Maximum Likelihood Estimation for Semiparametric Regression Models with Interval-Censored Multi-State Data

Yu Gu

Department of Biostatistics
University of North Carolina at Chapel Hill

March 20, 2023

Acknowledgments

- Dr. Donglin Zeng
- Dr. Gerardo Heiss
- Dr. Danyu Lin

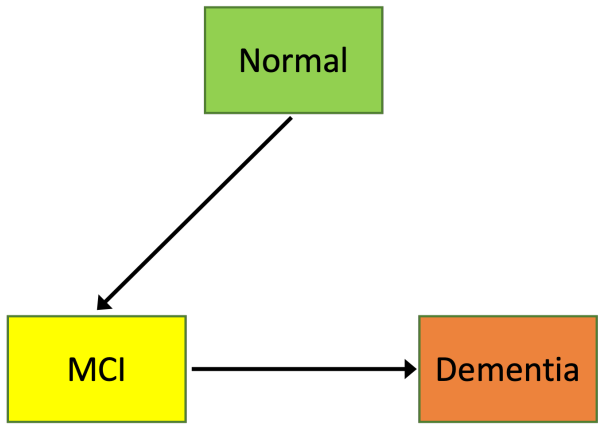
Gu, Y., Zeng, D., Heiss, G., Lin, D. Y. Maximum Likelihood Estimation for Semiparametric Regression Models with Interval-Censored Multi-State Data. https://arxiv.org/abs/2209.07708.

Outline

Outline

Multi-State Data

- Multi-state data arise frequently in studies of chronic diseases (e.g., dementia, cancer).
- Health status can be characterized by a finite number of disease states.
- ***Transition***: change from one state to another



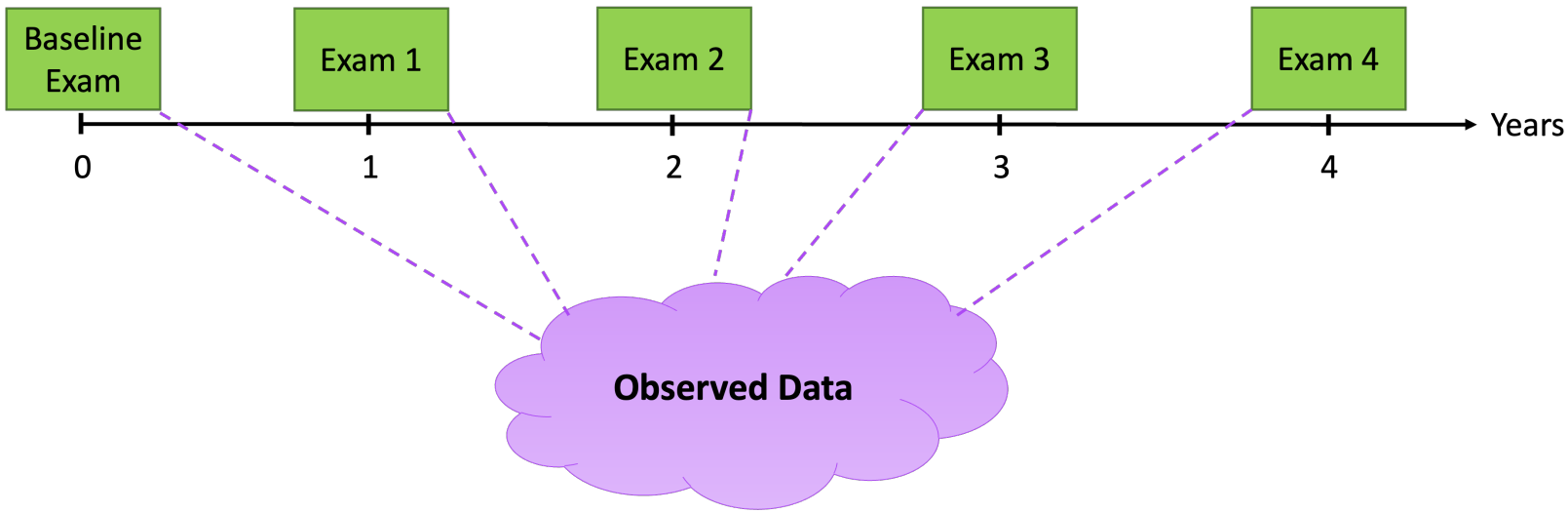
Significance

Analysis of multi-state data allows us to

- understand how a subject transitions from one state to another over time
- study the associations between risk factors and the disease process
- predict disease progression over time

Interval Censoring

- For economic and logistical reasons, subjects can only be examined periodically, such that transitions are only known to occur between two successive examinations.



- Such data are called ***interval-censored multi-state data***.

Analysis Challenges

- None of the transition times are directly observed.
- Trajectory of transitions from one examination to the next is unknown.
- Dependence among transitions of the same subject.

