If X and Y are independent, we have $E(X|\sigma(Y)) = E(X)$. //

An even stronger result is given in the next proposition:

Proposition A.2.4 *Independence*
Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$. The following statements are equivalent:

(1) X and Y are independent.

(2) $P_{X|Y=y} = P_X$ for P_Y–almost all y.

(3) $P_{X,Y} = P_X P_Y$.

(4) If the random variables have the (joint) cumulative distribution functions $F_{X,Y}$, F_X, and F_Y, then $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all x, y.

Proof: Bauer (1991, pp. 65, 401), Shao (1999, pp. 34). □

The above result can be generalized to vector–valued random variables. In particular we have that a vector $\mathbf{X} = (X_1, \ldots, X_K)$ of random variables is independent iff the joint probability $P_{X_1,\ldots,X_K}$ is equal to the product of the marginal distributions $\prod P_{X_k}$. From this it follows for the absolutely continous case that independence is equivalent to the factorization of the density

$$f(x_1, \ldots, x_K) = \prod_{k=1}^{K} f(x_k),$$

where $f(x_1, \ldots, x_K)$ is the density of the joint distribution $P_{X_1,\ldots,X_K}$ and $f(x_k)$ of the marginal distribution P_{X_k}.

Now, we turn to conditional independence which is slightly more complicated. However, the following definition of conditional independence for events is well–known.

Definition A.2.5 *Conditional independence*
Let $(\Omega, \mathcal{F}, P)$ be a probability space. For disjoint subsets $A, B, C \in \mathcal{F}$, with $P(C) \neq 0$ we say that A is *conditionally independent of B given C* if

$$P(A \cap B|C) = P(A|C)P(B|C),$$

where $P(A|C) = P(A \cap C)/P(C)$ and $P(B|C)$ analogously. //

In order to generalize this idea to the conditional independence of random variables, recall that a conditional probability is a specific conditional expectation.

Definition A.2.6 *Conditional independence for random variables*
Let X, Y, and Z be random variables on $(\Omega, \mathcal{F}, P)$. Then, we say that X and Y are *conditionally independent given Z* if for any $A \in \sigma(X)$ there exists a version of the conditional probability $P(A|Y, Z)$ which is $\sigma(Z)$–measurable. //

Conditional independence is denoted by $X \perp\!\!\!\perp Y|Z$ (Dawid, 1979, 1980). The above definition implies that if $X \perp\!\!\!\perp Y|Z$ then the probability distribution of X given knowledge on both values of Y and Z remains the same if Y is disgarded, i.e. Y contributes no more information than already included in Z. We have the following equivalent and more familiar characterizations of conditional independence.

Corollary A.2.7 *Conditional independence and factorization*
Let X, Y, and Z be random variables on $(\Omega, \mathcal{F}, P)$.

(1) If X, Y, and Z are discrete, the conditional independence $X \perp\!\!\!\perp Y|Z$ is equivalent to

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$$

for all x, y, z with $P(Z = z) > 0$.

(2) When the three variables admit a joint density w.r.t. a σ–finite product measure ν, we have the following equivalent expressions.

$$\begin{aligned} & X \perp\!\!\!\perp Y|Z \\ \Leftrightarrow \quad & f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z) \\ \Leftrightarrow \quad & f_{X|YZ}(x|y, z) = f_{X|Z}(x|z) \\ \Leftrightarrow \quad & f_{XYZ}(x, y, z) f_Z(z) = f_{XZ}(x, z) f_{YZ}(y|z). \end{aligned}$$

Proof: Dawid (1979), Lauritzen (1996, p. 28). □

It is easily checked that conditional independence satisfies the graphoid axioms defined in Chapter 1 (cf. Lauritzen, 1996).

Corollary A.2.8 *Graphoid axioms*
Let X, Y, and Z be random variables on $(\Omega, \mathcal{F}, P)$ and let h denote an arbitrary measurable function on the space of X. Conditional independence satisfies the following properties:

(C1) *Symmetrie:* $X \perp\!\!\!\perp Y|Z \Rightarrow Y \perp\!\!\!\perp X|Z$,

(C2) *Decomposition:* $X \perp\!\!\!\perp Y|Z$ and $U = h(X) \Rightarrow U \perp\!\!\!\perp Y|Z$,

(C3) *Weak union:* $X \perp\!\!\!\perp Y|Z$ and $U = h(X) \Rightarrow X \perp\!\!\!\perp Y|(Z,U)$,

(C4) *Contraction:* $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp W|(Y,Z) \Rightarrow X \perp\!\!\!\perp (W,Y)|Z$.

In addition, it is obvious that $X \perp\!\!\!\perp Y|X$ is always true, i.e. *redundancy* holds. Further, if the joint density of the involved variables w.r.t. a poduct measure is poisitve and continuous, we have

(C5) *Intersection:* $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ then $X \perp\!\!\!\perp (Y,Z)$.

