# Estimation and Model Selection for the Competing Risks Model with Masked Causes of Failure

Radu V. Craiu

Department of Statistics

University of Toronto

Collaborators: Thierry Duchesne (Laval), Thomas Lee (Colorado State).

## Talk's Outline

**Part 1** Inference for the competing risks model with masked failure causes.

1. Competing risks with masking.
2. Data and model.
3. EM algorithm - convergence, variance estimation.
4. Particular hypothesis of interest: proportionality, symmetry, time-varying masking probabilities.
5. Example and Robustness study.

**Part II** Model selection.

1. AIC, BIC, MDL for the competing risks model.
2. Simulation study.
3. Conclusions and future work.

## Competing risks

• Competing risk framework: items (individuals) may fail from one of $J$ causes;

• Normally, for each item we observe a failure or censoring time, and a unique cause of failure if item not censored;

• Under masking, cause of failure not determined uniquely for some of the items; all we know is that it belongs to one of the $G$ **proper masking groups.**

• Some items with a masked failure cause go to a 2nd stage analysis where unique failure cause is determined.

## Data and Notation

Suppose there are three possible causes of failure ($J$=3), and two **proper** masking groups $g_4 = \{1,2\}$ and $g_5 = \{1,2,3\}$. We denote $g_j = \{j\}$ for all $j = 1,2,3$. Denote $M = J+G = 5$

| Item no. | Time of failure/censoring | Cause of failure | Masking group | Censoring indicator |
|---|---|---|---|---|
| 1 | 0.0183 | 1 | 1 | 0 |
| 2 | 0.0427 | -1 | 4 | 0 |
| 3 | 0.0735 | -1 | 5 | 0 |
| 4 | 0.171 | 1 | 4 | 0 |
| 5 | 0.231 | - | - | 1 |
| 6 | 0.604 | 3 | 5 | 0 |

• To each item $1 \leq i \leq N$ we can associate the vector $(\delta_{i1}, \delta_{i2}, \delta_{i3})$ in which $\delta_{ij} = 1$ if item $i$ has failed because of cause $j$, and $\delta_{ij} = 0$ otherwise $\forall j = 1,2,3$.

• To each item $i$ we can associate the vector $\gamma_{i1}, \gamma_{i2}, ..., \gamma_{i5}$ of 0-1 variables that indicate the group masking the failure.

• A censored item will have all $\delta$'s equal to zero.

| Item no. | Time of failure/censoring | Cause of failure | Masking group | Censoring indicator |
|---|---|---|---|---|
| 1 | 0.0183 | 1 | 1 | 0 |
| 2 | 0.0427 | -1 | 4 | 0 |
| 3 | 0.0735 | -1 | 5 | 0 |
| 4 | 0.171 | 1 | 4 | 0 |
| 5 | 0.231 | - | - | 1 |
| 6 | 0.604 | 3 | 5 | 0 |

The above translates into:

| Item | Time | $\delta_{i1}$ | $\delta_{i2}$ | $\delta_{i3}$ | $\gamma_{ig_1}$ | $\gamma_{ig_2}$ | $\gamma_{ig_3}$ | $\gamma_{ig_4}$ | $\gamma_{ig_5}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0183 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0.0427 | - | - | 0 | 0 | 0 | 0 | 1 | 0 |
| 3 | 0.0735 | - | - | - | 0 | 0 | 0 | 0 | 1 |
| 4 | 0.171 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0.231 | 0 | 0 | 0 | 0 | 0 | 0 | - | - |
| 6 | 0.604 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |

- Complete data $(t_i, \delta_{i1}, \ldots, \delta_{iJ}, \gamma_{i1}, \ldots, \gamma_{iM}), \quad i = 1, \ldots, n.$

- Incomplete data: $\delta_{ij}$'s will be missing for some of the items.

- For right-censored items, we set all the $\delta_{ij}$'s equal to 0. For these items, the likelihood does not depend on the $\gamma_{ig}$'s.

## Complete data and Missing mechanism

**Observed data:** What we actually observe, i.e., stage 1 data for all the items, stage 2 data only for some items.

**Complete data:** What we would actually observe if every single item with a masked failure cause in stage 1 went on to a 2nd stage analysis.

**Missing at random** P(item $i$ is masked|OBS) does not depend on the missing $\delta_{ij}$.

## Model for the hazards

• Let $T$ be the r.v. of time to failure and $C$ be the r.v. of cause of failure.

• Failure can be due to only one cause at a time.

• The cause-specific hazard function for cause $j$ is

$$\lambda_j(t) = \lim_{h \downarrow 0} \frac{\mathsf{P}[t \leq T \leq t+h, C = j | T \geq t]}{h}.$$

$$S(t) = \mathsf{P}[T > t] = \exp\left\{ - \int_0^t \sum_{j=1}^{J} \lambda_j(u)\, du \right\}.$$

Use

$$\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} 1_k(t),$$

where $0 = a_0 < a_1 < \cdots < a_K = \infty$ and $1_k(t)$ is the indicator that $t \in (a_{k-1}, a_k]$.

• Weak parametrization that is flexible and mathematically convenient.

• Likelihood-based testing and estimation is possible.

## Masking Parameters

- Define the **masking probabilities**

$$\mathsf{P}_{g|j} = \mathsf{P}[\text{failure cause masked in } g | C = j].$$

- What we really want

1. The **diagnostic probabilities**

$$\pi_{j|g}(t) = \mathsf{P}[C = j | T = t, \text{ masked in } g].$$

A simple application of Bayes' rule yields

$$\pi_{j|g}(t) = \frac{\lambda_j(t)\mathsf{P}_{g|j}}{\sum_{l \in g} \lambda_l(t)\mathsf{P}_{g|l}}.$$

2. The **cumulative incidence functions**

$$F_j(t) = \mathsf{P}(T \leq t, C = j) = \int_0^t \lambda_j(u)S(u)du.$$

## Likelihood complete data

Let $\theta$ be vector containing the $\lambda_{jk}$'s and $\mathsf{P}_{g|j}$'s. **Goal:** Inference about $\theta$.

Let $\mathcal{G}_j = \{g : j \in g\}$ and $\mathcal{G}_j^* = \mathcal{G}_j / \{j\}$.

Log-likelihood under complete data:

$$l_C(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left\{ \left[ \delta_{ij} \ln \sum_{k=1}^{K} \lambda_{jk} 1_k(t_i) - \sum_{k=1}^{K} \lambda_{jk} \int_0^{t_i} 1_k(u)\, du \right] \right\}$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_{ij} \left[ (1 - \sum_{g \in \mathcal{G}_j^*} \gamma_{ig}) \ln(1 - \sum_{g \in \mathcal{G}_j^*} \mathsf{P}_{g|j}) + \sum_{g \in \mathcal{G}_j^*} \gamma_{ig} \ln \mathsf{P}_{g|j} \right].$$

- **ML estimators**

$$\widehat{\lambda}_{jk} = \frac{\sum_{i=1}^{n} \delta_{ij}\ 1_k(t_i)}{e_k}$$

and

$$\widehat{\mathsf{P}}_{g|j}^{(l)} = \frac{\sum_{i=1}^{n} \delta_{ij}\ \gamma_{ig}}{\sum_{i=1}^{n} \delta_{ij}},$$

where $e_k = \sum_{i=1}^{n} \int_0^{t_i} 1_k(u)\, du$.

