Dynamic path analysis

A new approach to analyzing time-dependent covariates

Egil Ferkingstad
(egil.ferkingstad@medisin.uio.no)

Department of Biostatistics
University of Oslo

SAMSI Workshop: Large Graphical Models and Random Matrices
November 9th, 2006

Joint work with Odd Aalen, Ørnulf Borgan and Johan Fosen

Introduction

- In survival and event history analysis, graphical models have been mostly absent. Hence there has not been much focus on the structure among variables beyond that between covariates and the response in a regression analysis.
- One main disadvantage of traditional graphical modelling is that the role of time is not explicitly considered. Here, we will introduce a new approach to graphical models, where we focus on how events and processes in the past influence the development in the future. We call this new approach dynamic path analysis.

Introduction

We are going to study the situation where the main outcome is a stochastic process, and where we can model its compensator as a linear function of a set of covariates. We will define a path analysis model, i.e. a set of hierarchical linear regression models where some covariates in one regression model will be the response in another regression model, thus forming a graph of vertices (variables) and edges showing how all variables are related to each other. One of these linear models will be the regression of the increment of the stochastic process. The path model will be fitted at each time we are collecting information about the process.

Case study

- Randomized trial of survival for 488 patients with liver cirrhosis
- Randomized to placebo or treatment with prednisone (a hormone)
- Consider only the 386 patients without ascites (excess fluid in the abdomen).
- Treatment: 191 patients, 94 deaths
  Placebo: 195 patients, 117 deaths
- A number of covariates registered at entry
- Prothrombin index was also registered at follow-up visits throughout the study

Cox regression analysis

Covariate	Coding
Treatment	Placebo=0; Prednisone=1
Sex	Female=0; Male=1
Age	Years (range 27-77)
Acetylcholinesterase	Range 42-556
Inflammation	Absent=0; Present=1
Baseline prothrombin	Percent of normal
Current prothrombin	Percent of normal

Cox regression analysis

Covariate	Model I	Model II
Treatment	-0.28 (0.14)	-0.06 (0.14)
Sex	0.27 (0.16)	0.31 (0.15)
Age	0.041 (0.008)	0.043 (0.008)
Acetylcholinesterase	-0.0019 (0.0007)	-0.0015 (0.0006)
Inflammation	-0.47 (0.15)	-0.43 (0.15)
Baseline prothrombin	-0.0014 (0.007)	
Current prothrombin		-0.054 (0.004)

Cox regression analysis

Covariate	Model I	Model II
Treatment	-0.28 (0.14)	-0.06 (0.14)
Sex	0.27 (0.16)	0.31 (0.15)
Age	0.041 (0.008)	0.043 (0.008)
Acetylcholinesterase	-0.0019 (0.0007)	-0.0015 (0.0006)
Inflammation	-0.47 (0.15)	-0.43 (0.15)
Baseline prothrombin	-0.0014 (0.007)	
Current prothrombin		-0.054 (0.004)

Model I gives an estimate of the total treatment effect

Cox regression analysis

Covariate	**Model I**	**Model II**
Treatment	-0.28 (0.14)	-0.06 (0.14)
Sex	0.27 (0.16)	0.31 (0.15)
Age	0.041 (0.008)	0.043 (0.008)
Acetylcholinesterase	-0.0019 (0.0007)	-0.0015 (0.0006)
Inflammation	-0.47 (0.15)	-0.43 (0.15)
Baseline prothrombin	-0.0014 (0.007)	
Current prothrombin		-0.054 (0.004)

Model I gives an estimate of the total treatment effect

Model II shows the importance of prothrombin

Purpose

Get a better understanding of how treatment (and other fixed covariates) partly have a direct effect on survival and partly an indirect effect operating via the internal time-dependent covariate (current prothrombin).

This will be achieved by a combining classical path analysis with Aalen's additive regression model to obtain a dynamic path analysis for censored survival data.

