Partially Missing Data: Interval-Censored Data

Interval Censoring

- Interval censored data is a special type of missing data where the outcome of interest is observed to like in some interval.
- Right censoring is a special case of interval censoring where its interval is a right-half line, i.e., $[a, \infty)$.
- Interval censoring results in partially missing data.
- We consider multiple types of interval-censored events in this module, where one or more outcomes are of interest at the same time.

Introduction

Multiple types of interval censored events

- Many clinical and epidemiological studies are concerned with multiple diseases.
  Symptomatic disease: right-censored.
  Asymptomatic disease: interval-censored.

- Example: Atherosclerosis Risk in Communities Study
  - Symptomatic cardiovascular diseases: **myocardial infarction (MI)** and **stroke**. The study follow-up is was up to 27 years.
  - Asymptomatic diseases: **diabetes** and **hypertension**. Participants were examined over five clinic visits at least three years apart.

- Existing methods have treated the right- and interval-censored events separately.

Panel count data

- Panel count data arise when recurrence of the same type event is examined intermittently.
- Examples include the number of losses of feedwater flow in a nuclear plant (Gaver & O'Muircheartaigh, 1987), the number of tumors in a cancer patient (Byar, 1980), and the number of damaged joints in a psoriatic arthritis patient (Siannis et al., 2006).
- Essentially, panel count data can be viewed as some aggregated data from a recurrent event process: we observe the number of recurrences over time intervals, instead of actual occurrence times of the event.
- We are interested to assess the effects of treatments or other covariates on the event using panel count data.

Existing methods

- Proportional means model assumes the mean count to be proportional across (time-independent) covariates over time.

- The model is analogous to generalized estimating equations for marginal models for longitudinal data, except that one needs to estimate some unknown cumulative baseline function.

- Some estimators (Sun & Wei, 2000) are easy to compute but entail assumptions for examination times.

- Iterative algorithms (Wellner & Zhang, 2007) are suggested but they are not stable computationally.

- Spline method (Lu et al. 2009) approximates the baseline function with splines; it requires some arbitrary choice of splines and knots.

Multivariate panel count data

- Multivariate panel count data are also common in medical practice.
- Examples include recurrence of different types of tumors and recurrence of cardiovascular events (stroke, heart attack, myocardial infraction).
- There has been little work for multivariate panel count data. For example, He et al. (2008) extends the proportional means model to this case.

Outline of this talk

- We propose semiparametric proportional hazards models for multiple types of interval censored data.
- We also study semiparametric regression models for multivariate (and univariate) panel count data.
- The models uses random effects to account for both within- and between- event dependence, and includes time-dependent covariates in the regression.
- We propose nonparametric maximum likelihood estimation methods for inference.
- We devise computationally efficient algorithms for computation.

Multiple Interval-Censored Events

Some Notation

- K_1 asymptomatic events: $T_1, \ldots, T_{K_1}$.
- K_2 symptomatic events: $T_{K_1+1}, \ldots, T_K$.
- $K = K_1 + K_2$.
- $X_k(\cdot)$: p-vector of possibly time-dependent external covariates.
- b_1, b_2: random variables.

Model

- For asymptomatic event time T_k $(k = 1, \ldots, K_1)$, the hazard function is given by

$$\lambda_k(t; X_k, b_1) = \lambda_k(t) \exp\left\{\beta_k^{\mathrm{T}} X_k(t) + b_1\right\}.$$

- For symptomatic event time T_k $(k = K_1 + 1, \ldots, K)$, the hazard function is given by

$$\lambda_k(t; X_k, b_1, b_2) = \lambda_k(t) \exp\left\{\beta_k^{\mathrm{T}} X_k(t) + \gamma_k b_1 + b_2\right\}.$$

- $b_1 \sim N(0, \sigma_1^2)$, $b_2 \sim N(0, \sigma_2^2)$.

Observed Data

- For $k = 1, \ldots, K_1$:
  - Interval-censored observation (L_k, R_k).

- For $k = K_1 + 1, \ldots, K$:
  - C_k: (noninformative) censoring time (e.g. end of the study).
  - We observe $Y_k \equiv \min(T_k, C_k)$ and $\Delta_k \equiv I(T_k \le C_k)$.

- Observed data:

$$\begin{aligned} \mathcal{O}_i &= \{L_{ik}, R_{ik}, X_{ik}(\cdot) : k = 1, \ldots, K_1\} \\ &\cup \{Y_{ik}, \Delta_{ik}, X_{ik}(\cdot) : k = K_1 + 1, \ldots, K\}, \quad i = 1, \ldots, n. \end{aligned}$$

Observed-data Likelihood

$$\prod_{i=1}^{n} \int_{b_{i1}} \int_{b_{i2}} \prod_{k=1}^{K_1} \left[\exp\left\{ -\int_0^{L_{ik}} e^{\beta_k^{\mathrm{T}} X_{ik}(s) + b_{i1}} d\Lambda_k(s) \right\} - \exp\left\{ -\int_0^{R_{ik}} e^{\beta_k^{\mathrm{T}} X_{ik}(s) + b_{i1}} d\Lambda_k(s) \right\} \right]$$

$$\times \prod_{k=K_1+1}^{K} \left[\left\{ e^{\beta_k^{\mathrm{T}} X_{ik}(Y_{ik}) + \gamma_k b_{i1} + b_{i2}} \lambda_k(Y_{ik}) \right\}^{\Delta_{ik}} \exp\left\{ -\int_0^{Y_{ik}} e^{\beta_k^{\mathrm{T}} X_{ik}(s) + \gamma_k b_{i1} + b_{i2}} d\Lambda_k(s) \right\} \right]$$

$$\times \phi(b_{i1}; \sigma_1^2) \phi(b_{i2}; \sigma_2^2) db_{i1} db_{i2}$$

- $\phi(b_{ij}; \sigma_j^2) = (2\pi\sigma_j^2)^{-1/2} \exp\{-b_{ij}^2/(2\sigma_j^2)\}.$

