Shared Frailty Graphical Survival Models

Anna Gottard and Carla Rampichini

Department of Statistics "Giuseppe Parenti"
University of Florence

International Conference on
STATISTICAL LATENT VARIABLES MODELS IN THE HEALTH SCIENCES
Perugia, Italy, 6-8 September 2006
Bolivia – Poorest Country in South America

Bolivia is characterized by
- High child abandonment rate
- High infant mortality rate

We want to study the relations between
- poverty status and fertility in young women
- infant mortality and characteristics of the mother

- In both models: survival times, explanatory r.v. and clustered unobserved heterogeneity
- Interest: the dependence structure of all the random quantities involved
Ingredients

- **Shared frailty survival models**
  - useful for unobserved heterogeneity shared by clusters of individuals
  - can be viewed as multilevel models

- **Graphical models**
  - useful whenever the interest is on the dependence structure of a multivariate set of random quantities
  - study the joint distribution of involved variables
  - provide intuitive understanding of the complex association structure
  - help to communicate

- **Multilevel graphical models** (Gottard & Rampichini, to appear)
  - chain graphs when data have a multilevel structure

- **Graphical duration models** (Gottard, 2002)
  - include nodes for marked point processes as terminal nodes of a chain graphs
Shared frailty survival graphical models can be viewed as the intersection of multilevel graphical models with duration graphical models.

- useful whenever the interest is on the dependence structure of a multivariate set of random quantities, including survival times and random variables
- provide intuitive understanding of the complex association structure of clustered survival data
- study the joint distribution of involved variables with survival times
- help to communicate assumptions and results
Single–Point Process - Survival models

$Y(t)$: single–point process, takes values on $S_Y = \{0, 1\}$, where

- $0$ is a transient state, with $Y(t = 0) = 0$
- $1$ is an absorbing state.

Event/failure = transition between states.
$T$: time–to–event = duration = survival time \[ T \in (0, \tau] \]
Survival models

⇒ Distribution function (assumed absolutely continuous):

\[ F(t) = P(T \leq t) \]

⇒ Survivor function: probability that the event occurs after \( t \):

\[ S(t) = P(T > t) = 1 - F(t) \]

⇒ The hazard function is the instantaneous rate of an event in \([t, t + dt)\):

\[ h(t) = \frac{dF(t)}{S(t)} = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid T \geq t)}{dt} \]

⇒ Cox proportional hazard model (Cox, 1972)

\[ h(t \mid X) = h_0(t) \exp(\beta_0 + \beta X) \]
Frailty models

\[\Rightarrow\text{SOMETIMES: observed explanatory variables/processes cannot explain ALL the variability in the observed time to event.}\]

\[\Rightarrow\text{UNEXPLAINED HETEROGENEITY can be caused by unobserved relevant risk factors.}\]

\[\Downarrow\]

FRAILTY

\[\Rightarrow\text{Unexplained heterogeneity can be SHARED among individuals \Rightarrow frailty is common to several individuals, forming a group.}\]

\[\Downarrow\]

SHARED FRAILTY
Cox shared frailty model & conditional independence

- Cox proportional hazard model (Cox, 1972) with shared Gamma frailty (Clayton, 1978)

\[ h_{ij}(t \mid U_j, X) = h_0(t) \cdot U_j \exp(\beta_0 + \beta X) \]

\[ U_j \sim \text{Gamma}, \text{ with } E(U_j) = 1 \text{ and } V(U_j) = \theta \]

- Units in the same cluster share the same frailty \( U_j \)

  - units of the same cluster \( j, \forall i \neq i', \; i, i' = 1, 2, \ldots, n_j \):

    \[ Y_{ij}(t) \perp \perp Y_{ij'}(t) \quad Y_{ij}(t) \perp \perp Y_{ij'}(t) \mid X \quad Y_{ij}(t) \perp \perp Y_{ij'}(t) \mid X, U_j \]

  - units in different clusters, \( \forall j \neq j', \; j, j' = 1, 2, \ldots, J \)

    \[ Y_{ij}(t) \perp \perp Y_{ij'}(t) \quad Y_{ij}(t) \perp \perp Y_{ij'}(t) \mid X \quad Y_{ij}(t) \perp \perp Y_{ij'}(t) \mid X, U_j \]
Graphical Survival models with shared frailty

Shared frailty models belong to the class of multilevel models

Instead, **Graphical multilevel models** represent in the graph the set of r.v.

\[(Y_j, U_j, X_j) \Rightarrow (Y_j(t), U_j, X_j)\]
Graphical Survival models with shared frailty (cont’d)

- to depict \((Y_j(t), U_j, X_j)\) define

  - **Individual node** representing a random variable of a specific statistical unit.
  - **Grouping latent node** representing an unobserved r.v. \(U_j\) such that, for all \(Y_{ij}, Y_{i'j} \in ch(U_j)\),

\[
Y_{ij} \perp \perp Y_{i'j} \mid U_j, X \quad \text{and} \quad Y_{ij} \perp \perp Y_{i'j} \mid X
\]

- The **pure response variable** is actually a single–point process \(Y(t)\) as in graphical duration models

  ⇒ Graphical duration models: \(Y(t)\) can included in a graph by a **single node**.
Random intercept survival chain graph

\[ Y_{1j}(t) \perp \perp Y_{2j}(t) \mid X, U_j \]
1. Fertility in Bolivia

**Main goal:** to study *time to first childbirth* together with poverty indicators, educational level and area of residence of the woman.