Proof: The proof of (C5) can be found in Lauritzen (1996, p. 29). □

The above definitions and results can be generalized to random vectors $\mathbf{X} = \mathbf{X}_V = (X_1, \ldots, X_K)$, $V = \{1, \ldots, K\}$. Then, we write $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B|\mathbf{X}_C$ or briefly $A \perp\!\!\!\perp B|C$ to denote that the subvector $\mathbf{X}_A$ is conditionally independent of $\mathbf{X}_B$ given $\mathbf{X}_C$. For the graphoid axioms we then mostly have that $h(\mathbf{X})$ denotes a subvector of $\mathbf{X}$ so that e.g. the property of decomposition may alternatively be formulated as: $A \perp\!\!\!\perp B|C$ and $D \subset A$ then $D \perp\!\!\!\perp B|C$. The latter interpretation is used in Chapter 1.

Appendix B

Stochastic processes

In this appendix, we give a brief overview over the terminology used in the preceding chapters and state some basic results, in particular the Doob-Meyer decomposition. Note, that most of the counting process and martingale framework has been developed with view to the asymptotic behaviour of statistics based on counting processes as well as for handling censored observations. Since these aspects are not central to the theory of local independence graphs, we do not go into details.

B.1 Basic notions

The following definitions and results are mainly based on Fleming and Harrington (1984), Andersen et al. (1993), and Brémaud (1981). We consider a continuous time parameter $\mathcal{T}$ with $\mathcal{T} = [0, \tau)$ or $\mathcal{T} = [0, \tau]$, $0 < \tau \leq \infty$.

Definition B.1.1 *Stochastic process and some properties*
A *stochastic process* Y is a family $\{Y(t)|t \in \mathcal{T}\}$ of random variables $Y(t)$ each defined on a probability space $(\Omega, \mathcal{F}, P)$ with $Y(t) \in \mathcal{S}$ for all $t \in \mathcal{T}$, where $\mathcal{S}$ denotes the set of states.
The process Y is called *cadlag (continu à droite, limité à gauche)* if its sample paths $\{Y(t, \omega)|t \in \mathcal{T}\}$ are right–continuous with left–hand limits for almost all $\omega \in \Omega$.
Further, Y is said to be

(1) *integrable* if $\sup_{0 \leq t < \infty} E(|Y(t)|) < \infty$,

(2) *square integrable* if $\sup_{0 \leq t < \infty} E(Y(t)^2) < \infty$,

(3) *bounded* if there exists a finite constant c such that $P(\sup_{0\leq t<\infty}|Y(t)|<c)=1$.

(3) *finite variation process* if $P(\int_0^t|Y(ds)|<\infty)=1$ for all $t\in\mathcal{T}$. //

As seen in the previous chapters, counting processes provide an appropriate basis for event history analysis.

Definition B.1.2 *Counting process*
A cadlag stochastic process N with state space $\mathcal{S}=\{0,1,2,\ldots\}$, zero at time zero, paths which are piecewise constant and non–decreasing with jumps $\Delta N(t)=N(t)-N(t^-)=1$ is called a *counting process.*
A multivariate process $\mathbf{N}=(N_1,\ldots,N_K)$, where the components N_k are counting processes, $k=1,\ldots,K$, is called *multivariate counting process* if no two components jump at the same time. //

The following definition of filtrations allows a rigorous formulation of the concept of information accruing over time.

Definition B.1.3 *Filtration / usual conditions / adapted process*
Let $(\Omega,\mathcal{F},P)$ be a probability space. A family of σ–algebras $\{\mathcal{F}_t|t\in\mathcal{T}\}$ with $\mathcal{F}_t\subset\mathcal{F}$ for all $t\in\mathcal{T}$ is called a *filtration* if it is

(1) increasing, i.e. $\mathcal{F}_s\subseteq\mathcal{F}_t$ for all $s\leq t,\ s,t\in\mathcal{T}$, and

(2) right–continuous, i.e.
$$\mathcal{F}_s=\bigcap_{t>s}\mathcal{F}_t \text{ for all } s\in\mathcal{T}.$$

Left– and right–hand limits of a filtration are defined as

$$\mathcal{F}_{t^-}=\sigma\{\bigcup_{h>0}\mathcal{F}_{t-h}\}=\bigvee_{h>0}\mathcal{F}_{t-h} \qquad \text{and} \qquad \mathcal{F}_{t^+}=\bigcap_{h>0}\mathcal{F}_{t+h}.$$

For notational convenience we mostly write $\mathcal{F}_t$ instead of $\{\mathcal{F}_t|t\in\mathcal{T}\}$ to denote a filtration.
A filtration $\{\mathcal{F}_t|t\in\mathcal{T}\}$ satisfies the *usual conditions* if it is complete, i.e. for all $A\subset B\in\mathcal{F}$: $P(B)=0\Rightarrow A\in\mathcal{F}_0$.
Finally, a stochastic process Y is said to be *adapted* to a filtration $\mathcal{F}_t$ if $Y(t)$ is $\mathcal{F}_t$–measurable for all $t\in\mathcal{T}$. //

A filtration generated by a process is called the history of that process.