Observed-data Likelihood

Interval-censored asymptomatic event time

$$\prod_{i=1}^{n}\int_{b_{i1}}\int_{b_{i2}}\prod_{k=1}^{K_1}\left[\exp\left\{-\int_0^{L_{ik}} e^{\beta_k^{\mathrm{T}}X_{ik}(s)+b_{i1}}d\Lambda_k(s)\right\}-\exp\left\{-\int_0^{R_{ik}} e^{\beta_k^{\mathrm{T}}X_{ik}(s)+b_{i1}}d\Lambda_k(s)\right\}\right]$$

$$\times\prod_{k=K_1+1}^{K}\left[\left\{e^{\beta_k^{\mathrm{T}}X_{ik}(Y_{ik})+\gamma_k b_{i1}+b_{i2}}\lambda_k(Y_{ik})\right\}^{\Delta_{ik}}\exp\left\{-\int_0^{Y_{ik}} e^{\beta_k^{\mathrm{T}}X_{ik}(s)+\gamma_k b_{i1}+b_{i2}}d\Lambda_k(s)\right\}\right]$$

$$\times\phi(b_{i1};\sigma_1^2)\phi(b_{i2};\sigma_2^2)db_{i1}db_{i2}$$

- $\phi(b_{ij};\sigma_j^2)=(2\pi\sigma_j^2)^{-1/2}\exp\{-b_{ij}^2/(2\sigma_j^2)\}.$

Observed-data Likelihood

$$\prod_{i=1}^{n}\int_{b_{i1}}\int_{b_{i2}}\prod_{k=1}^{K_1}\left[\exp\left\{-\int_0^{L_{ik}} e^{\beta_k^{\mathrm{T}}X_{ik}(s)+b_{i1}}d\Lambda_k(s)\right\}-\exp\left\{-\int_0^{R_{ik}} e^{\beta_k^{\mathrm{T}}X_{ik}(s)+b_{i1}}d\Lambda_k(s)\right\}\right]$$

$$\times\prod_{k=K_1+1}^{K}\left[\left\{e^{\beta_k^{\mathrm{T}}X_{ik}(Y_{ik})+\gamma_k b_{i1}+b_{i2}}\lambda_k(Y_{ik})\right\}^{\Delta_{ik}}\exp\left\{-\int_0^{Y_{ik}} e^{\beta_k^{\mathrm{T}}X_{ik}(s)+\gamma_k b_{i1}+b_{i2}}d\Lambda_k(s)\right\}\right]$$

$$\times\phi(b_{i1};\sigma_1^2)\phi(b_{i2};\sigma_2^2)db_{i1}db_{i2}$$

Right-censored symptomatic event time

- $\phi(b_{ij};\sigma_j^2)=(2\pi\sigma_j^2)^{-1/2}\exp\{-b_{ij}^2/(2\sigma_j^2)\}$.

Nonparametric Maximum Likelihood Estimation (NPMLE)

- Λ_k: step function that jumps only at $t_{k1}, \ldots, t_{k,m_k}$.
    - For $k = 1, \ldots, K_1$: ordered sequence of all L_{ik} and R_{ik} with $R_{ik} < \infty$.
    - For $k = K_1 + 1, \ldots, K$: ordered sequence of all Y_{ik} with $\Delta_{ik} = 1$.
- Jump sizes $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{k,m_k})$.
- The objective function

$$
\begin{aligned}
L_n(\theta, \mathcal{A}) = \prod_{i=1}^{n} \int_{b_{i1}} \int_{b_{i2}} \prod_{k=1}^{K_1} & \left\{ \exp\left(- \sum_{t_{kl} \le L_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}} \right) - \exp\left(- \sum_{t_{kl} \le R_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}} \right) \right\} \\
\times \prod_{k=K_1+1}^{K} & \left[\left\{ \Lambda_k\{Y_{ik}\} e^{\beta_k^{\mathrm{T}} X_{ik}(Y_{ik}) + \gamma_k b_{i1} + b_{i2}} \right\}^{\Delta_{ik}} \exp\left(- \sum_{t_{kl} \le Y_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + \gamma_k b_{i1} + b_{i2}} \right) \right] \\
& \times \phi(b_{i1}; \sigma_1^2) \phi(b_{i2}; \sigma_2^2) d b_{i1} d b_{i2}.
\end{aligned}
$$

- $\theta = (\beta_1, \ldots, \beta_K, \gamma_{K_1+1}, \ldots, \gamma_K, \sigma_1^2, \sigma_2^2)$, $\mathcal{A} = (\lambda_1, \ldots, \lambda_K)$.

Nonparametric Maximum Likelihood Estimation (NPMLE)

- Treat b_{i1} and b_{i2} as complete data and propose an EM algorithm.

- M-step: update parameters by maximizing the conditional expectation of

$$\sum_{i=1}^{n}\sum_{k=1}^{K_1}\log\left\{\exp\left(-\sum_{t_{kl}\leq L_{ik}}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{ikl}+b_{i1}}\right)-\exp\left(-\sum_{t_{kl}\leq R_{ik}}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{ikl}+b_{i1}}\right)\right\}$$

$$+\sum_{k=K_1+1}^{K}\left[\Delta_{ik}\left\{\log\Lambda_k\{Y_{ik}\}+\beta_k^{\mathrm{T}}X_{ik}(Y_{ik})+\gamma_k b_{i1}+b_{i2}\right\}\right.$$

$$\left.-\sum_{t_{kl}\leq Y_{ik}}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{ikl}+\gamma_k b_{i1}+b_{i2}}\right].$$

Nonparametric Maximum Likelihood Estimation (NPMLE)

- Treat b_{i1} and b_{i2} as complete data and propose an EM algorithm.
- M-step: update parameters by maximizing the conditional expectation of

$$\sum_{i=1}^{n}\sum_{k=1}^{K_1}\log\left\{\exp\left(-\sum_{t_{kl}\le L_{ik}}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{ikl}+b_{i1}}\right)-\exp\left(-\sum_{t_{kl}\le R_{ik}}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{ikl}+b_{i1}}\right)\right\}$$

$$+\sum_{k=K_1+1}^{K}\left[\Delta_{ik}\left\{\log\Lambda_k\{Y_{ik}\}+\beta_k^{\mathrm{T}}X_{ik}(Y_{ik})+\gamma_k b_{i1}+b_{i2}\right\}\right.$$

$$\left.-\sum_{t_{kl}\le Y_{ik}}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{ikl}+\gamma_k b_{i1}+b_{i2}}\right].$$

Introducing Poisson Random Variables

- We introduce independent Poisson variables

$$W_{ikl} \sim \text{Poisson}\left(\lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}}\right)$$

for $t_{kl} \leq R_{ik}^*$, $R_{ik}^* = L_{ik} I(R_{ik} = \infty) + R_{ik} I(R_{ik} < \infty)$.