## Likelihood missing data

- $J = 3$, $g = \{1, 2, 3\}$. item $i$ is masked in $g$ then

$$(\delta_{i1}, \delta_{i2}, \delta_{i3}) \sim \text{Multin}(1, \pi_{1|g}, \pi_{2|g}, \pi_{3|g}).$$

Let $\mathcal{M}$ denote set of $i$'s for which we have missing data. Define for an item $i \in \mathcal{M}$, $g_i$ the masking group for $i$. The log-likelihood of missing data given observed data is

$$l_{\mathcal{M}|OBS}(\theta) =$$
$$\sum_{i \in \mathcal{M}} \left\{ \sum_{j \in g_i^*} \delta_{ij} \ln \pi_{j|g_i}(t_i) + (1 - \sum_{j \in g_i^*} \delta_{ij}) \ln(1 - \sum_{j \in g_i^*} \pi_{j|g_i}(t_i)) \right\}$$

where $g_i^*$ denotes all the causes but one in masking group $g_i$ (e.g. if $g_i = \{1, 2, 3\}$, then $g_i^*$ could be any of $\{1, 2\}$, $\{1, 3\}$ or $\{2, 3\}$).

## Observed data likelihood

$$
\begin{aligned}
& l_{OBS}(\theta) = E[l_C(\theta)|OBS] - E[l_{\mathcal{M}}(\theta)|OBS] \\
= & \sum_{i=1}^{n}\sum_{j=1}^{J}\left\{\left[E[\delta_{ij}|OBS]\,\ln\sum_{k=1}^{K}\lambda_{jk}1_k(t_i) - \sum_{k=1}^{K}\lambda_{jk}\int_0^{t_i}1_k(u)\,du\right]\right. \\
+ & \left. E[\delta_{ij}|OBS]\left[(1-\sum_{g\in\mathcal{G}_j^*}\gamma_{ig})\ln(1-\sum_{g\in\mathcal{G}_j^*}\mathsf{P}_{g|j}) + \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln\mathsf{P}_{g|j}\right]\right\} \\
& - \sum_{i\in\mathcal{M}}\sum_{j\in g_i}E[\delta_{ij}|OBS]\ln\pi_{j|g_i}(t_i),
\end{aligned}
$$

## EM Algorithm

To obtain $\widehat{\theta}$ maximizing $l_{OBS}(\theta)$, we use the EM algorithm:

**Initial step:** Set

$$\widehat{\lambda}_{jk}^{(0)} = \frac{\sum_{i=1}^{n} 1[\delta_{ij} \text{ observed and equal to } 1]}{e_k},$$

and $\widehat{\mathsf{P}}_{g|j}^{(0)} = 1/\#\mathcal{G}_j$.

• **E-step:** Compute $E_{\widehat{\theta}^{(n-1)}}[\delta_{ij}|OBS]$ using

$$E_{\theta'}[\delta_{ij}|OBS] = \frac{\lambda'_j(t_i)\mathsf{P}'_{g_i|j}}{\sum_{l\in g_i}\lambda'_l(t_i)\mathsf{P}'_{g_i|l}} = \pi'_{j|g_i}(t_i)$$

if cause of item $i$ is masked in $g_i$ and no second stage. Otherwise $\delta_{ij} \in \{0, 1\}$.

• **M-step:** Set

$$\widehat{\lambda}_{jk}^{(l)} = \frac{\sum_{i=1}^{n} E_{\widehat{\theta}^{(l-1)}}[\delta_{ij}|OBS]\ 1_k(t_i)}{e_k}$$

and

$$\widehat{\mathsf{P}}_{g|j}^{(l)} = \frac{\sum_{i=1}^{n} E_{\widehat{\theta}^{(l-1)}}[\delta_{ij}|OBS]\ \gamma_{ig}}{\sum_{i=1}^{n} E_{\widehat{\theta}^{(l-1)}}[\delta_{ij}|OBS]}.$$

## Convergence and asymptotic variance

**Assumption :** The division of the time line into $K$ intervals $(a_0, a_1]$, ..., $(a_{K-1}, a_K)$ is such that, in each interval and for each cause $j \in \{1, \ldots, J\}$ we have at least one failure which is masked in a group that contains $j$.

$\Rightarrow$ The algorithm will converge to a stationary point (Wu, Ann. Stat. 1983, Theorem 2).

$\Rightarrow$ The algorithm will converge even if no items are sent to the second stage as long as the hazards are not proportional.

We use SEM (Meng and Rubin, JASA 1991) to calculate the asymptotic variance of the estimators found.

## Hypotheses of interest

• Assuming **proportional hazards** and with second stage data Flehinger et al. (Biomtrka, 1998) developed nonparametric methods for estimation.

• Under proportional hazards Goetghebeur and Ryan (Biomtrka, 1995) use estimating equations for consistent estimation.

• Under no second stage data the parameters are unidentifiable if **hazards are proportional**, basically due to an overparametrization.

• Our model will detect unidentifiability because the EM will be extremely unstable.

• We can bypass unidentifiability by assuming different end points (across causes) for the intervals $[a_{k-1}, a_k]$.

• Previous analyses have used the **symmetry assumption** $\mathrm{P}_{g|j} = \mathrm{P}_g$ to avoid this problem (Dinse, JASA 1986).

• However, no testing for proportional hazards and symmetry are proposed. **Until now!**

## Likelihood Ratio Tests

Let $\widehat{\theta}$, $\widehat{\theta}_{PH}$ and $\widehat{\theta}_{SYM}$ denote the MLEs under the general model, the proportional hazards assumption (PH) and the symmetry assumption (SYM), respectively.

We can perform likelihood ratio tests of the PH and SYM hypotheses:

For $H_0: \ \lambda_{jk} = \phi_j \lambda_{1k}$, we compute

$$r = 2[l_{OBS}(\widehat{\theta}) - l_{OBS}(\widehat{\theta}_{PH})]$$

and reject $H_0$ at level $\alpha$ if $r \geq \chi^2_{\alpha;(J-1)(K-1)}$.

For $H_0: \ \mathrm{P}_{g|j} = \mathrm{P}_g$, we compute

$$r = 2[l_{OBS}(\widehat{\theta}) - l_{OBS}(\widehat{\theta}_{SYM})]$$

and reject $H_0$ at level $\alpha$ if $r \geq \chi^2_{\alpha;\sum_{k=1}^{G} \#g_k - G}$. where $G$ is the number of proper masking groups and # denotes cardinality.

## Reliability of hard drives

Flehinger et al. (Lifetime Data Anal., 2002) present a dataset related to the reliability of hard drives. They follow 1000 items for four years; during that period they observe 172 failures, 66 of which are masked.

Possible failure causes: 1, 2 and 3.

Observed proper masking groups: $\{1,3\}$ and $\{1,2,3\}$.

Flehinger et al. construct the observed likelihood directly assuming Weibull cause-specific hazards.