Definition B.1.4 *(Internal) history*
Let Y be a stochastic process on $(\Omega, \mathcal{F}, P)$. The family of σ–algebras $\mathcal{F}_t = \sigma\{Y(s)|s \leq t\}$ is termed the *internal history* of Y. Any larger filtration $\mathcal{G}_t$, i.e. $\mathcal{G}_t \supset \mathcal{F}_t$ for all $t \in \mathcal{T}$, is a *history* of Y. //

Note that if Y is a counting process then its internal history is indeed right–continuous (Andersen et al., 1993, p. 61). Further, any process is obviously adapted to its internal history.

Most results in the counting process framework can be given for quite general processes as long as they fulfill specific conditions at least 'locally'. In order to specify this notion of 'local' we need the definition of a stopping time.

Definition B.1.5 *Stopping time*
Let $\mathcal{F}_t$ be a filtration on $(\Omega, \mathcal{F}, P)$. A nonnegative random variable T on $(\Omega, \mathcal{F}, P)$ is called a $\mathcal{F}_t$*–stopping time* if $\{T \leq t\} \in \mathcal{F}_t$ for all $t \in \mathcal{T}$. //

The martingale property may now be defined either for the whole process or only locally.

Definition B.1.6 *(Local) martingale*
A cadlag stochastic process M on $(\Omega, \mathcal{F}, P)$ with a filtration $\mathcal{F}_t$ is called a $\mathcal{F}_t$–*martingale* if it is

(1) integrable,

(2) adapted to $\mathcal{F}_t$, and

(3) satisfies the *martingale property* $E(M(t+s)|\mathcal{F}_t) = M(t)$ a.s. for all $s, t \geq 0$.

It is called a $\mathcal{F}_t$*–submartingale* if the last condition is replaced by the *submartingale property* $E(M(t+s)|\mathcal{F}_t) \geq M(t)$ a.s. for all $s, t \geq 0$.
The process M is called a *local* $\mathcal{F}_t$–(sub)martingale if there exists an increasing sequence of $\mathcal{F}_t$–stopping times $\{T_n|n \in \mathbb{N}\}$ with $P(T_n > t) \to 1$ for $n \to \infty$ for all $t \in \mathcal{T}$, such that the stopped processes $M_n = \{M(t \wedge T_n)|t \in \mathcal{T}\}$ are $\mathcal{F}_t$–(sub)martingales for all $n = 1, 2, \ldots$, where $(t \wedge s) = \min(t, s)$. //

Two simple properties of a martingale provide insight into the way martingales are used in the counting process framework for event history analysis. First, if M is a $\mathcal{F}_t$–martingale then

$$E(M(t)|\mathcal{F}_{t^-}) = M(t^-),$$

i.e. the martingale property carries over to the left–hand limits. Second, we have that

$$E(M(dt)|\mathcal{F}_{t^-}) = 0,$$

i.e. given the strict pre–t history the expected increment is zero. Both properties are cited from Fleming and Harrington (1991, p. 24).
Note that any counting process is a local submartingale w.r.t. its internal filtration (Andersen et al., 1993, p. 73). Local integrability and local boundedness are defined similar to the local martingale property. Obviously, any martingale is a local martingale.

For a local submartingale Y it is often possible to find an increasing process A such that the difference is a martingale. A condition for the existence of A is its predictability as defined next.

Definition B.1.7 *Predictable process*
Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $\mathcal{F}_t$ that satisfies the usual conditions. The σ–algebra on $[0,\infty) \times \Omega$ generated by all sets $[0] \times A, A \in \mathcal{F}_0$, and $(s,u] \times A$, $0 \leq s < u < \infty, A \in \mathcal{F}_s$, is called the *$\mathcal{F}_t$–predictable σ–algebra.*
A stochastic process Y on $(\Omega, \mathcal{F}, P)$ is called *$\mathcal{F}_t$–predictable* if, as a mapping from $[0,\infty) \times \Omega$ to $\mathbb{R}$, it is measurable w.r.t. the $\mathcal{F}_t$–predictable σ–algebra. //

Simpler characterizations of predictable processes are summarized in the following remark (Fleming and Harrington, 1991, p. 32).

Remark B.1.8 *Characterization of predictable processes*
(1) If Y is a $\mathcal{F}_t$–predictable process then $Y(t)$ is $\mathcal{F}_{t^-}$–measurable for all $t > 0$.
(2) If $\mathcal{F}_t$ is a filtration and Y a left–continuous real–valued process adapted to $\mathcal{F}_t$. Then, Y is $\mathcal{F}_t$–predictable (cf. also Brémaud, 1981, p. 9). //

Note that with the first part of the above remark we have for a predictable process $Y(t)$ that $E(Y(t)|\mathcal{F}_{t^-}) = Y(t)$ a.s.