- $A_{ik} = \sum_{t_{kl} \leq L_{ik}} W_{ikl} \sim \text{Poisson}\left(\sum_{t_{kl} \leq L_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}}\right)$.
- $B_{ik} = \sum_{L_{ik} < t_{kl} \leq R_{ik}} W_{ikl} \sim \text{Poisson}\left(\sum_{L_{ik} < t_{kl} \leq R_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}}\right)$.
- $P(A_{ik} = 0, B_{ik} > 0 | b_{i1}, X)$

$$\begin{aligned}
&= \exp\left(-\sum_{t_{kl} \leq L_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}}\right) \left\{1 - \exp\left(-\sum_{L_{ik} < t_{kl} \leq R_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}}\right)\right\} \\
&= \exp\left(-\sum_{t_{kl} \leq L_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}}\right) - \exp\left(-\sum_{t_{kl} \leq R_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl} + b_{i1}}\right)
\end{aligned}$$

EM Algorithm

- Observed data:

$$\widetilde{\mathcal{O}}_i = \{A_{ik} = 0, B_{ik} > 0 : k = 1, \ldots, K_1\}$$

$$\cup\{Y_{ik}, \Delta_{ik} : k = K_1 + 1, \ldots, K\}.$$

- Missing data: b_{i1}, b_{i2}, and W_{ikl} $(i = 1, \ldots, n; k = 1, \ldots, K_1; l = 1, \ldots, m_k, t_{kl} \leq R^*_{ik})$.
- E-step: evaluate $\widehat{E}(W_{ikl})$, $\widehat{E}\{\exp(b_{i1})\}$, $\widehat{E}\{\exp(\gamma_k b_{i1} + b_{i2})\}$, $\widehat{E}(b_{i1})$, $\widehat{E}(b^2_{i1})$, and $\widehat{E}(b^2_{i2})$ through Gaussian-Hermite quadratures.

EM Algorithm

- M-step: maximize the conditional expectation of the complete-data log-likelihood function.

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{k=1}^{K_1}\sum_{l=1}^{m_k} I(t_{kl} \le R_{ik}^*) \Big\{ -\widehat{E}\left(\log W_{ikl}!\right) + \widehat{E}\left(W_{ikl}\right) \Big(\log \lambda_{kl} + \beta_k^{\mathrm{T}} X_{ikl}\Big) \\
&\qquad + \widehat{E}\left(W_{ikl} b_{i1}\right) - \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl}} \widehat{E}\left(e^{b_{i1}}\right) \Big\} \\
&+ \sum_{k=K_1+1}^{K} \Big[\Delta_{ik} \Big\{ \log \Lambda_k\{Y_{ik}\} + \beta_k^{\mathrm{T}} X_{ik}(Y_{ik}) + \gamma_k \widehat{E}\left(b_{i1}\right) + \widehat{E}\left(b_{i2}\right) \Big\} \\
&\qquad - \sum_{t_{kl} \le Y_{ik}} \lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{ikl}} \widehat{E}\left(e^{\gamma_k b_{i1} + b_{i2}}\right) \Big].
\end{aligned}
$$

Inference: Profile Likelihood Approach

- We show that the resulting estimators are consistent and the parametric components are asymptotically normal and efficient.
- Variance estimation based on profile log-likelihood function

$$pl_n(\theta) = \max_{\mathcal{A}} \log L_n(\theta, \mathcal{A}).$$

  - Evaluated by the proposed EM algorithm but only update $\Lambda_1, \ldots, \Lambda_K$ in the M-step.
- We estimate the covariance matrix of $\widehat{\boldsymbol{\theta}}$ by the inverse of

$$\sum_{i=1}^{n} \left\{ \left. \frac{\partial pl_i(\theta)}{\partial \theta} \right|_{\theta = \widehat{\boldsymbol{\theta}}} \right\}^{\otimes 2}.$$

  - pl_i is the ith subject's contribution to pl_n.

Dynamic Prediction

- A subject with covariate X.
- Event history at time $t > 0$

$$\begin{aligned}\mathcal{O}(t) &= \{L_k(t), R_k(t) : k = 1, \ldots, K_1\} \\ &\quad \cup \{Y_k(t), \Delta_k(t) : k = K_1 + 1, \ldots, K\}.\end{aligned}$$

- Two measures.
  - Survival function / cumulative incidence function.
  - Risk score.

Predict Cumulative Incidence Function

- Survival function: $P(T_k \geq t|X)$.
- In some studies, **one of the symptomatic event is terminal (e.g., death) such that its occurrence precludes the development of other events**.
- At time $s > 0$, we observe event history $\mathcal{O}(s)$ and $T_k > s$, $T_K > s$. So we predict $P(T_k \leq t, T_k \leq T_K|\mathcal{O}(s), X)$ by

$$\int_b P(s < T_k \leq t, T_k \leq T_K|X, b)\widehat{P}(b|\mathcal{O}(s), X)db.$$

- $\widehat{P}(b|\mathcal{O}(s), X) \propto \widehat{P}(O(s)|b, X)\phi(b_1, \widehat{\sigma}_1^2)\phi(b_2, \widehat{\sigma}_2^2)$.