Using our method, we can assess the fit of the Weibull distribution, and test the symmetry and proportional hazards assumptions.

Estimates of the $P_{g|j}$'s:

Our estimates

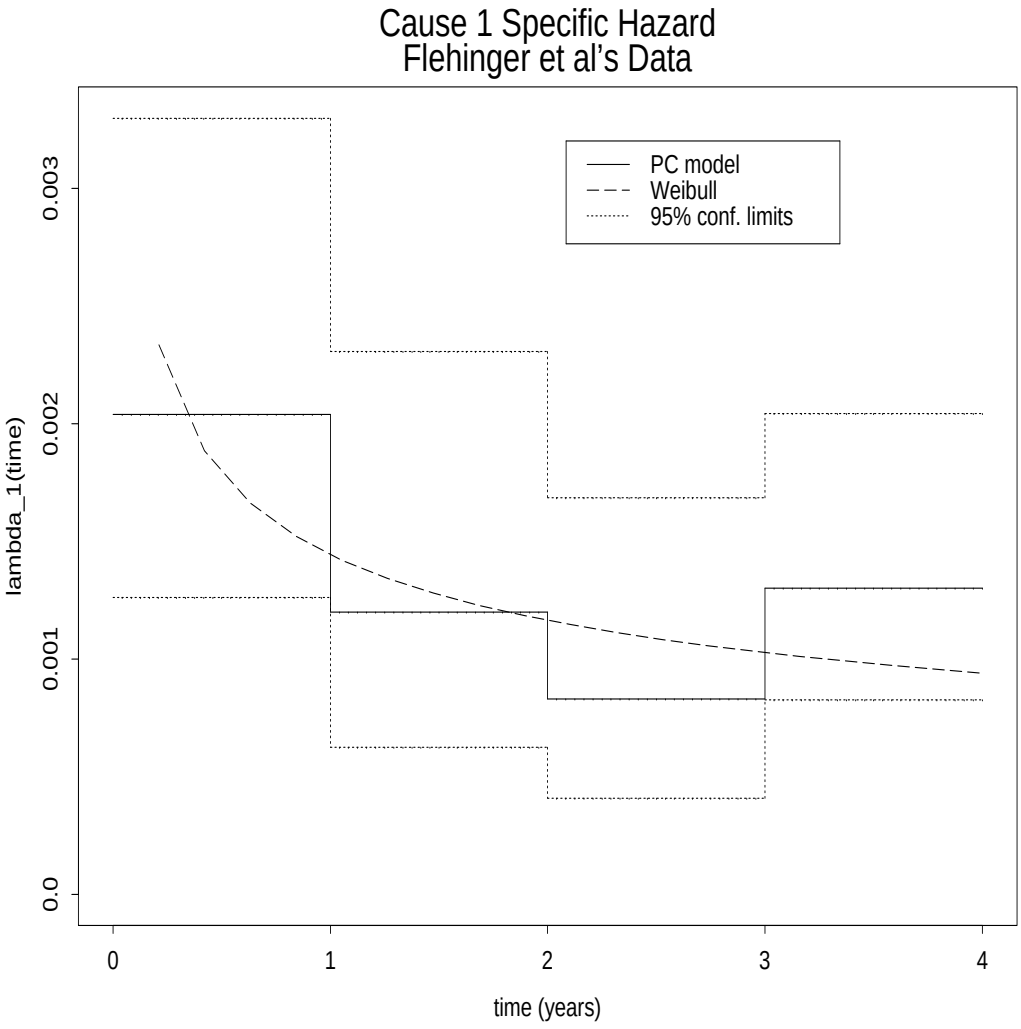
| | $j=1$ | $j=2$ | $j=3$ |
|---|---|---|---|
| $g=\{1\}$ | 0.282 | 0 | 0 |
| $g=\{2\}$ | 0 | 0.543 | 0 |
| $g=\{3\}$ | 0 | 0 | 0.116 |
| $g=\{1,3\}$ | 0.410 (0.079) | 0 | 0.445 (0.056) |
| $g=\{1,2,3\}$ | 0.308 (0.077) | 0.457 (0.119) | 0.439 (0.057) |