B.2 The Doob–Meyer decomposition

Different formulations of the Doob–Meyer decomposition can be found in the literature. We mainly cite from Fleming and Harrington (1991), where most of the proofs can be found. Heuristically the decomposition implies under very general conditions that any process can be decomposed into a predictable part, which may be regarded as the predictor in a regression model, and a kind of error term given as martingale.

Theorem B.2.1 *Doob–Meyer decomposition*
Let $\{Y(t)|t \in \mathcal{T}\}$ be a nonnegative right–continuous local $\mathcal{F}_t$–submartingale on $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}_t$ is a filtration satisfying the usual conditions. Then, there exists a unique increasing right–continuous predictable process A such that $A(0) = 0$ a.s., $P(A(t) < \infty) = 1$ for all $t > 0$, and $Y - A$ is a right–continuous local martingale. The process A is called the *compensator* of Y.

Proof: Fleming and Harrington (1991, p. 58). □

Applied to counting processes, the Doob–Meyer decomposition reads as follows.

Theorem B.2.2 *Doob–Meyer decomposition for counting processes*
Let $\{N(t)|t \in \mathcal{T}\}$ be an arbitrary counting process on $(\Omega, \mathcal{F}, P)$. Then, there exists a unique right–continuous predictable increasing process Λ such that $\Lambda(0) = 0$ a.s. $P(\Lambda(t) < \infty) = 1$ for all $t > 0$, and $M = N - \Lambda$ is a right–continuous local $\mathcal{F}_t$–martingale, where $\mathcal{F}_t$ is the internal history of N.
The processes Λ is locally bounded, which implies that M is a local square integrable martingale. In addition, the jumps of the compensator are smaller than one, i.e. $\Delta\Lambda(t) = \Lambda(t) - \Lambda(t^-) \leq 1$ a.s. for all $t \geq 0$.

Proof: Fleming and Harrington (1991, p. 61). □

For the interpretation of the compensator it might be helpful to know that $E(N(t)) = E(\Lambda(t))$, i.e. the expected compensator at t is the expected number of events up to time t, and $\Lambda(dt) = E(N(dt)|\mathcal{F}_{t^-})$, i.e. the increment of the compensator at t is the probability of a jump at time t given the strict pre–t history (Fleming and Harrington, 1991, p. 62).

The next results are important for characterizing the relation between different counting processes.

Theorem B.2.3 *Predictable covariation process*
Let M_1 and M_2 be two local square integrable $\mathcal{F}_t$–martingales on $(\Omega, \mathcal{F}, P)$. Then, there exists a compensator of the process $M_1 M_2$ called the *predictable covariation process of* M_1 *and* M_2 denoted by $\langle M_1, M_2 \rangle$. //

Proof: Andersen et al. (1993, p. 68). □

For the special case of $M_1 = M_2 = M$ the process $\langle M, M \rangle = \langle M \rangle$ is called the *predictable variation process* and is the compensator of M^2. A heuristic interpretation of $\langle M_1, M_2 \rangle$ follows from

$$\langle M_1, M_2 \rangle(dt) = cov(M_1(dt), M_2(dt) | \mathcal{F}_{t^-}), \tag{B.1}$$

i.e. the increments of the covariance process correspond to the covariance of the increments of M_1 and M_2 given the strict pre–t history. The increments are uncorrelated if $\langle M_1, M_2 \rangle = 0$.

Definition B.2.4 *Orthogonal martingales*
Let M_1 and M_2 be two local square integrable $\mathcal{F}_t$–martingales on $(\Omega, \mathcal{F}, P)$. If $\langle M_1, M_2 \rangle = 0$ then M_1 and M_2 are said to be *orthogonal.* //

Applying the previous remark (B.1) we have that orthogonal martingales M_1 and M_2 are uncorrelated if $M_1(0)$ and $M_2(0)$ are uncorrelated.

If $N_1(t)$ and $N_2(t)$ are counting processes and $M_k = N_k - \Lambda_k$, $k = 1, 2$, are the martingales resulting from the Doob–Meyer decomposition then we have (Andersen et al., 1993, p. 74)

$$\begin{aligned} \langle M_k \rangle &= \Lambda_k - \int \Delta\Lambda_k \, d\Lambda_k, \quad k = 1, 2 \\ \langle M_1, M_2 \rangle &= -\int \Delta\Lambda_1 \, d\Lambda_2. \end{aligned}$$

The second equality implies that the martingales of any two different counting processes are orthogonal provided that the compensators are both continuous.

Further results as required e.g. for the proofs of the asymptotic distribution of the statistical procedures presented in Chapter 5 are not cited here. These would involve some deeper insight in the theory of stochastic integrals and martingale transforms which is beyond the scope of this section. The correponding background may again be found in the references cited above, i.e. Andersen et al. (1993), Fleming and Harrington (1991), and in particular Brémaud (1981).