Dynamic Prediction: Risk Score

- Risk score for a symptomatic event time T_k:

$$\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} X_k(s) + \widehat{\gamma}_k \widehat{b}_1(s) + \widehat{b}_2(s).$$

- $\widehat{b}(s) \equiv (\widehat{b}_1(s), \widehat{b}_2(s))$ is a suitable estimator of b given the event history $\mathcal{O}(s)$.
    - Posterior mean or mode of b.
    - An imputed value from the posterior distribution.

- The risk score using the posterior mean is given by

$$\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} X_k(s) + \int_b \{\widehat{\gamma}_k b_1(s) + b_2(s)\} \widehat{P}(b|\mathcal{O}(s), X) db.$$

Method for Panel Count Data

Some notations

- We consider a random sample of n subjects with K types of recurrent events.
- Let $N_{ki}(t)$ denote the number of the kth type of event the ith subject has experienced by time t.
- Let $X_i(t)$ denote a set of potentially time-dependent external covariates for the ith subject.

Semiparametric regression model

- Our model assumes that $N_{ki}(t)$ is a non-homogeneous Poisson process with intensity function

$$\lambda_{ki}(t) = \lambda_k(t) e^{\beta_k^{\mathrm{T}} X_i(t) + b_{ki}^{\mathrm{T}} Z_i(t) + \xi_i^{\mathrm{T}} \widetilde{Z}_i(t)}. \tag{1}$$

- b_{ki} is a p-dimensional random effect for the kth type of event; ξ_i is a q-dimensional random effect shared by the K types of events.

- Both $Z_i(t)$ and $\widetilde{Z}_i(t)$ are covariates associated with random effects and they contain the constant one and part of $X_i(t)$.

- b_{ki} and ξ_i are independent and follow zero-mean normal distribution with covariance matrices Σ_k and Ψ, respectively.

Observed panel data

- In panel count data, we observe the event counts at a sequence of examination times, $N_{ki}(U_{ki1}), \ldots, N_{ki}(U_{kim_{ki}})$ or, equivalently,

$$\Delta_{kij} = N_{ki}(U_{kij}) - N_{ki}(U_{ki,j-1}), \quad (j = 1, \ldots, m_{ki}).$$

- Here, $0 < U_{ki1} < \cdots < U_{kim_{ki}} = C_{ki}$ are the m_{ki} examination times for $N_{ki}(\cdot)$, C_{ki} is the end of follow-up, $U_{ki0} = 0$, and $N_{ki}(0) = 0$.

- We assume that the examination times are non-informative (independent of the events given covariates).

Likelihood function

- The likelihood is proportional to

$$\prod_{i=1}^{n}\left[\int_{\xi_i}\phi(\xi_i;\Psi)\prod_{k=1}^{K}\int_{b_{ki}}\phi(b_{ki};\Sigma_k)\prod_{j=1}^{m_{ki}}\frac{\{\int_{U_{ki,j-1}}^{U_{kij}}e^{\beta_k^{\mathrm{T}}X_i(t)+b_{ki}^{\mathrm{T}}Z_i(t)+\xi_i^{\mathrm{T}}\widetilde{Z}_i(t)}d\Lambda_k(t)\}^{\Delta_{kij}}}{\Delta_{kij}!}\right.$$

$$\left.\times\exp\left\{-\int_0^{C_{ki}}e^{\beta_k^{\mathrm{T}}X_i(t)+b_{ki}^{\mathrm{T}}Z_i(t)+\xi_i^{\mathrm{T}}\widetilde{Z}_i(t)}d\Lambda_k(t)\right\}db_{ki}d\xi_i\right].$$

- $\phi(\cdot;\Sigma)$ denotes the multivariate normal density with mean 0 and covariance matrix Σ.

Nonparametric maximum likelihood estimation

- We adopt NPMLE approach for estimation.
- In this approach, the baseline function, $\Lambda_k(\cdot)$, is a step function with jumps at all examination time points associated with event k.
- Then the likelihood becomes

$$\prod_{i=1}^{n}\left\{\int_{\xi_i}\phi(\xi_i;\Psi)\prod_{k=1}^{K}\int_{b_{ki}}\phi(b_{ki};\Sigma_k)\prod_{j=1}^{m_{ki}}\frac{\left(\sum_{l:t_{kl}\in(U_{ki,j-1},U_{kij}]}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{kil}+b_{ki}^{\mathrm{T}}Z_{kil}+\xi_i^{\mathrm{T}}\widetilde{Z}_{kil}}\right)^{\Delta_{kij}}}{\Delta_{kij}!}\right.$$

$$\left.\times\exp\left(-\sum_{l:t_{kl}\leq C_{ki}}\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{kil}+b_{ki}^{\mathrm{T}}Z_{kil}+\xi_i^{\mathrm{T}}\widetilde{Z}_{kil}}\right)db_{ki}d\xi_i\right\}, \qquad (2)$$

where $X_{kil}=X_i(t_{kl})$, $Z_{kil}=Z_i(t_{kl})$, and $\widetilde{Z}_{kil}=\widetilde{Z}_i(t_{kl})$.

More on NPMLE

- The advantage of NPMLE is
  (a) we use all information told by the data, since those time points are only possible time for Λ_k's estimation;
  (b) there is minimal assumption for Λ' so without subjective choice as used in spline approach.

- The challenge is how to compute NPMLE in the presence of hundreds of parameters.

Algorithm

- We adopt Poissonization tricks which have been used for analyzing interval censored data before.

- Consider independent Poisson random variables W_{kil} $(k = 1, \ldots, K; i = 1, \ldots, n; l = 1, \ldots, m_k)$

$$W_{ikl} \sim Poisson\left(\lambda_{kl} e^{\beta_k^{\mathrm{T}} X_{kil} + b_{ki}^{\mathrm{T}} Z_{kil} + \xi_i^{\mathrm{T}} \widetilde{Z}_{kil}}\right).$$

- The key fact is that the likelihood for Δ_{ij} is the same as the likelihood for

$$\sum_{l: t_{kl} \in (U_{ki,j-1}, U_{kij}]} W_{kil}.$$

- Thus, we can treat W's and the random effects as missing data and apply the EM algorithm for maximizing the likelihood function.