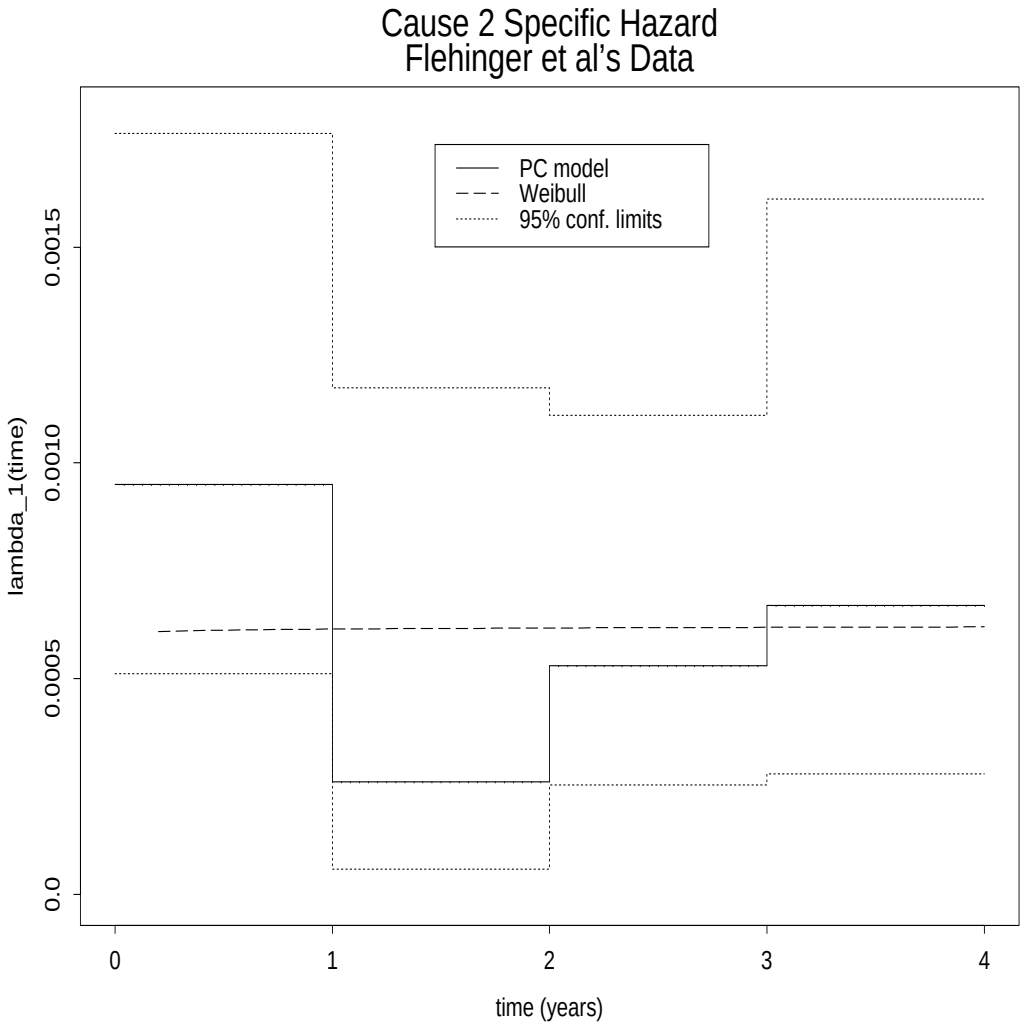
Flehinger et al. - Weibull

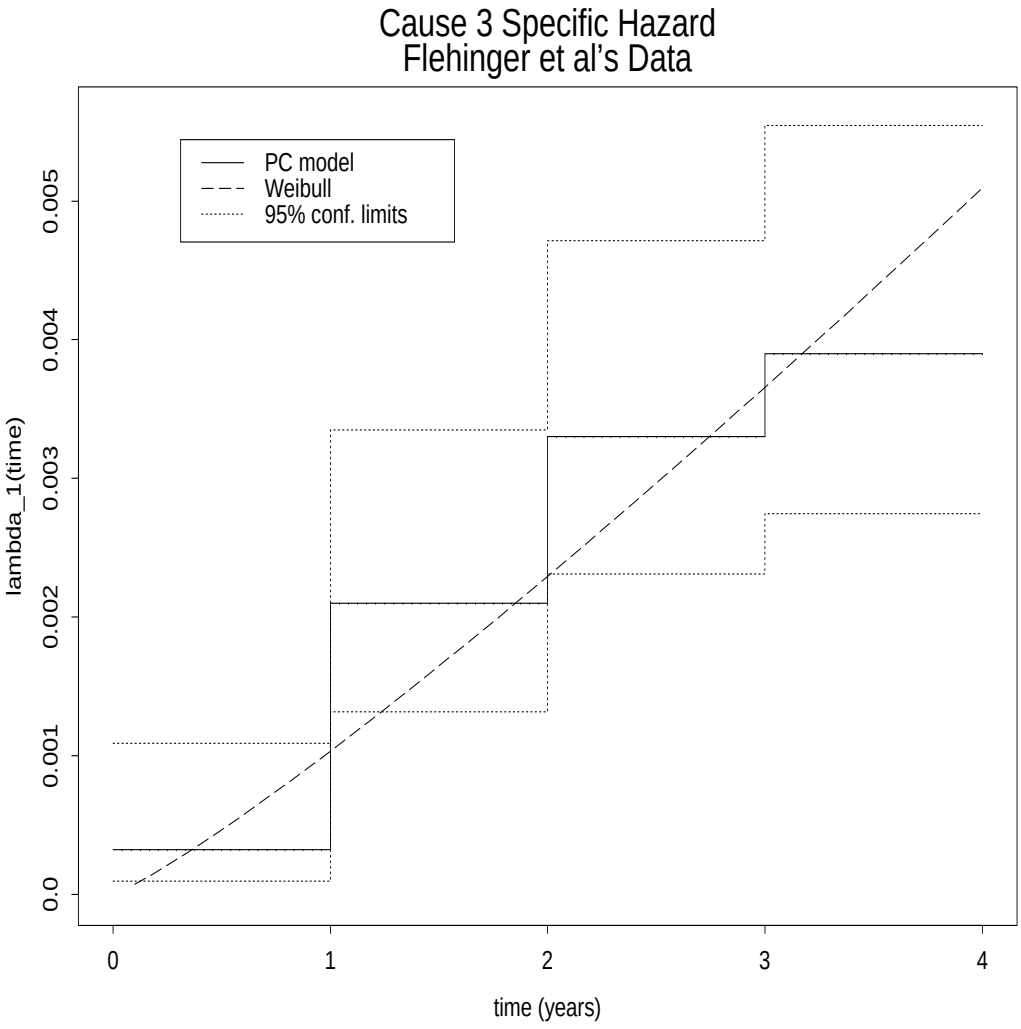
| | | | |
|---|---|---|---|
| $g=\{1\}$ | 0.278 | 0 | 0 |
| $g=\{2\}$ | 0 | 0.531 | 0 |
| $g=\{3\}$ | 0 | 0 | 0.118 |
| $g=\{1,3\}$ | 0.412 | 0 | 0.446 |
| $g=\{1,2,3\}$ | 0.310 | 0.469 | 0.436 |

P-value, likelihood ratio test for PH: 0.000004

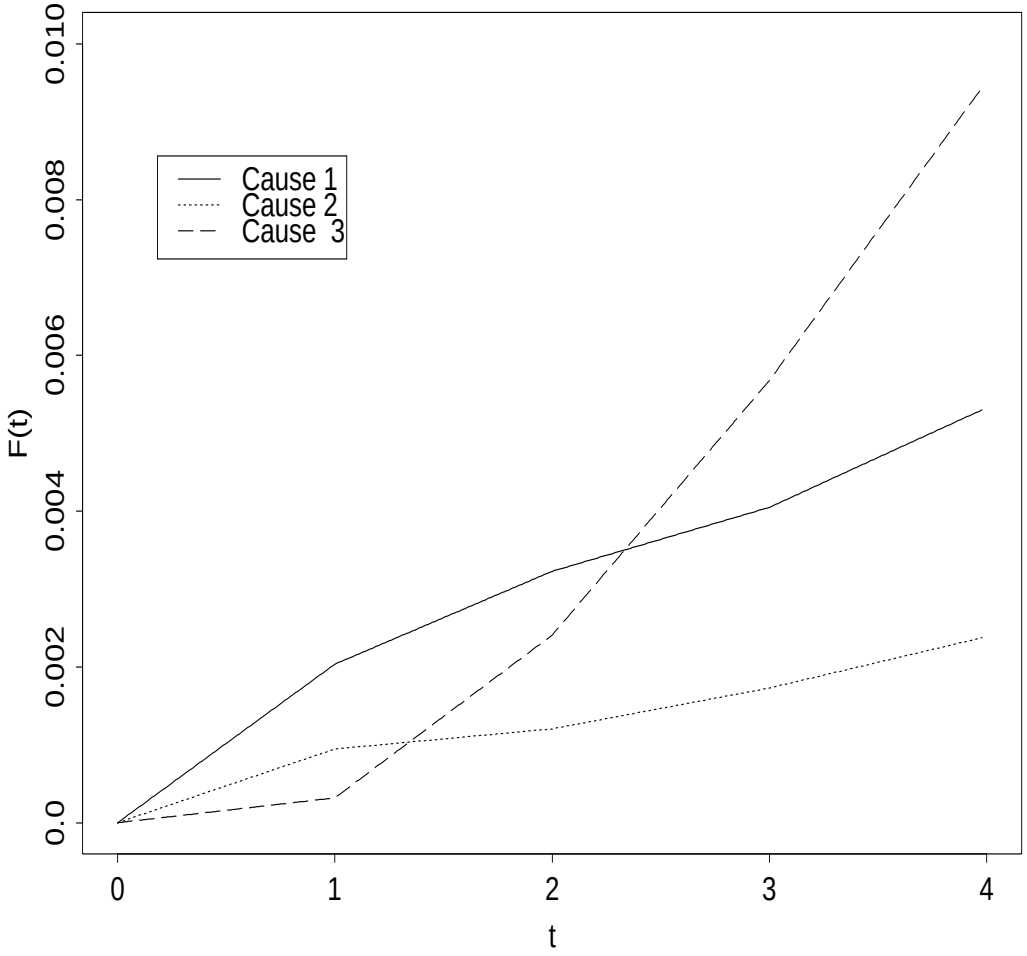
P-value, likelihood ratio test for SYM: 0.0254