Appendix C

Notations

In this part of the appendix, the symbols and notations which are repeatedly used in the previous chapters are summarized and briefly explained.

General

Random variables are mostly denoted by X, whereas Y stands for stochastic processes. They are defined on a probability space $(\Omega, \mathcal{F}, P)$. In the multivariate case bold types are used. In particular $\mathbf{X}_V = (X_1, \ldots, X_K)$ or $\mathbf{Y}_V = (Y_1, \ldots, Y_K)$ denote the situation, where the components are represented as vertices $k \in V$ in a graph with vertex set V. An upper index, e.g. X^i, indicates the sample unit and a lower index, e.g. X_k or $\mathbf{X}_A$, a component or subsets of the multivariate vector. Realizations of random variables are denoted by lower case letters, e.g. x or y. Further, we employ the following symbols:

$P(\cdot)$	probability function
$F(\cdot)$	cumulative distribution function
$f(\cdot)$	density
$S(\cdot)$	survival function
$E(\cdot)$	expectation
$Cov(\cdot)$	covariance
$\mathcal{F}$, $\mathcal{G}$	σ–fields
$\sigma(X)$	σ–field generated by X
$E(\cdot\|\mathcal{F})$	conditional expectation

$X \perp\!\!\!\perp Y \mid Z$	X is conditionally independent of Y given Z
$\mathbf{1}\{\cdot\}$	indicator function
$\delta_{x,y}$	Kronecker symbol, i.e. $\delta_{x,y} = \mathbf{1}\{x = y\}$
$\mathbf{I}$	identity matrix
$\mathbf{y}_{-k}$	$\mathbf{y}_{V \setminus \{k\}}$
$t \wedge s$	$\min(t, s)$
$t \vee s$	$\max(t, s)$
$\mathbb{R}$	real numbers
$\mathcal{N}(\mu, \Sigma)$	normal distribution with mean vector μ and covariance matrix Σ

Graphs

G	graph, $G = (V, E)$
V	set of vertices, mostly $V = \{1, \ldots, K\}$
$V_R \subset V$	set of vertices denoting the random variables
$V_P \subset V$	set of vertices denoting the processes
$j, k \in V$	vertices
$V_1, \ldots, V_J$	partition of V
E	set of edges
$(j, k) \in E$	directed edge from j to k
$\{j, k\} \in E$	undirected edge
$E^u(A)$	set of undirected edges on $A \subset V$, $E^u(A) = \{\{j, k\} \mid j, k \in A, j \neq k\}$
$E^d(A)$	set of directed edges on $A \subset V$, $E^d(A) = \{(j, k) \mid j, k \in A, j \neq k\}$
G_A	subgraph on $A \subset G$, $G_A = (A, E^A)$, $E^A = E \cap (E^d(A) \cup E^u(A))$
$\upsilon(j)$	undirected path component containing $j \in V$
$\Upsilon(G)$	set of all undirected path components in G
$\xi(j)$	cycle component containing $j \in V$
$\Xi(G)$	set of all cycle components in G
π	trail
DAG	directed acyclic graph
$\mathrm{pa}(A)$	parents of A in G
$\mathrm{ch}(A)$	children of A in G
$\mathrm{nb}(A)$	neighbors of A in G
$\mathrm{bd}(A)$	boundary of A in G, $\mathrm{bd}(A) = \mathrm{pa}(A) \cup \mathrm{nb}(A)$

$\mathrm{cl}(A)$	closure of A in G, $\mathrm{cl}(A) = \mathrm{bd}(A) \cup A$
$\mathrm{an}(A)$	ancestors of A in G
$\mathrm{de}(A)$	descendants of A in G
$\mathrm{nd}(A)$	non–descendants of A in G, $\mathrm{nd}(A) = V \backslash (\ \mathrm{de}(A) \cup A)$
$\mathrm{An}(A)$	smallest ancestral set containing A
$G^{\sim}$	undirected version of G
G^m	moralized version of G
G^B	graph G modified by deleting all directed edges starting in $B \subset V$
$\perp\!\!\!\perp_G$	graph separation
IR	irrelevance relation
IR_δ	δ–separation
$\wedge$	meet operation
$\vee$	join operation
F	intervention vertex in an intervention graph
ϕ	the value of F when it is idle