Full data and likelihood

- Specifically,

$$(\xi_i, b_{ki}, W_{ki}) \;\; (k = 1, \ldots, K; i = 1, \ldots, n)$$

are missing data, and

$$\{X_i(\cdot), \sum_{l:t_{kl}\in(U_{ki,j-1},U_{kij}]} W_{kil}\}$$

are observed data.

- The complete-data log-likelihood is

$$\sum_{i=1}^{n}\left[-\frac{1}{2}\log(2\pi)^q|\Psi| - \frac{1}{2}\xi_i^{\mathrm{T}}\Psi^{-1}\xi_i + \sum_{k=1}^{K}\left\{-\frac{1}{2}\log(2\pi)^p|\Sigma_k| - \frac{1}{2}b_{ki}^{\mathrm{T}}\Sigma_k^{-1}b_{ki}\right\}\right.$$
$$+\sum_{k=1}^{K}\sum_{l=1}^{m_k} I(t_{kl} \leq C_{ki})\left\{W_{kil}\left(\log\lambda_{kl} + \beta_k^{\mathrm{T}}X_{kil} + b_{ki}^{\mathrm{T}}Z_{kil} + \xi_i^{\mathrm{T}}\widetilde{Z}_{kil}\right)\right.$$
$$\left.\left. -\lambda_{kl}e^{\beta_k^{\mathrm{T}}X_{kil} + b_{ki}^{\mathrm{T}}Z_{kil} + \xi_i^{\mathrm{T}}\widetilde{Z}_{kil}} - \log W_{kil}!\right\}\right]. \qquad (3)$$

More on EM algorithm

- E-step involves the conditional distribution of W's and the random effects given the observed data for computing some conditional expectations.

- The conditional mean of functions of W's is explicit; the conditional mean of function of the random effects can be evaluated using Gaussian-Hermite quadratures.

- Big advantage of the EM algorithm is in the M-step:
  - β_k can be updated by a weighted Poisson regression.
  - More importantly, the jump sizes for Λ_k can be updated explicitly:

$$\lambda_{kl} = \frac{\sum_{i=1}^{n} I(C_{ki} \geq t_{kl}) \widehat{E}(W_{kil})}{\sum_{i=1}^{n} I(C_{ki} \geq t_{kl}) \widehat{E}\left(e^{\beta_k^{\mathrm{T}} X_{kil} + b_{ki}^{\mathrm{T}} Z_{kil} + \xi_i^{\mathrm{T}} \widetilde{Z}_{kil}}\right)}.$$

- Therefore, the EM algorithm is immune to the large number of parameters in NPMLE.

Asymptotic properties

- (Consistency) Under some regularity conditiosn, $\|\widehat{\theta} - \theta_0\| \to 0$ almost surely, where $\| \cdot \|$ is the Euclidean norm. In addition, $\sum_{k=1}^{K} \sup_{t \in [0, \tau_k]} |\widehat{\Lambda}_k(t) - \Lambda_{k0}(t)| \to 0$ almost surely.

- $n^{1/2}(\widehat{\theta} - \theta_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

Variance estimation

- Standard profile likelihood theory also implies that the asymptotic variances can be estimated consistently based on the profile likelihood function.

- To compute the profile likelihood function given β's and Σ's, we apply the same EM algorithm so only Λ_k's is updated, which is fast since its jump sizes have explicit form during the iteration.

- The limiting covariance matrix of $n^{1/2}(\widehat{\theta} - \theta)$ can be consistently estimated by the inverse of

$$n^{-1}\left(\sum_{i=1}^{n}\left\{\frac{pl_i(\widehat{\theta}+h_n e_j) - pl_i(\widehat{\theta})}{h_n}\right\}\left\{\frac{pl_i(\widehat{\theta}+h_n e_l) - pl_i(\widehat{\theta})}{h_n}\right\}\right).$$

- Empirically, we find $h_n = n^{-1/2}$ or $5n^{-1/2}$ works well.

Simulation Study

Simulation Study with Multiple Interval-Censored Events

- $X_1 \sim \text{Unif}(0,1)$.
- $X_2(t) = B_1 I(t \leq V) + B_2 I(t > V)$, where $B_1, B_2 \sim$Bernoulli(0.5), $V \sim \text{Unif}(0,\tau)$, and $\tau = 4$.
- $X_k = (X_1, X_2)^{\mathrm{T}}$.
- $K_1 = 2, K_2 = 3$. Last event is a terminal event.
- $\Lambda_1(t) = 0.5t$, $\Lambda_k(t) = \log\{1 + t/(k-1)\}$ for $k = 2, \ldots, 5$.
- Censoring time $C \sim$Unif$(2\tau/3, \tau)$.
- Potential examination times $U_m \sim U_{m-1} + 0.1 +$Unif$(0, 0.5)$ with $U_0 = 0$.

Simulation Results

Table: Summary Statistics for the Simulation Studies

	True	$n = 200$				$n = 500$			
	Value	Bias	SE	SEE	CP	Bias	SE	SEE	CP
β_{11}	0.5	0.024	0.505	0.515	0.959	0.002	0.309	0.308	0.950
β_{12}	0.4	0.020	0.291	0.286	0.946	0.007	0.172	0.173	0.956
β_{21}	0.5	0.022	0.521	0.517	0.955	0.013	0.307	0.312	0.955
β_{22}	-0.2	-0.013	0.295	0.288	0.944	-0.008	0.184	0.175	0.944
β_{31}	-0.5	-0.014	0.457	0.477	0.957	-0.012	0.288	0.289	0.950
β_{32}	0.5	0.005	0.245	0.260	0.963	0.004	0.155	0.158	0.950
β_{41}	-0.5	-0.020	0.483	0.505	0.963	-0.012	0.301	0.305	0.952
β_{42}	0.5	0.014	0.268	0.276	0.953	0.004	0.168	0.168	0.949
β_{51}	0.3	-0.004	0.440	0.451	0.959	0.003	0.272	0.275	0.959
β_{52}	-0.2	0.007	0.230	0.240	0.957	0.000	0.142	0.146	0.957
γ_3	0.25	-0.002	0.225	0.226	0.967	-0.006	0.143	0.140	0.953
γ_4	0.25	0.008	0.261	0.277	0.977	0.003	0.171	0.170	0.961
γ_5	0.25	0.003	0.244	0.252	0.974	0.000	0.155	0.155	0.956
σ_1^2	1	0.077	0.424	0.573	0.970	0.041	0.242	0.323	0.973
σ_2^2	1	-0.047	0.289	0.332	0.988	-0.027	0.177	0.192	0.976

Bias: mean (β_k) or median (γ_k, σ_j^2) bias.