Cause 1 Specific Hazard
Flehinger et al's Data
PC model
Weibull
95% conf. limits
lambda_1(time)
0.0
0.001
0.002
0.003
0
1
2
3
4
time (years)


Cause 2 Specific Hazard
Flehinger et al's Data
PC model
Weibull
95% conf. limits
lambda_1(time)
0.0
0.0005
0.0010
0.0015
0
1
2
3
4
time (years)


Cause 3 Specific Hazard
Flehinger et al's Data
PC model
Weibull
95% conf. limits
lambda_1(time)
0.0
0.001
0.002
0.003
0.004
0.005
0
1
2
3
4
time (years)

Incidence functions for the three causes



## Robustness Study

Simulate from:

- models with 3 proper groups $g_4 = \{1, 2\}$, $g_5 = \{1, 3\}$, $g_6 = \{1, 2, 3\}$.

- piecewise constant $M_1 - M_4$ models (all with 3 intervals);

- Weibull $W_1 - W_4$ models.

- 30% of masked items go to second stage.

- Model $PC_2$ is constructed with 2 intervals defined by the median of failure times.

- Model $PC_4$ has 4 intervals constructed with the 25th, 50th and 75th percentiles of the failure times.

- Model $WEI_3$ uses 3 intervals constructed with the 33rd and 67th percentiles of the failure times

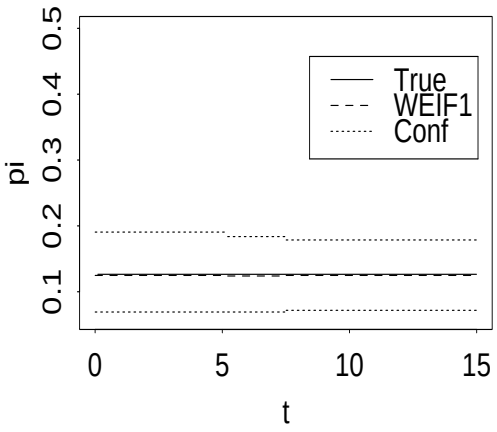
- Model $WEI_4$ has 4 intervals defined by the 25th, 50th and 75th percentiles of the failure times.

**Robustness - Estimation of $P_{\{1,2,3\}|1}$**

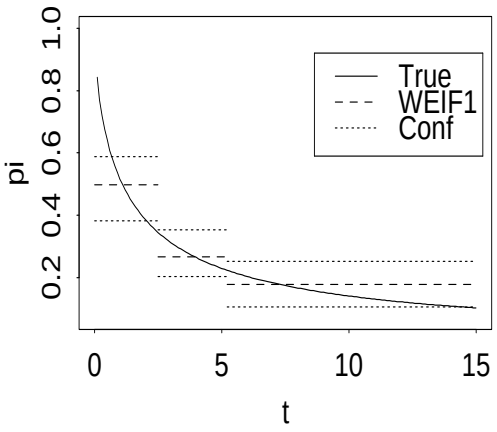
| Model | Monte Carlo average ($SE_{\text{SEM}}$) | | True value ($SE_{\text{MC}}$) |
|---|---|---|---|
| | $PC_2$ | $PC_4$ | |
| $M_1$ | 0· 200 (0· 031) | 0· 200 (0· 031) | 0· 200 (0· 030) |
| $M_2$ | 0· 195 (0· 031) | 0· 195 (0· 033) | 0· 200 (0· 028) |
| $M_3$ | 0· 196 (0· 034) | 0· 199 (0· 031) | 0· 200 (0· 031) |
| $M_4$ | 0· 201 (0· 035) | 0· 201 (0· 035) | 0· 200 (0· 032) |
| | $WEI_3$ | $WEI_4$ | |
| $W_1$ | 0· 195 (0· 042) | 0· 195 (0· 040) | 0· 200 (0· 046) |
| $W_2$ | 0· 196 (0· 048) | 0· 196 (0· 044) | 0· 200 (0· 050) |
| $W_3$ | 0· 200 (0· 033) | 0· 201 (0· 033) | 0· 200 (0· 038) |
| $W_4$ | 0· 196 (0· 031) | 0· 196 (0· 031) | 0· 200 (0· 031) |

$\pi_{1|\{1,2,3\}}(t)$

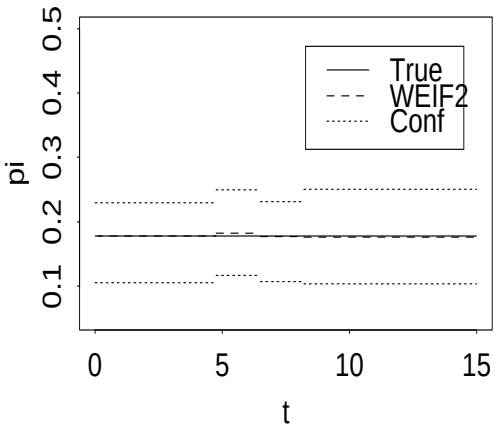
(a) Model W_1



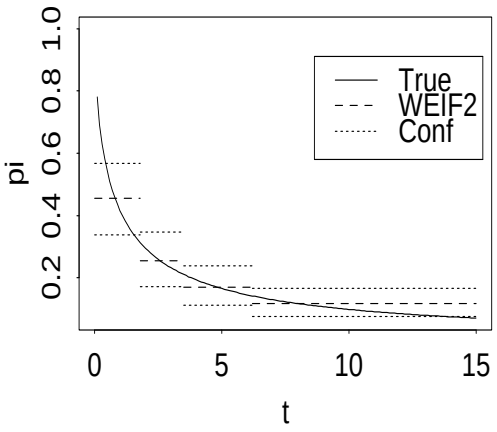

(b) Model W_3




(c) Model W_2




(d) Model W_4



## Robustness - Testing

We look at the robustness of the LRT tests to misspecification of the hazard rates model.

| Model | $A_{PH}$ | $A_{SYM}$ |
|---|---|---|
| $M_1\backslash W_1$ | True | False |
| $PC_2$ | 6\2 | 100\100 |
| $PC_4$ | 7\7 | 100\100 |
| $M_2\backslash W_2$ | True | True |
| $PC_2$ | 8\3 | 6\9 |
| $PC_4$ | 7\5 | 6\10 |
| $M_3\backslash W_3$ | False | True |
| $PC_2$ | 100\100 | 3\8 |
| $PC_4$ | 100\100 | 2\8 |
| $M_4\backslash W_4$ | False | False |
| $PC_2$ | 100\100 | 100\100 |
| $PC_4$ | 100\100 | 100\100 |

## Misspecified Hazard Rates

• True model $M_0$ with two competing risks each having constant cause-specific hazards, $\lambda_1$ and $\lambda_2$ on the interval $[0, t_{max}]$.