Stochastic processes

$\mathcal{T}$	time parameter, $\mathcal{T} = [0, \tau]$ or $\mathcal{T} = [0, \tau)$ with $0 < \tau \leq \infty$
$\{Y(t) \mid t \in \mathcal{T}\}$	stochastic process, briefly: Y
$Y(t^+)$	$\lim_{h \downarrow 0} Y(t+h)$
$Y(t^-)$	$\lim_{h \downarrow 0} Y(t-h)$
$\Delta Y(t)$	jump of Y, $\Delta Y(t) = Y(t) - Y(t^-)$
$Y^T(t)$	stopped process, i.e. $Y^T(t) = Y(T \wedge t)$
$\mathcal{S}$	state space
$q, r \in \mathcal{S}$	states
$\alpha(t)$	hazard rate or transition intensity
$\{\mathcal{F}_t \mid t \in \mathcal{T}\}$	filtration on $(\Omega, \mathcal{F}, P)$, briefly: $\mathcal{F}_t$
$\bigvee \mathcal{F}_t$	smallest σ–field generated by the arguments
$\mathcal{F}_{t^-}$	left limit of filtration, $\mathcal{F}_{t^-} = \bigvee_{h>0} \mathcal{F}_{t-h}$
$Y_j \not\to Y_k \mid \mathbf{Y}_{V \backslash \{j,k\}}$	Y_k locally independent of Y_j given the remaining components
$\{M(t) \mid t \in \mathcal{T}\}$	martingale, briefly: M
$\langle M \rangle$	predictable variation process of a martingale
$\langle M_1, M_2 \rangle$	predictable covariation process

Time series

VAR(M)	vector auto regressive model of M-th order
$f^*(\cdot)$	spectral matrix
Σ_ϵ	covariance matrix of residuals in VAR models
$Y_j \not\mapsto Y_k[\mathbf{Y}_V]$	noncausality
$Y_j \not\sim Y_k[\mathbf{Y}_V]$	instantaneous noncausality

Marked point processes (MPP)

$\mathcal{E} = \{e_1, \ldots, e_K\}$	mark space
$\{(T_s, E_s) \vert T_s \in \mathcal{T}, E_s \in \mathcal{E}\}$	marked point process
$\{N(t) \vert t \in \mathcal{T}\}$	counting process, briefly N
$\mathbf{N} = (N_1, \ldots, N_K)$	multivariate counting process
$N_k(t) = N(t; e_k)$	mark specific counting process
$\Lambda_k(t) = \Lambda(t; e_k)$	mark specific compensator
$\Lambda^C(t)$	continuous part of $\Lambda(t)$, i.e. $\Lambda(t) = \sum_{s \le t} \Delta\Lambda(s) + \Lambda^C(t)$
$\lambda_k(t) = \lambda(t; e_k)$	mark specific intensity process, $\Lambda(t) = \int \lambda(s)\, ds$
H_t	history process, $H_t = \{(T_s, E_s) \vert T_s \le t, E_s \in \mathcal{E}\}$
$\mathbb{H}$	set of all possible histories
$\mathcal{H}$	$\sigma(\mathbb{H})$
$\hat{\pi}(B)$	$P(H_t \in B \vert \hat{H}_t)$
$\mu_t(\tilde{B})$	prediction process $P(Y \in \tilde{B} \mid \mathcal{F}_t)$

Markov processes

$N_{qr}(t) = N(t; (q, r))$	counting process for transitions from state q to state r
$\Lambda_{qr}(t) = \Lambda(t; (q, r))$	compensator of $N_{qr}(t)$
$\alpha_{qr}(t) = \alpha(t; (q, r))$	transition intensity for transitions from state q to state r
$\mathbf{Y} \sim (Y_1, \ldots, Y_k)$	composable Markov process
CFMP	composable finite state Markov process
$N_k(t; (\mathbf{y}, y'_k))$	counting process for transitions of CFMP from state $\mathbf{y}$ to $\mathbf{y}'$, where $\mathbf{y}'$ differs from $\mathbf{y}$ only in the k-th component
$\alpha_k(t; (\mathbf{y}, y'_k))$	transition intensity for a CFMP

Likelihood

$L(t\|H_t)$	likelihood of a MPP based on the history H_t
$L_k(t\|H_t^{\mathrm{cl}(k)})$	mark specific likelihood
$T_{s(k)}$	times of occurrences of events e_k
$L(t\|H_t^{V\setminus A}\|\|\hat{H}_t^A)$	likelihood based on observing $H_t^{V\setminus A}$ and fixing $\hat{H}_t^A$

(Non/semi–) parametric models

$\boldsymbol{\theta}$	multidimensional parameter, mostly $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)$
Θ	parameter space
$\hat{\theta}$	estimator
θ^0	true value of parameter
$\mathcal{I}(\tau\|\theta)$	information matrix
$A_k(t)$	cumulative intensity for event e_k, i.e. $A_k(t) = \int \alpha_k(s)\, ds$
$\hat{A}_k(t)$	Nelson–Aalen estimator
$r(\cdot)$	relative risk
β	regression parameter
Z	predictable process independent of θ, mostly the number at risk for e_k

Bibliography

Aalen, O. O. (1978). Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics 6*, 701 – 726.

Aalen, O. O. (1987). Dynamic Modelling and Causality. *Scandinavian Actuarial Journal*, 177 – 190.

Aalen, O. O., Ø. Borgan, N. Keiding, and J. Thormann (1980). Interaction between Life History Events: Nonparametric Analysis of Prospective and Retrospective Data in the Presence of Censoring. *Scandinavian Journal of Statistics 7*, 161 – 171.