- Data from the Bolivia DHSIII - 3rd Demographic and Health Survey (1998)- on 11,187 women leaving in Bolivia
- Sub-sample of women aged 14-35 years (7777 women)
- Variable of interest: time to first childbirth (4420 events)
- Explanatory variables: indicators of poverty, family status, individual characteristics
- Clusters: area of residence (33 areas, $n_j \min=19$, $\max=622$, average=236). Areas are characterized by different ethnic groups (not measured)
Fertility in Bolivia: variables and blocks ordering

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pure explanatory</td>
<td>( X_1 ) House facilities |</td>
</tr>
<tr>
<td></td>
<td>( X_{11} ) Roof (ref=other, poor 22%) |</td>
</tr>
<tr>
<td></td>
<td>( X_{12} ) Floor (ref=cement, of earth 37%, rich 8%) |</td>
</tr>
<tr>
<td></td>
<td>( X_{13} ) Electricity (ref=no, yes 66%) |</td>
</tr>
<tr>
<td>2. Intermediate</td>
<td>( X_2 ) N. of years spent at school (0-19, mean=6.57) |</td>
</tr>
<tr>
<td>3. Grouping latent node</td>
<td>( U_j ) area of residence |</td>
</tr>
<tr>
<td>4. Pure response</td>
<td>( Y(t) ) time to first childbirth |</td>
</tr>
</tbody>
</table>

**Factorization:**

\[
\prod_{i=1}^{n_j} f_{ij}^Y(t \mid U_j, X_1, X_2) f(U_j) f(X_2 \mid X_1) f(X_{11}, X_{12}, X_{13})
\]
The estimation of the chain graph model requires many steps.

The estimation strategy is based on univariate appropriate regression models according to the dependent variable scale and to the recursive nature of the chain graph (Cox and Wermuth, 1996).

The shared frailty model is fitted by means of the \texttt{stcox} procedure of \textsc{Stata}, with a Gamma frailty.
Fertility in Bolivia: the resulting graph

House facilities: $X_{11} = $ Poor roof $X_{12} = $ Floor $X_{13} = $ Electricity

$X_2 = $ Education (years)
Estimated hazard functions: different house facilities

Hazard functions: House Facilities rich vs poor

- Average educational level (6.57 years)
- **Poor HF**: poor roof=1, floor of earth=1, electricity=0
- **Rich HF**: rich floor=1, electricity=1

‘Poor HF’ women show:
- a higher hazard rate at each age
- first childbirth earlier
- higher difference in central age
Estimated hazard functions: different $U_j$ values

Hazard for rich house facilities

Hazard for poor house facilities

★ Differences among areas are greater for central ages

1. high hazard group $u = 1 + \hat{\theta}$
2. average group $u = 1$
3. low hazard group $u = 1 - \hat{\theta}$
Infant mortality in Bolivia

- Data: Bolivia DHSIII - 3rd (1998)
- Women aged 40 and over of La Paz, north Bolivia (394 women)
- Pure response point process: children survival times
- 1990 children, 293 deaths: infant mortality rate = 14.72%
- Clusters: children of the same mother ($n_j$ min=1, max=13, average=5)
- Explanatory variables only at cluster level (mother)
Infant mortality: variables and blocks ordering

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pure explanatory</td>
<td>$Z_1$</td>
</tr>
<tr>
<td></td>
<td>$Z_{11}$ mother lives in the countryside (ref=no, yes 35%)</td>
</tr>
<tr>
<td></td>
<td>$Z_{12}$ Floor (ref=cement, poor (of earth) 32%, rich 1%)</td>
</tr>
<tr>
<td>2. Intermediate</td>
<td>$Z_2$ N. of years spent at school (0-19, mean=4.78)</td>
</tr>
<tr>
<td>3. Intermediate</td>
<td>$X(t)$ age at the 1th childbirth (13-42, mean=22.6)</td>
</tr>
<tr>
<td>single point process</td>
<td></td>
</tr>
<tr>
<td>3. Grouping latent node</td>
<td>$U_j$ mother ID</td>
</tr>
<tr>
<td>4. Pure response</td>
<td>$Y(t)$ time to child death</td>
</tr>
</tbody>
</table>

NB: survival times of children of the same mother are viewed as repeated measures!
Infant mortality: the resulting graph

Pure explanatory variables:

- $Z_{11} =$ resident in countryside  \hspace{1cm} Z_{12} =$ Floor
- $Z_2 =$ Education (years)
- $X(t) =$ age at 1$^{st}$ childbirth
Hazard functions: women with 1\textsuperscript{st} childbirth=22.6 years

Women by years of education

Average educated women (4.8 years) for different $U_j$ values

1. none 0 years
2. average 4.8 years
3. high 10 years

$\star$ Differences among women are greater for lower values of \textit{time} (age at death in month) 
$(\theta = 0.31$, LRT for $\theta = 0: \chi_{(01)} = 10.91$, $Prob \geq \chi_{(01)} = 0.000)$
Concluding remarks and future work

- The analysis is at an initial stage:
  - quality of the data
  - more socio-demographic indicators and other variables
  - test the proportional assumption

- Unobserved heterogeneity:
  - constant (at subject and/or group level) vs varying, i.e. $U_j(t)$ as a point process at group level
  - also intermediate variables can have unobserved heterogeneity subject and/or group level

- Further investigate graphical properties