SE: empirical standard error or mean absolute deviation (γ_k, σ_j^2).

SEE: mean (β_k) or median (γ_k, σ_j^2) standard error estimator.

CP: empirical coverage percentage of the 95% confidence interval based on Wald method. For σ_j^2, the confidence intervals are based on the log transformation.

Estimated Baseline Functions

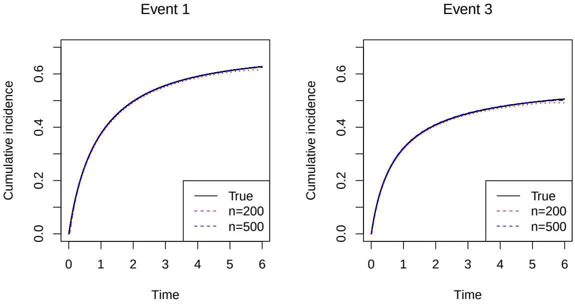

Figure: Estimation of the baseline cumulative incidence function when no event history is available.

Dynamic Prediction

- At the first visit time $s = 1$
  - Event 2 has occurred.
    - $(L_2, R_2) = (0, 1)$.
  - Events 1, 3, 4, and 5 have not.
    - $(L_1, R_1) = (1, \infty)$.
    - $(Y_3, \Delta_3) = (1, 1)$.
    - $(Y_4, \Delta_4) = (1, 1)$.
    - $(Y_5, \Delta_5) = (1, 1)$.

Prediction

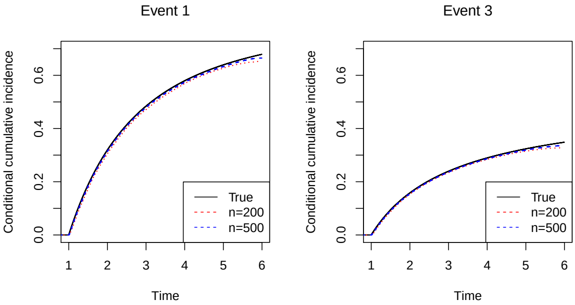

Figure: Estimation of the baseline cumulative incidence function conditional on the event history.

Simulation Study for Panel Count Data

- We considered two types of recurrent events with intensity functions $0.7(1+0.7t)^{-1}e^{\beta_{11}X_1+\beta_{12}X_2+b_1+\xi}$ and $0.4e^{\beta_{21}X_1+\beta_{22}X_2+b_2+\xi}$.

- X_1 and X_2 are independent Ber(0.5) and Un[0,1], respectively.

- The number of the examination times, m_{ik}, is 1, 2, or 3 randomly, and we generated m_{ki} time points from $\text{Un}(0, 3-0.1m_{ki})$ and let U_{kij} be the jth ordered time point plus $\text{Un}(0.1(j-1), 0.1j)$.

- All U_{kij} are between 0 and 3, and all adjacent examination times are separated by at least 0.1.

- On average, there are 1.7 events of the first type and 2.3 events of the second type for each subject.

Simulation result

						Joint analysis						
		$n = 200$				$n = 400$				$n = 800$		
	Est	SE	SEE	CP	Est	SE	SEE	CP	Est	SE	SEE	CP
$\beta_{11} = 0.5$	0.498	0.179	0.183	96	0.498	0.127	0.128	95	0.499	0.089	0.090	95
$\beta_{12} = -0.5$	-0.500	0.310	0.317	95	-0.498	0.218	0.218	95	-0.499	0.154	0.152	95
$\beta_{21} = 0$	-0.002	0.156	0.161	96	-0.002	0.113	0.112	95	0.001	0.080	0.079	95
$\beta_{22} = 0.6$	0.602	0.275	0.282	95	0.602	0.195	0.196	95	0.602	0.137	0.137	95
$\sigma_1^2 = 0.5$	0.487	0.161	0.181	98	0.493	0.113	0.122	98	0.496	0.078	0.084	97
$\sigma_2^2 = 0.4$	0.387	0.134	0.155	98	0.395	0.096	0.105	97	0.396	0.068	0.072	97
$\psi^2 = 0.25$	0.246	0.099	0.115	97	0.246	0.070	0.077	97	0.249	0.050	0.053	97

						Separate analysis						
		$n = 200$				$n = 400$				$n = 800$		
	Est	SE	SEE	CP	Est	SE	SEE	CP	Est	SE	SEE	CP
$\beta_{11} = 0.5$	0.503	0.181	0.182	95	0.499	0.126	0.128	95	0.499	0.089	0.090	95
$\beta_{12} = -0.5$	-0.499	0.311	0.311	95	-0.499	0.218	0.216	95	-0.499	0.155	0.151	95
$\sigma_1^2 = 0.75$	0.741	0.161	0.177	97	0.741	0.113	0.121	96	0.746	0.079	0.084	96
$\beta_{21} = 0$	-0.004	0.159	0.159	95	-0.003	0.112	0.112	95	-0.001	0.079	0.079	95
$\beta_{22} = 0.6$	0.605	0.277	0.280	95	0.605	0.193	0.196	95	0.601	0.138	0.138	95
$\sigma_2^2 = 0.65$	0.634	0.127	0.142	97	0.643	0.089	0.097	97	0.646	0.063	0.068	96

Computation detail

- With the convergence threshold at 10^{-3} and the maximum number of iterations at 1,000, the EM algorithm converged in more than 99.8% of the replicates for $n = 200$ and always converged for $n = 400$ and 800.
- It took approximately 30, 100, and 500 seconds to analyze one dataset of size 200, 400, and 800, respectively.