Outline

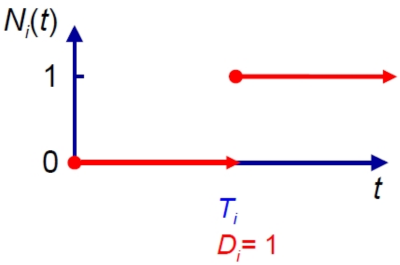
- Brief introduction to counting processes, intensity processes and martingales
- Brief review of Aalen's additive regression model for censored survival data
- Dynamic path analysis explained by means of the cirrhosis example
- Concluding comments

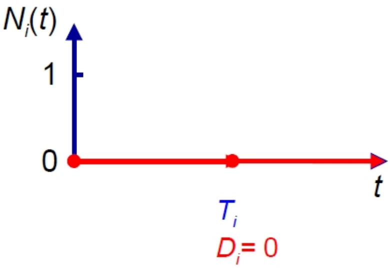
Counting processes

Have censored survival data $(\underbrace{T_i}_{\text{Censored survival data}}, \underbrace{D_i}_{\text{Status indicator (censored=0; failure=1)}})$

$N_i(t)$ counts the observed number of failures for individual i as a function of (study) time t.

Two possible outcomes for $N_i(t)$:





Intensity processes and martingales

- Denote by F_{t-} all information available to the researcher "just before" time t (on failures, censorings, covariates, etc.)
- The intensity process $\lambda_i(t)$ of $N_i(t)$ is given by

$$\lambda_i(t)\mathrm{d}t = P(\mathrm{d}N_i(t) = 1|F_{t-}) = E(\mathrm{d}N_i(t)|F_{t-})$$

  where $\mathrm{d}N_i(t)$ is the increment of N_i over $[t, t + \mathrm{d}t]$.
- Cumulative intensity process: $\Lambda_i(t) = \int_0^t \lambda_i(s)\mathrm{d}s$.

Intensity processes and martingales

- Introduce $M_i(t) = N_i(t) - \Lambda_i(t)$.

$$\begin{aligned} E(\mathrm{d}M_i(t)|F_{t-}) &= E(\mathrm{d}N_i(t) - \lambda_i(t)\mathrm{d}t|F_{t-}) \\ &= E(\mathrm{d}N_i(t)|F_{t-}) - \lambda_i(t)\mathrm{d}t \\ &= \lambda_i(t)\mathrm{d}t - \lambda_i(t)\mathrm{d}t \\ &= 0 \end{aligned}$$

$\Rightarrow M_i(t)$ is a martingale.

- Note that $\underbrace{\mathrm{d}N_i(t)}_{\textit{observation}} = \underbrace{\lambda_i(t)\mathrm{d}t}_{\textit{signal}} + \underbrace{\mathrm{d}M_i(t)}_{\textit{noise}}$.
- For statistical modeling we focus on $\lambda_i(t)$.

Aalen's additive regression model

- Intensity process for individual i:

$$\lambda_i(t) = \alpha_i(t) R_i(t),$$

  where $\alpha_i(t)$ is the hazard rate and $R_i(t)$ an at risk indicator.
- $x_{i1}(t), \ldots, x_{ip}(t)$: (possibly) time-dependent covariates for individual i (assumed predictable)
- Aalen's non-parametric additive model is given by

$$\alpha_i(t) = \beta_0(t) + \beta_1(t) x_{i1}(t) + \cdots + \beta_p(t) x_{ip}(t).$$

  Here, $\beta_0(t)$ is the baseline hazard, while $\beta_j(t)$ is the excess risk at t per unit increase of $x_{ij}(t)$ for $j = 1, \ldots, p$.

Aalen's additive regression model

It is difficult to estimate the $\beta_j(t)$ non-parametrically, so we focus on the cumulative regression functions:

$$B_j(t) = \int_0^t \beta_j(s)\mathrm{d}s$$

At each time s we have a linear model conditional on "the past" F_{s-} (with $x_{i0}(t) = 1$)

$$\underbrace{\mathrm{d}N_i(s)}_{\text{observation}} = \sum_{j=0}^{p} \underbrace{x_{ij}(s)Y_i(s)}_{\text{covariates}} \underbrace{\mathrm{d}B_j(s)}_{\text{parameters}} + \underbrace{\mathrm{d}M_i(s)}_{\text{noise}}.$$

Estimate the increments $\mathrm{d}B_j(s)$ by ordinary least squares at each time s when a failure occurs.