Aalen, O. O. and S. Johansen (1978). An Empirical Transition Matrix for Nonhomogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics 5*, 141 – 150.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1982). Linear Nonparametric Tests for Comparison of Counting Processes, with Application to Censored Survival Data (with Discussion). *International Statistical Review 50*, 219 – 258.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

Andersson, S. A., D. Madigan, and M. D. Perlman (2001). An Alternative Markov Property for Chain Graphs. *Scandinavian Journal of Statistics*, to appear.

Arjas, E. (1989). Survival Models and Martingale Dynamics (with Discussion). *Scandinavian Journal of Statistics 16*, 177 – 225.

Arjas, E. and M. Eerola (1992). On Predictive Causality in Longitudinal Studies. *Journal of Statistical Planing and Inference 34*, 361 – 386.

Arjas, E., P. Haara, and I. Norros (1992). Filtering the Histories of a Partially Observed Marked Point Process. *Stochastic Processes and Applications 40*, 225 – 250.

Bauer, H. (1991). *Wahrscheinlichkeitstheorie*. de Gruyter, Berlin.

Brémaud, P. (1981). *Point Processes and Queues: Martingale Dynamics*. Springer, New York.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. McGraw Hill, New York.

Caputo, A. (1998). *Eine alternative Familie von Modellverteilungen für Kovarianz– und Konzentrationsgraphen*. Ph. D. thesis, University of Munich.

Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Cox, D. R. (1972). Regression Models and Life–tables (with Discussion). *Journal of the Royal Statistical Society, Series B 34*, 187 – 220.

Cox, D. R. (1975). Partial Likelihood. *Biometrika 62*, 269 – 276.

Cox, D. R. and N. Wermuth (1996). *Multivariate Depencencies — Models, Analysis and Interpretation*. Chapman and Hall, London.

Cozman, F. G. and P. Walley (1999). Graphoid Properties of Epistemic Irrelevance and Independence. Technical report, University of Sao Paulo, Brasil.

Dahlhaus, R. (2000). Graphical Interaction Models for Multivariate Time Series. *Metrika 51*, 157 – 172.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B 41*, 1 – 31.

Dawid, A. P. (1980). Conditional Independence for Statistical Operations. *The Annals of Statistics 8*, 598 – 617.

Dawid, A. P. (1998). Conditional Independence. In S. Kotz, C. B. Read, and D. L. Banks (Eds.), *Encyclopedia of Statistical Sciences, Update Volume 2*, pp. 146 – 155. Wiley–Interscience, New York.

Dawid, A. P. (2000). Causal Inference without Counterfactuals. *Journal of the American Statistical Association 95*, 407 – 448.

Edwards, D. (2000). *Introduction to Graphical Modelling* (2 ed.). Springer, New York.

Edwards, D. and T. Havránek (1985). A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika 72*, 339 – 351.

Eerola, M. (1994). *Probabilistic Causality in Longitudinal Studies*, Volume 92 of *Lecture Notes in Statistics*. Springer, Berlin.

Eichler, M. (1999). *Graphical Models in Time Series Analysis*. Ph. D. thesis, University of Heidelberg.

Eichler, M. (2000). Causality Graphs for Stationary Time Series. Technical report, University of Heidelberg.

Fleming, T. R. and D. P. Harrington (1984). *Counting Processes and Survival Analysis*. Wiley, New York.

Florens, J. P. and D. Fougère (1996). Noncausality in Continuous Time. *Econometrica 64*, 1195 – 1212.

Frydenberg, M. (1990a). Marginalization and Collapsibility in Graphical Interaction Models. *The Annals of Statistics 18*, 790 – 805.

Frydenberg, M. (1990b). The Chain Graph Markov Property. *Scandinavian Journal of Statistics 17*, 333 – 353.

Frydenberg, M. and S. L. Lauritzen (1989). Decomposition of Maximum Likelihood in Mixed Graphical Interaction Models. *Biometrika 76*, 539 – 555.

Galles, D. and J. Pearl (1996). Axioms of Causal Relevance. Technical Report 240, University of California, Los Angeles.

Gänssler, P. and W. Stute (1977). *Wahrscheinlichkeitstheorie*. Springer, Berlin.

Gill, R. D. (1984). Understanding Cox's Regression Model: A Martingale Approach. *Journal of the American Statistical Association 79*, 441 – 447.

Gottard, A. (1998). *Analisi di Processi Stocastici interdependenti mediante Modelli Grafici di durata*. Ph. D. thesis, University of Florence.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross–spectral Methods. *Econometrica 37*, 424 – 438.

Heinicke, A. (1999). *Statistische Modellierung und Erfassung multivariater Assoziationsstrukturen anhand von KS–Verteilungen.* Logos, Berlin.