• Working model has cut points $0 = a_0 < a_1 < a_2 = t_{max}$

• For each cause $j = 1, 2$ the cause-specific hazard is

$$\lambda_j(t) = \lambda_{j1} 1_{(0,a_1]}(t) + \lambda_{j2} 1_{(a_1,t_{max}]}(t). \qquad (1)$$

Denote $n_{jk}$ the number of items that failed due to cause $j$ in the interval $(a_{k-1}, a_k]$ for each $k = 1, 2$.

Under the true model, $M_0$, $\widehat{\lambda}_1 = \frac{n_{11}+n_{12}}{e_1+e_2}$ and under model $M_1$, $\widehat{\lambda}_{11} = \frac{n_{11}}{e_1}$.

• $\widehat{\lambda}_{11}$ and $\widehat{\lambda_1}$ asymptotically unbiased for $\lambda_1$.

• The variance of $\widehat{\lambda}_{11}$ is larger than the variance of $\widehat{\lambda}_1$ for N large.

## Model Selection without Masking

• There are $K-1$ points to be determined; define $A_K = \{a_1, \ldots, a_{K-1}\}$. $a_0 = 0$ and $a_K$ =last observed time.

• Akaike Information Criterion (AIC): best fitting model is defined as the minimizer of an estimator of the Kullback–Leibler (KL) distance measure between a fitted model and the "true" model

$$\text{AIC}(A_K) = -2l_{OBS}(\theta) + 2JK.$$

• Bayesian Information Criterion (BIC): approximately equivalent to choosing the model with the largest posterior probability with respect to an uniform prior.

$$\text{BIC}(A_K) = -2l_{OBS}(\theta) + JK \log N.$$

• Minimum description length criterion: best fitting model as the one that produces the shortest code length of the data.

## Minimum Description Length (MDL) Principle

• The MDL principle (Rissanen 1989) is to split the code length for a set of data into two components: (i) a fitted model plus (ii) the data "conditioned on" the fitted model; i.e., the part in the data that is not explained by the fitted model

$$\begin{aligned} CL(\text{“data”}) &= CL(A_K, \widehat{\theta}) + CL(\text{“data”}\,|A_K, \widehat{\theta}) \\ &= CL(A_K) + CL(\widehat{\theta}) + CL(\text{“data”}\,|A_K, \widehat{\theta}). \end{aligned}$$

where $\widehat{\theta} = (\widehat{\lambda}_{11}, \ldots, \widehat{\lambda}_{JK})$

$$\text{MDL}(A_K) = \sum_{k=1}^{K} \log n_k + \frac{J}{2} \sum_{k=1}^{K} \log(n_k + n_{k+1} \ldots + n_K) - l_{OBS}(\theta), \tag{2}$$

where $n_k$ is the number of failures observed inside the interval $(a_{k-1}, a_k]$.

• Unlike AIC or BIC, in MDL the penalty for each interval is not the same.

• The penalty for the $k$th interval is a function of its "width" $n_k$.

• The "late" intervals (i.e., large $k$) are penalized more than those "early" intervals (i.e., small $k$).

## Simulation Results

• Three sample sizes $N = (100, 200, 800)$

• Three values for the probability $p$ that a masked item is sent to second stage analysis, $p \in \{0.3, 0.6, 1.0\}$

• Paired Wilcoxon tests were also applied to test if the difference between the median MSE values of any two methods is significant or not at 1.25% significance level.

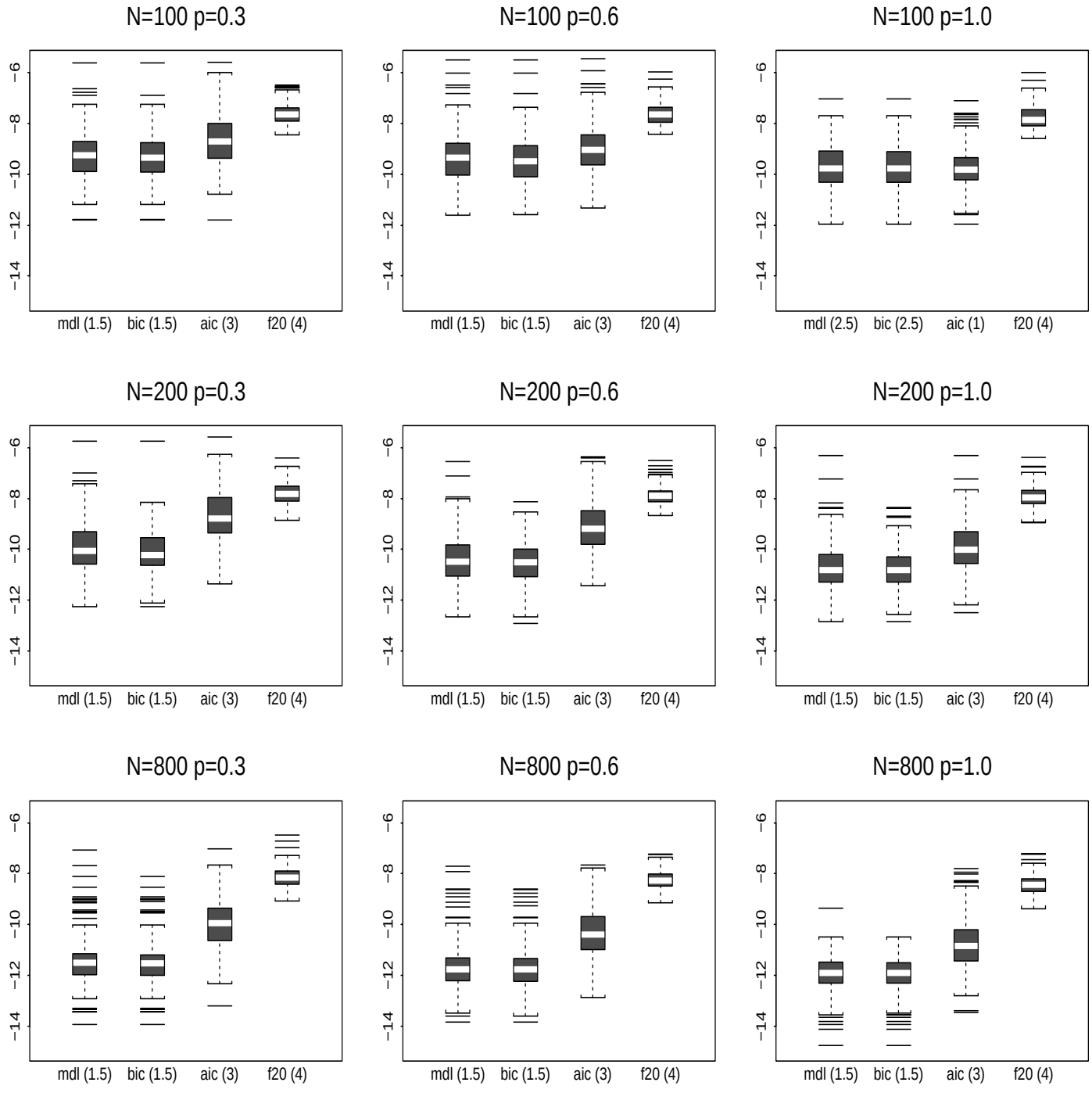
• All models have three causes of failure and three masking groups $g_4 = \{1, 2\}$, $g_5 = \{1, 3\}$ and $g_6 = \{1, 2, 3\}$.