Aalen's additive regression model

- Estimate $B_j(t)$ by adding the estimated increments at all time s up to time t
- The vector of the $\hat{B}_j(t)$ is a multivariate "Nelson-Aalen type" estimator
- The statistical properties can be derived using results on counting processes, martingales, and stochastic integrals (see e.g. Andersen *et al.*, Springer, 1993).
- Software:
  - "aareg" in Splus (not in R)
  - "addreg" for Splus and R at www.med.uio.no/imb/addreg/
  - "aalen" for Splus and R at www.biostat.ku.dk/∼ts/timereg.html

Dynamic path modelling

- Dynamic path diagrams are defined in analogy with classical path diagrams
- A dynamic path diagram is a set of *time-indexed* directed acyclic graphs (DAGs) $G(t) = (V(t), E(t)),\ t \in [0, \infty)$, where $V(t)$ is the set of vertices and $E(t)$ the set of edges at time t.
- At any time t, $V(t)$ is partitioned into a *covariate set* $V_c(t) = \{X_1(t), \ldots, X_p(t)\}$ and an *outcome process* Y(t), i.e. $V(t) = V_c(t) \cup \{Y(t)\}$ where $Y(t) \notin V_c(t)$.
- The same $X_i(t) \in V(t)$ for all t
- The partition into covariate set and outcome process is the same for all t
- The set of included edges $E(t)$ may vary with time (edges may appear and disappear)

Dynamic path modelling

- We assume that, for all t,

$$E(t) \subseteq (V_c(t) \times V_c(t)) \cup (V_c(t) \times \{Y(t)\}).$$

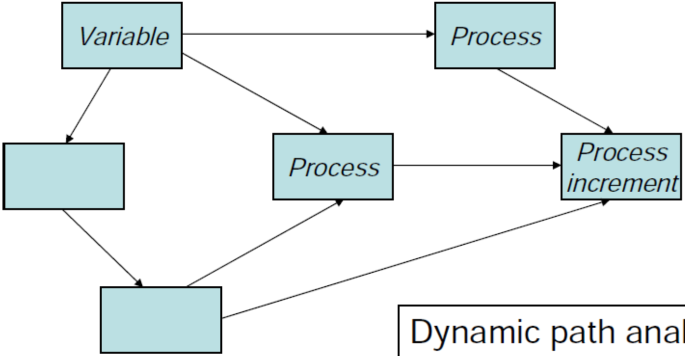
  This simply means that all edges are allowed, except edges pointing from the outcome to a covariate.

- We have a sequence of dynamic path models, one for each time t when we collect information.
- Estimation is done by recursive least squares regression, as usual in path analysis, where the increment of the outcome process $Y(t)$ is regressed onto its parents in the graph using the additive regression model (also a least squares method)
- The additivity of the Aalen regression model makes the dynamic path analysis possible.

Dynamic path modelling

- Associated with each edge is a regression coefficient:
  - for a regression of $X_j(t)$ onto $X_h(t)$ we have a (ordinary linear) regression coeffisient $\psi_{hj}(t)$ (corresponding to an edge $X_h(t) \rightarrow X_j(t)$ within the covariate set)
  - for a regression of $\mathrm{d}Y(t)$ onto $X_j(t)$ we have an (additive) regression coefficient $\mathrm{d}B_j(t)$ (corresponding to an edge $\mathrm{d}B_j(t) \rightarrow \mathrm{d}Y(t)$ from the covariate set to the outcome process)
- A direct effect corresponds to a path of length one in the graph
- An indirect effect corresponds to a path of length greater than one.
- Because of the linear structure of this model, the direct and indirect effects add up to the total effect. An indirect effect is simply the product of the (ordinary or additive) regression functions along the corresponding path.

Dynamic path diagram



Dynamic path analysis:
The boxes contain variables
or stochastic processes,
the final box always a process.
The graph coefficients are
functions of time.

Dynamic path modelling — cirrhosis example

Let's first only consider treatment and current prothrombin. We fit additive models:

1. with treatment as only covariate (marginal model):

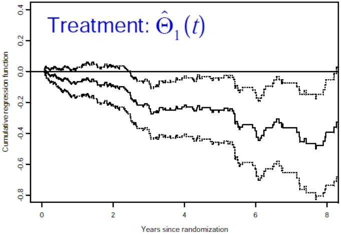
$$\begin{tikzcd} \text{Treatment} \arrow[r, "\mathrm{d}\Theta_1(t)"] & \mathrm{d}N_i(t) \end{tikzcd}$$

2. with treatment and current prothrombin (dynamic model):

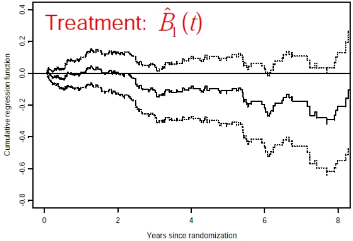
$$\begin{tikzcd} \text{Treatment} \arrow[rr, "\mathrm{d}B_1(t)"] & & \mathrm{d}N_i(t) \\ & \text{Current prothrombin} \arrow[ur, "\mathrm{d}B_2(t)"] & \end{tikzcd}$$