Klein, J. P., N. Keiding, and E. A. Copelan (1993). Plotting Summary Predictions in Multivariate Survival Models: Probabilities of Relapse and Death in Remission for Bone Marrow Transplantation Patients. *Statistics in Medicine 12*, 2315 – 2332.

Koehler, K. J. and J. T. Symanowski (1995). Constructing Multivariate Distributions with Specific Marginal Distributions. *Journal of Multivariate Analysis 55*, 261 – 282.

Koster, J. T. A. (1996). Markov Properties for Non–Recursive Causal Models. *Annals of Statistics 24*, 2148 – 2177.

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

Lauritzen, S. L. (2000). Causal Inference from Graphical Models. In O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg (Eds.), *Complex Stochastic Systems.* Chapman and Hall, London.

Lauritzen, S. L. and N. Wermuth (1989). Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative. *The Annals of Statistics 17*, 31 – 57.

Lynggaard, H. and K. H. Walther (1993). Dynamic Modelling with Mixed Association Models. Master's thesis, Aalborg University.

Norros, I. (1985). Systems weakened by Failures. *Stochastic Processes and Applications 20*, 181 – 196.

Parner, J. (1999). *Causal Reasoning and Time–dependent Confounding.* Ph. D. thesis, University of Copenhagen.

Parner, J. and E. Arjas (1999). Causal Reasoning from Longitudinal Data. Technical report, Rolf Nevanlinna Institute, University of Helsinki.

Parner, J. and N. Keiding (1998). Time–dependent Treatment Initiation and Informative Right–censoring. Technical Report 13, Department of Biostatistics, University of Copenhagen.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, San Mateo.

Pearl, J. (1993). Graphical Models, Causality and Intervention. *Statistical Science 8*, 266 – 269.

Pearl, J. (1995). Causal Diagramms for Empirical Research. *Biometrika 82*, 669 – 710.

Pearl, J. (2000). *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Pearl, J. and A. Paz (1987). Graphoids: A Graph–Based Logic for Reasoning about Relevancy Relations. In B. Du Boulay, D. Hogg, and L. Steel (Eds.), *Advances in Artificial Intelligence – II*, pp. 357 – 363.

Pearl, J. and J. M. Robins (1995). Probabilistic Evaluation of Sequential Plans from Causal Models with Hidden Variables. In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence*, Volume 11, pp. 444 – 453. Morgan Kaufmann, San Francisco.

Robins, J. M. (1986). A new Approach to Causal Inference in Mortality Studies with sustained Exposure Periods — Application to Control of the Healthy Worker Effect. *Mathematical Modelling 7*, 1393 – 1512.

Robins, J. M. (1989). The Control of Confounding by Intermediate Variables. *Statistics in Medicine 8*, 679 – 701.

Robins, J. M. (1998). Structural Nested Failure Time Models. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*, pp. 4372 – 4389. Wiley, Chichester.

Schweder, T. (1970). Composable Markov Processes. *Journal of Applied Probability 7*, 400 – 410.

Shao, J. (1999). *Mathematical Statistics*. Springer, New York.

Simpson, E. H. (1951). The interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Series B 13*, 238 – 241.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*, Volume 81 of *Lecture Notes in Statistics*. Springer, New York.

Tjøstheim, D. (1981). Granger–Causality in Multiple Time Series. *Journal of Econometrics 17*, 157 – 176.

Verma, T. and J. Pearl (1990). Causal Networks: Semantics and Expressiveness. In R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer (Eds.), *Uncertainty and Artificial Intelligence*, pp. 69 – 76. North Holland, Amsterdam.

Wermuth, N. and S. L. Lauritzen (1990). On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models (with Discussion). *Journal of the Royal Statistical Society, Series B 52*, 21 – 72.

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

Acknowledgements

First of all I would like to thank my supervisor Iris Pigeot. She supported and encouraged my work in several ways. In particular, she enabled me to get into contact with other researchers working on related topics which was essential for developing the basic ideas of this thesis. Further, I very much appreciate that she granted the freedom I needed while providing assistance whenever it was required. Without her advice this thesis would have probably become a 'never–ending story'.

Furthermore, I am in dept to many other persons who inspired and facilitated my work. Above all David Edwards who brought the potential of local independence to my attention, Sabine Loesgen who offered a 'cloistered' environment in Switzerland where I could complete the graph–theoretic proofs, and Christian Hennig who made a lot of valuable comments on preliminary versions of my thesis.

In addition, I am grateful for the pleasant atmosphere of work and social life that I shared with my colleagues. My sincere thanks for this to Angelika Caputo, Angelika Blauth, Eva–Maria Fronk, and Astrid Heinicke.

I further would like to thank my mother for teaching me how to work effectively and for her continuous confidence in my abilities, my father for stimulating my mathematical interests, and my grandparents who patiently accepted that I could not spend as much time with them as they deserved during the last four years.

Financial support by the SFB 386 of the Deutsche Forschungsgemeinschaft is also gratefully acknowledged.