1. M1: model with PC hazards with three intervals

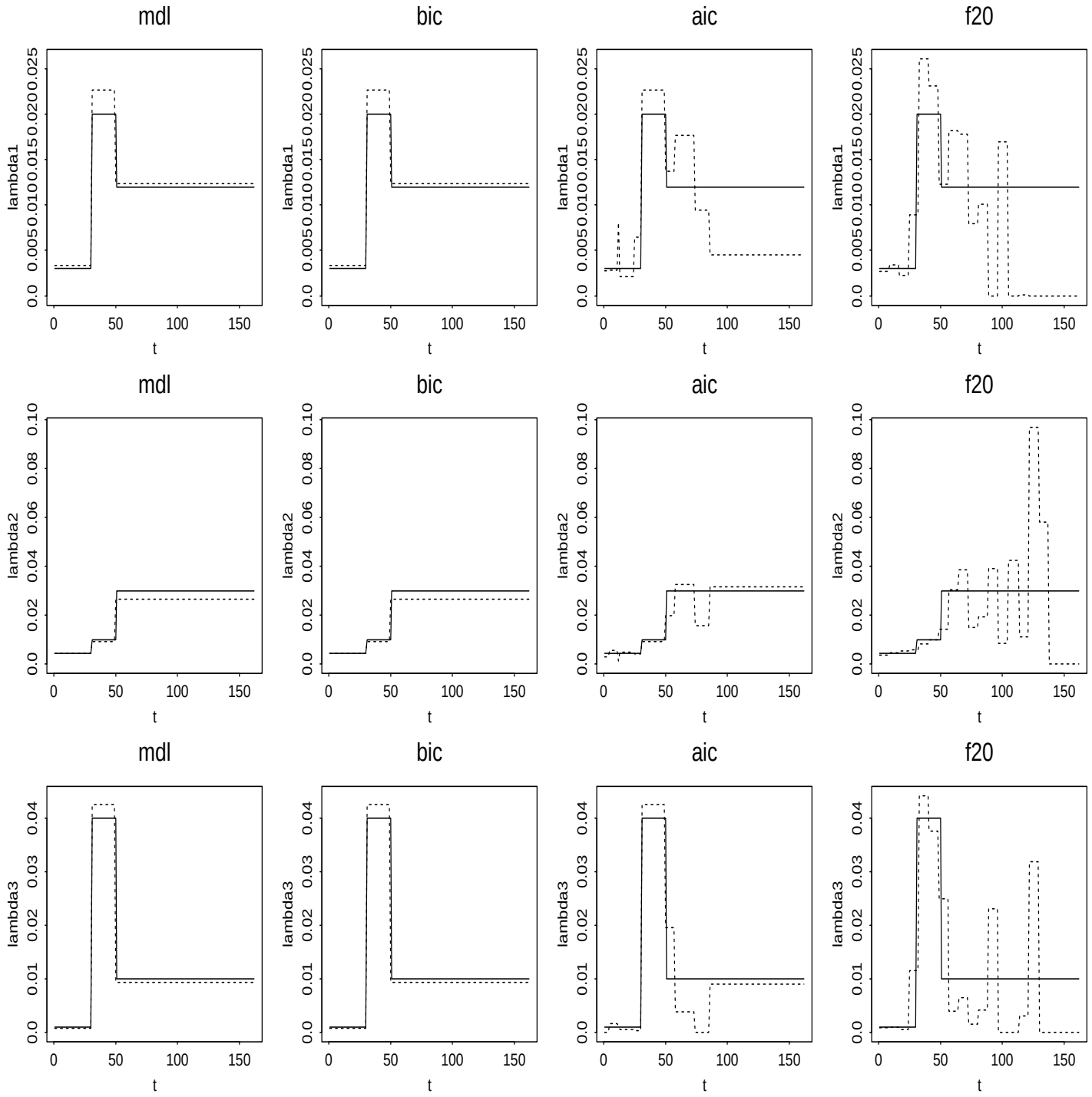
2. M2: model with PC hazards with seven intervals

3. W: Weibull distributed hazards

# Model M1 - MSE

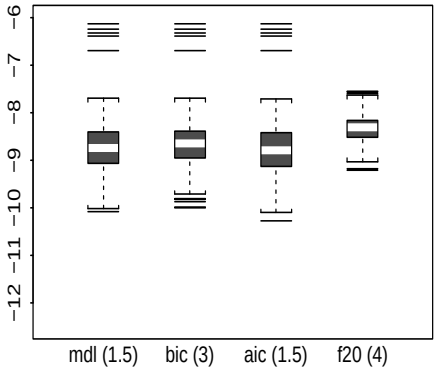

N=100 p=0.3
N=100 p=0.6
N=100 p=1.0
N=200 p=0.3
N=200 p=0.6
N=200 p=1.0
N=800 p=0.3
N=800 p=0.6
N=800 p=1.0
−6
−8
−10
−12
−14
mdl (1.5)
bic (1.5)
aic (3)
f20 (4)
mdl (2.5)
bic (2.5)
aic (1)