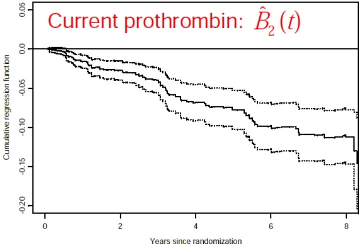
Dynamic path modelling — cirrhosis example

(i) marginal model:



(ii) dynamic model:



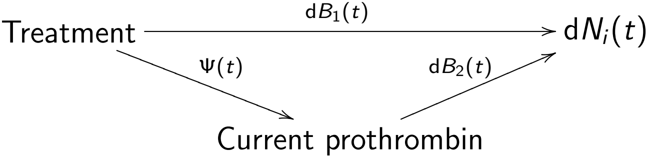


Total effect of treatment is underestimated in the dynamic model

Current prothrombin has a strong effect on mortality

Dynamic path modelling — cirrhosis example

Using dynamic path analysis we see how the two analyses fit together:



Treatment

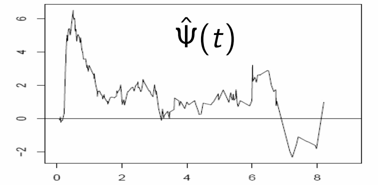
$dB_1(t)$

$dN_i(t)$

$\Psi(t)$

$dB_2(t)$

Current prothrombin

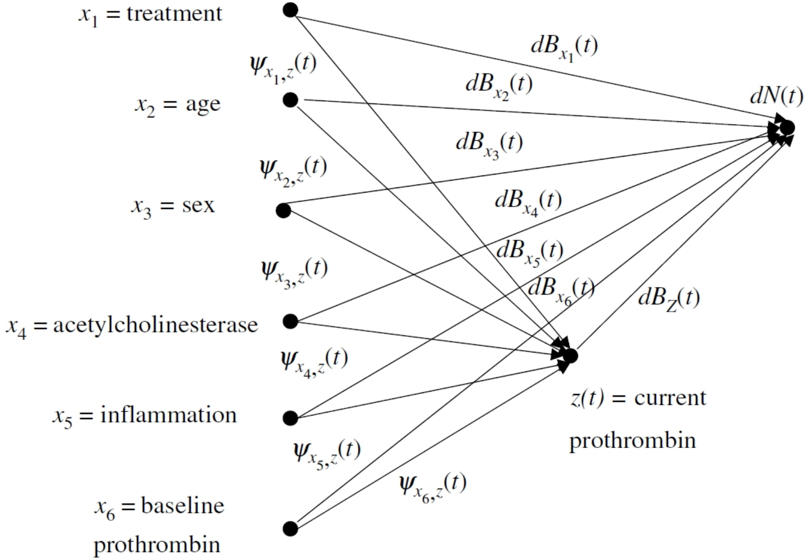


Treatment increases prothrombin, and high prothrombin reduces mortality. Part of the treatment effect is *mediated* through prothrombin.

Due to additivity and least squares estimation:

$$\underbrace{d\hat{\Theta}_1(t)}_{\text{total effect}} = \underbrace{d\hat{B}_1(t)}_{\text{direct effect}} + \underbrace{\hat{\Psi}(t)\cdot d\hat{B}_2(t)}_{\text{indirect effect}}$$

Dynamic path diagram — cirrhosis example



DAG of one possible model of the liver cirrhosis data

Model selection

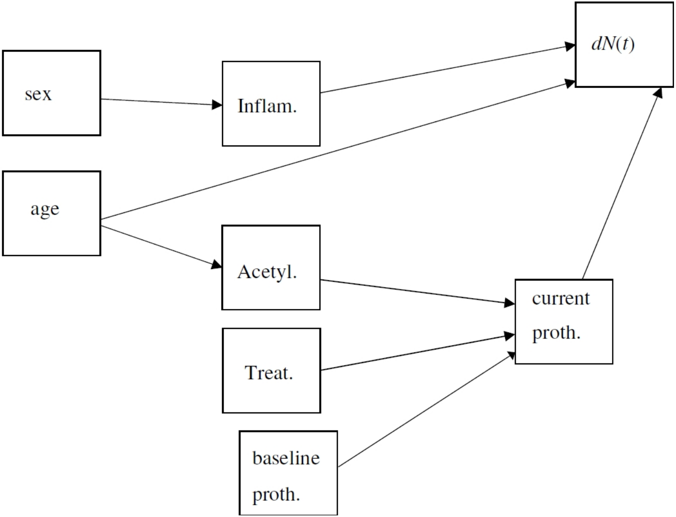
Model selection for dynamic path analysis is quite complicated:

- Need to decide the maximal set E of edges to include in a model
- Need to decide which of the edges in E that should be present at the various points in time.