Estimated baseline function

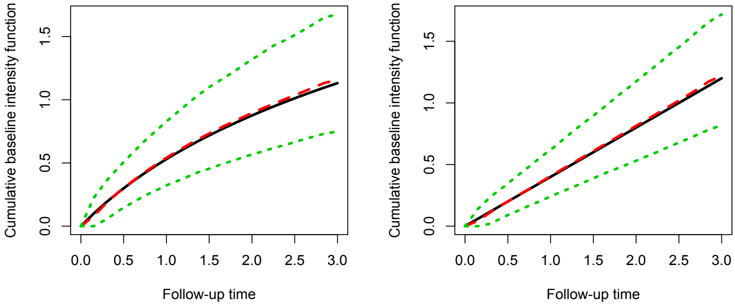

Figure: Simulation results on estimating the cumulative baseline intensity functions $\Lambda_1(\cdot) = \log(1 + 0.7t)$ (left panel) and $\Lambda_2(\cdot) = 0.4t$ (right panel): the solid and dashed curves pertain to the true value and averaged estimate, respectively; the dotted curves pertain to the lower and upper 2.5 percentiles of the estimates.

Real Data Applications

Application to ARIC Study

- Epidemiological cohort study conducted in four U.S. communities: Forsyth County, NC; Jackson, MS; Minneapolis, MN; and Washington County, MD.
- 15,792 participants
- A baseline examination between 1987 and 1989; four subsequent examinations in 1990-1992, 1993-1995, 1996-1998, and 2011-2013.
    - Interval-censored observations for diabetes and hypertension.
- Reviews of hospital records.
    - Right-censored observations on MI, stroke, and death.

Application to ARIC Study

- Covariates: cohort×race, sex, and five baseline risk factors: age, body mass index (BMI), glucose level, systolic blood pressure, and smoking status.
- 8,728 subjects without prevalence cases or missing covariates included in this analysis.

Table: Distribution of Observations for the Events in the ARIC Study

Event	Incidence Case	Non-case During Follow-up	No Information
Diabetes	1508 (17.3%)	6771 (77.6%)	449 (5.1%)
Hypertension	4081 (46.8%)	4202 (48.1%)	445 (5.1%)

Event	Incidence Case	Non-case
MI	726 (8.3%)	8002 (91.7%)
Stroke	445 (5.1%)	8283 (94.9%)
Death	2503 (28.7%)	6225 (71.3%)

Application to ARIC Study

Table: Estimation Results for the Random Effects in the ARIC Study

Parameter	Estimate	Std error	p-value
γ_{MI}	0.7145	0.1258	<0.0001
γ_{Stroke}	0.9045	0.1450	<0.0001
γ_{Death}	0.7184	0.1026	<0.0001
σ_1^2	0.5801	0.1215	<0.0001
σ_2^2	1.1465	0.1165	<0.0001

Application to ARIC Study

Table: Estimation Results for the Regression Parameters of the Asymptomatic Events in the ARIC Study

	Diabetes			Hypertension		
Covariate	Estimate	Std error	*p*-value	Estimate	Std error	*p*-value
Forsyth County, white	−0.5332	0.1817	0.0033	−0.5032	0.0615	<0.0001
Jackson, black	−0.1356	0.1806	0.4530	−0.1075	0.0673	0.1104
Minneapolis, white	−0.9415	0.1802	<0.0001	−0.5747	0.0579	<0.0001
Washington County, white	−0.3778	0.1778	0.0336	−0.3798	0.0592	<0.0001
Age	−0.0093	0.0057	0.1025	0.0166	0.0036	<0.0001
Male	−0.0655	0.0593	0.2694	−0.2329	0.0396	<0.0001
BMI	0.0911	0.0059	<0.0001	0.0254	0.0044	<0.0001
Glucose	0.1075	0.0033	<0.0001	0.0004	0.0023	0.8744
Systolic blood pressure	0.0096	0.0026	0.0003	0.0780	0.0022	<0.0001
Smoker	0.4576	0.0674	<0.0001	0.3134	0.0468	<0.0001

NOTE: The blacks in Forsyth County form the reference group for the cohort × race variables.

Application to ARIC Study

Table: Estimation Results for the Regression Parameters of the Symptomatic Events in the ARIC Study

	MI			Stroke		
Covariate	Estimate	Std error	*p*-value	Estimate	Std error	*p*-value
Forsyth County, white	0.0467	0.2477	0.8504	0.1308	0.3688	0.7228
Jackson, black	−0.3121	0.2681	0.2444	0.6622	0.3755	0.0778
Minneapolis, white	−0.1052	0.2476	0.6710	0.0507	0.3688	0.8907
Washington County, white	0.1953	0.2457	0.4266	0.5013	0.3653	0.1700
Age	0.0805	0.0078	<0.0001	0.1121	0.0099	<0.0001
Male	0.9279	0.0901	<0.0001	0.4050	0.1071	0.0002
BMI	0.0273	0.0101	0.0068	−0.0010	0.0123	0.9356
Glucose	0.0059	0.0046	0.2007	0.0215	0.0057	0.0002
Systolic blood pressure	0.0135	0.0036	0.0002	0.0192	0.0047	<0.0001
Smoker	1.2378	0.0888	<0.0001	1.0023	0.1127	<0.0001

NOTE: The blacks in Forsyth County form the reference group for the cohort × race variables.

Application to ARIC Study: Prediction with Risk Scores

- Evaluate the performance of the proposed risk score and compare it with the risk score from the proportional hazards model.

- Training set.
  - Fit the proposed model to obtain parameter estimators.
  - Fit the proportional hazards model to obtain parameter estimators $\widehat{\boldsymbol{\beta}}_{PH}$.

- Testing set.
  - Calculate the risk scores of MI, stroke, and death based on the event history at examinations 2, 3, and 4.
  - Calculate the risk score $\widehat{\boldsymbol{\beta}}_{PH}^{\mathrm{T}} X$.

- Evaluate the performance of the prediction using C-index.