# Model M1 - $\widehat{\lambda}$

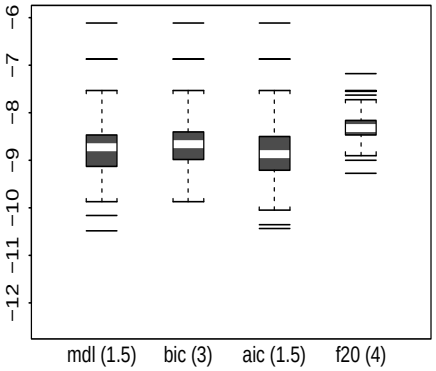


# Model M2 - MSE

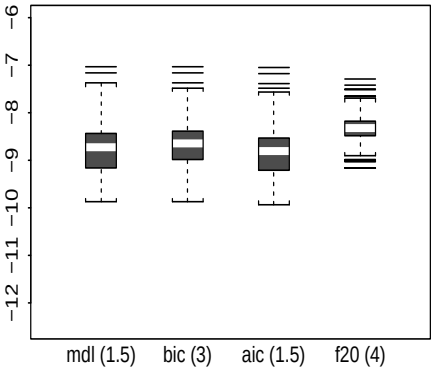
N=100 p=0.3



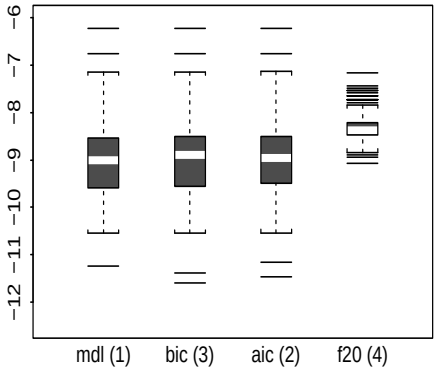

N=100 p=0.6



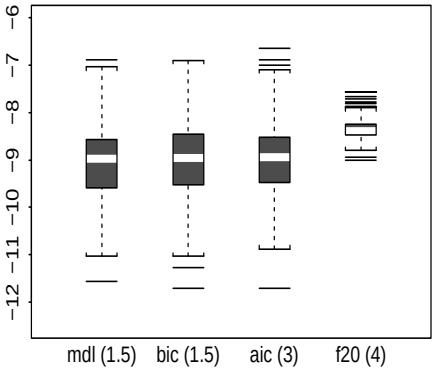

N=100 p=1.0



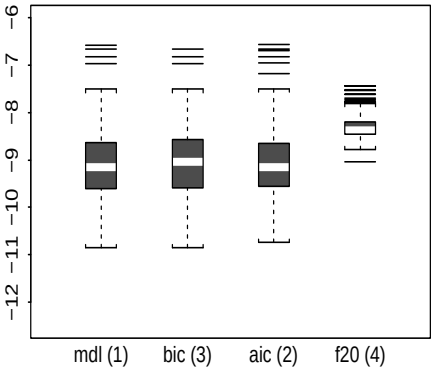

N=200 p=0.3



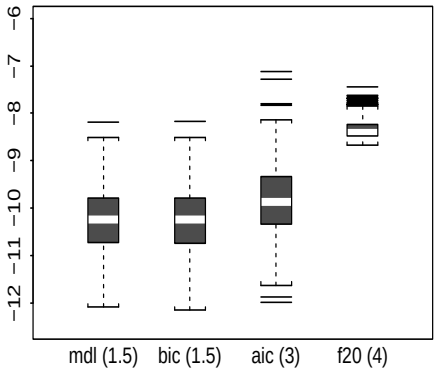

N=200 p=0.6



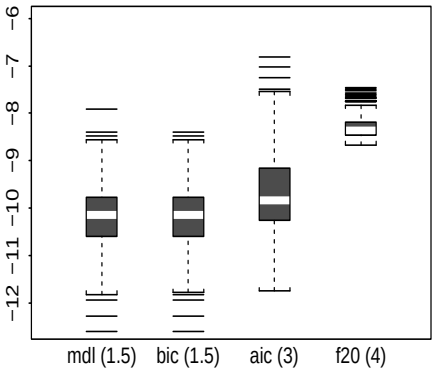

N=200 p=1.0



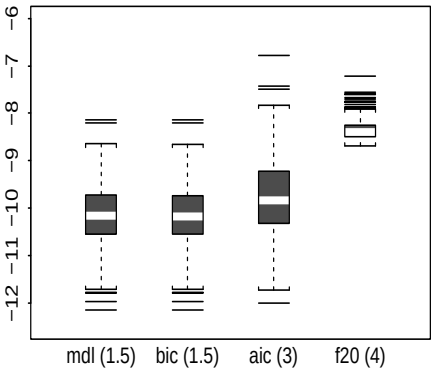

N=800 p=0.3



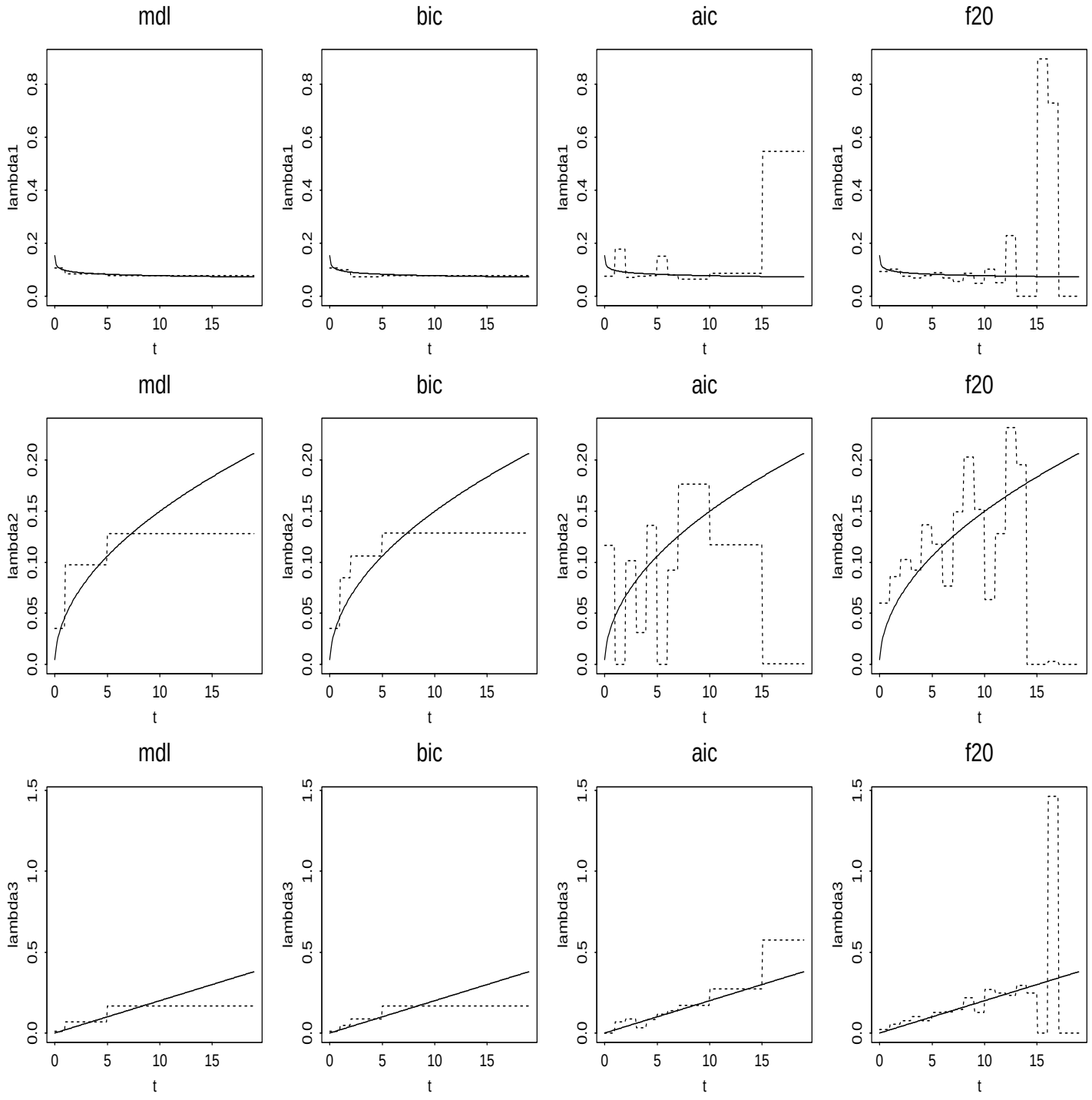

N=800 p=0.6




N=800 p=1.0



# Model W - $\hat{\lambda}$



## Future work

• Design: how to choose which masked items should be sent to a second stage analysis.

• Model selection: how to adapt for masked data.

• Data without second stage.

• Extend to hazard functions that are piecewise linear, splines, etc.