Ad hoc model selection procedure for the cirrhosis data:

- Include edge from a covariate to the outcome process $N(t)$ if it has a significant direct effect using the common test for effects in the additive hazards model (Aalen 1989)
- Include edge between covariates if it is significant for most of the event times in an interval of at least half a year, based on Wald test statistics from ordinary linear regression.
- Use bootstrapping for validating the model choice: if a bootstrap confidence interval contains zero, then the covariate is insignificant.

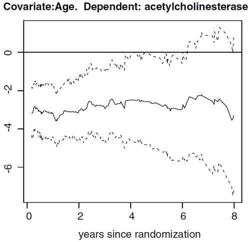
Dynamic path diagram — cirrhosis example

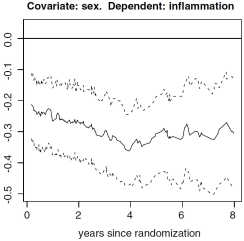


The path model best fitting the data

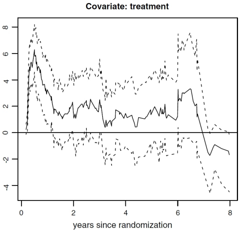
Results from cirrhosis example

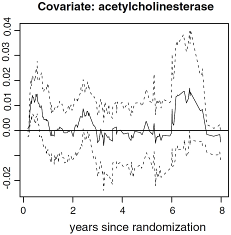
Direct effects on acetyl and inflammation (ordinary least squares):

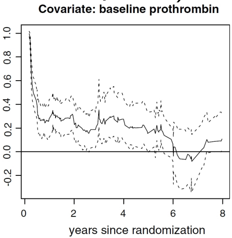




Direct effects on current prothrombin (ordinary least squares):

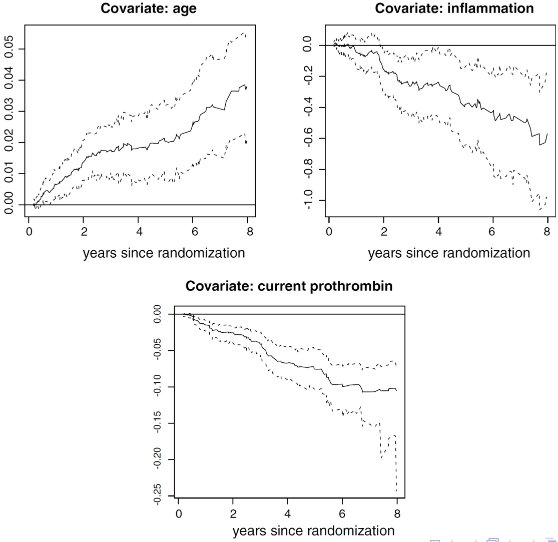






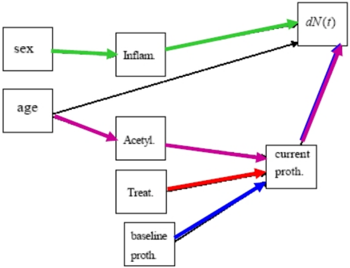
Results from cirrhosis example

Direct cumulative effects on death (additive regression):

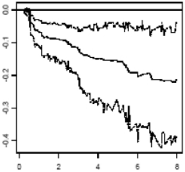


Results from cirrhosis example

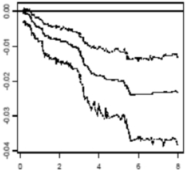
Indirect cumulative effects on death:



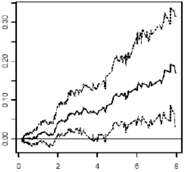
Treatment



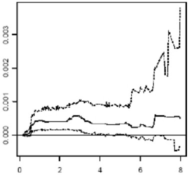
Baseline prothrombin



Sex



Age



General conclusions

- Aalen's additive regression model is a useful supplement to Cox's regression model
- Additivity and least squares estimation make dynamic path analysis feasible, including the concepts direct, indirect and total treatment effects
- Dynamic path analysis may be extended to recurrent event data with e.g. the previous number of events as an internal time-dependent covariate
- Much methodological work remains to be done on dynamic path analysis, e.g. on methods for model selection