Application to ARIC Study: Prediction with Risk Scores

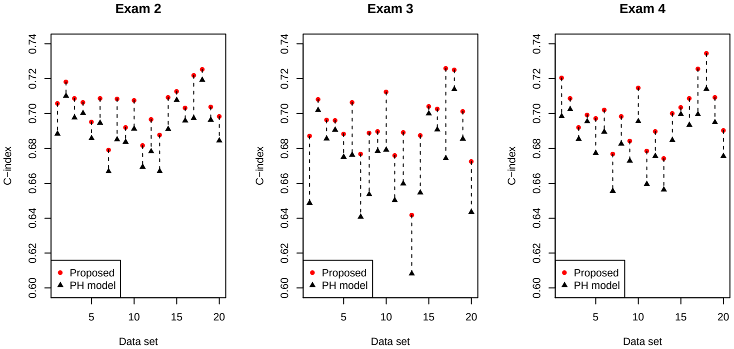

Figure: Plot of the estimates of the C-index of MI at each examination in the ARIC study.

Application to ARIC Study: Prediction with Cumulative Incidence Function

- Cumulative incidence functions of stroke given different event histories (baseline, year 3, and year 6).
- For comparison, we estimated cumulative incidence function of stroke under the univariate model of Fine and Gray (1999).
  - Does not condition on the event history and thus reflects the population average.

Application to ARIC Study: Prediction with Cumulative Incidence Function

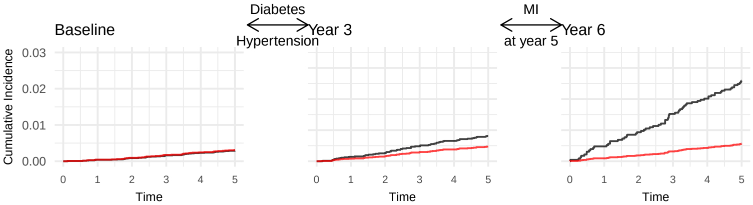

Figure: Estimation of the cumulative incidence of stroke for a 50-year-old white female smoker residing in Forsyth County, NC, with BMI 40 kg/m^2, glucose 98 mg/dl, and systolic blood pressure 113 mmHg: black curves pertain to proposed model with MI developed at year 5 and diabetes and hypertension developed between baseline and year 3; red curves pertain to Fine and Gray model.

Skin Cancer Trial

- We considered a cancer chemoprevention trial designed to evaluate the effectiveness of difluoromethylornithine in reducing the recurrence of skin cancer in patients with a history of non-melanoma skin cancer (Bailey et al., 2010).

- A total of 143 patients was randomly assigned to difluoromethylornithine, and 147 to placebo.

- The patients were scheduled to be examined every 6 months for the development of two types of non-melanoma skin cancer: basal cell carcinoma and squamous cell carcinoma.

Analysis specifics

- The actual examination times varied among patients, and the number of examinations ranged from 1 to 17.

- The number of new basal cell tumors ranged from 0 to 16, and the number of new squamous cell tumors ranged from 0 to 23.

- Covariates in the model include the treatment indicator for difluoromethylornithine, the number of prior skin tumors at baseline, gender, and age at diagnosis dichotomized as ≥ 65 versus < 65 years, and the random effects consist of scalar b_1 and b_2 and ξ.

- The EM algorithm converged within 300 iterations.

Result from data analysis

	Basal cell carcinoma			Squamous cell carcinoma			Any cancer		
	Est	St error	*p*-value	Est	St error	*p*-value	Est	St error	*p*-value
Treatment	-0.168	0.183	0.359	-0.146	0.265	0.582	-0.121	0.149	0.415
Prior tumors	0.104	0.013	< 0.001	0.109	0.016	< 0.001	0.108	0.007	< 0.001
Male	0.120	0.178	0.498	0.635	0.262	0.015	0.255	0.151	0.090
Age $\geq$ 65	-0.147	0.187	0.433	0.852	0.284	0.003	0.188	0.158	0.236

Findings from data analysis

- Treatment reduced the risk of both types of skin cancer, although not statistically significant. The number of prior tumors was positively associated with the risk of both types of cancer. Males and older patients had higher risk of squamous cell carcinoma. Gender and age were not significantly associated with basal cell carcinoma.
- The variance of the type-specific random effect indicates strong correlations for the recurrence of the same type of cancer over time.
- The variance of the shared random effect was estimated at 0.128, with standard error of 0.192, suggesting a relatively weak dependence between the two types of cancer.
- Bailey et al. (2010) reported a non-significant treatment effect on reducing squamous cell carcinoma but a significant treatment effect on basal cell carcinoma.

Prediction from data analysis

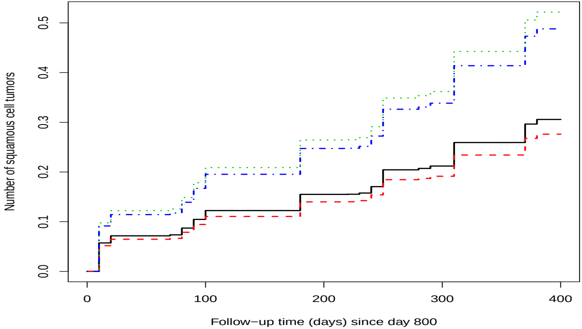

Figure: Prediction of squamous cell carcinoma based on the event history by day 800 for a patient older than 65 who is treated by difluoromethylornithine: the solid, dashed, dotted, and dot-dashed curves correspond to (1) no prior tumor at baseline but two basal cell tumors and one squamous cell tumor by day 800, (2) no prior tumor at baseline but one squamous cell tumor by day 800, (3) ten prior tumors at baseline and two new basal cell tumors and one new squamous cell tumor by day 800, and (4) ten prior tumors at baseline and one new squamous cell tumor by day 800.

Remarks

Concluding remarks

- We have presented a unified framework for semiparametric regression analysis of univariate and multivariate interval censored events and panel count data.

- Joint modelling can be used when the examination times are informative, e.g., they depend on the patient's health status.

- Competing risk events such as death can be incorporated into the data analysis.

- Extension to incomplete panel count data, e.g., the count is missing, is possible.