Existing Methods

Nonhomogeneous

Time-homogeneous Markov models

- Kalbfleisch & Lawless (1985); Satten (1999); Cook et al. (2002, 2004)
- simple; parametric; implemented in the msm package
- homogeneous assumption is unrealistic

Piecewise constant transition intensities

Nonparametric

- Gentleman et al. (1994); Saint-Pierre et al. (2003); Jackson (2011); Lawless & Nazeri Rad (2015)
- relatively simple; parametric
- restrictive; sensitive to the choice of change points

Spline-based intensities + penalized likelihood

- Joly & Commenges (1999); Machado & van den Hout (2018); Machado et al. (2021)
- more flexible; semiparametric
- tuning parameters (e.g., knots); inconsistent estimators

Overview of this work

- We provide a new framework to study semiparametric regression models for general interval-censored multi-state data.
- Our models use random effects to capture the dependence among transitions and accommodate time-dependent covariates.
- We adopt ***nonparametric maximum likelihood estimation*** (NPMLE) for inference.
- We devise a stable EM algorithm to compute the estimators.

Outline

Semiparametric Regression Model

We specify the following proportional intensity model:

$$\lambda_{ijk}(t; \boldsymbol{X}_i, \boldsymbol{b}_i) = \lambda_{jk}(t) \exp\left\{\beta_{jk}^{\mathsf{T}} \boldsymbol{X}_i(t) + \boldsymbol{b}_i^{\mathsf{T}} \boldsymbol{Z}_i(t)\right\} \tag{1}$$

- $\lambda_{jk}(\cdot)$: arbitrary baseline intensity function
- β_{jk}: unknown regression parameters
- $\boldsymbol{X}_i(\cdot)$: potentially time-dependent covariates
- $\boldsymbol{b}_i \sim N_d(\boldsymbol{0}, \boldsymbol{\Sigma}(\gamma))$: random effects
- $\boldsymbol{Z}_i(\cdot)$: consists of 1 and covariates that are part of $\boldsymbol{X}_i(\cdot)$

Observed Data

- Examination times: $(U_{i0}, U_{i1}, \ldots, U_{iM_i})$
- Observed states: $(S_{i0}, S_{i1}, \ldots, S_{iM_i})$
- Covariates: $\boldsymbol{X}_i(t)$

Likelihood

Under the noninformative censoring and conditional Markov assumptions, the likelihood is proportional to

$$\prod_{i=1}^{n} \int_{\boldsymbol{b}_i} \prod_{l=1}^{M_i} \boldsymbol{P}(U_{i,l-1}, U_{il}; \boldsymbol{X}_i, \boldsymbol{b}_i)^{(S_{i,l-1}, S_{il})} \phi(\boldsymbol{b}_i; \gamma) d\boldsymbol{b}_i \tag{2}$$

- $\boldsymbol{P}(u, v; \boldsymbol{X}_i, \boldsymbol{b}_i)$ denotes the transition probability matrix between times u and v for the ith subject.
- $\phi(\boldsymbol{b}; \gamma)$ denotes the density function of $N_d(\boldsymbol{0}, \boldsymbol{\Sigma}(\gamma))$.

Compute Transition Probability

- The transition probability matrix is given by

$$\boldsymbol{P}(u, v; \boldsymbol{X}_i, \boldsymbol{b}_i) = \prod_{u<t\leqslant v} \{\boldsymbol{I}_K + d\boldsymbol{A}(t; \boldsymbol{X}_i, \boldsymbol{b}_i)\}$$

- $\boldsymbol{A}(t; \boldsymbol{X}_i, \boldsymbol{b}_i)$ is the cumulative transition intensity matrix, with

$$\boldsymbol{A}(t; \boldsymbol{X}_i, \boldsymbol{b}_i)^{(j,k)} = \int_0^t \exp\{\beta_{jk}^{\mathsf{T}} \boldsymbol{X}_i(s) + \boldsymbol{b}_i^{\mathsf{T}} \boldsymbol{Z}_i(s)\} d\Lambda_{jk}(s).$$

- $\Lambda_{jk}(t) = \int_0^t \lambda_{jk}(s) ds$.

NPMLE

- We treat Λ_{jk} as step functions with nonnegative jumps at all examination times.

- Then the transition probability matrix

$$\boldsymbol{P}_i(u,v) = \prod_{u<t\leqslant v} \{\boldsymbol{I}_K + d\boldsymbol{A}_i(t)\}$$

$$\Downarrow$$

$$\widetilde{\boldsymbol{P}}_i(u,v) = \prod_{u<t_q\leqslant v} \{\boldsymbol{I}_K + \delta\boldsymbol{A}_i(t_q)\}.$$

- $\delta\boldsymbol{A}_i(t_q)$ involves the jump sizes λ_{jkq}, which are high-dimensional and lack analytical expressions.

Poissonization

- We introduce independent latent Poisson random variables W_{ijkq} with means $\{\delta \boldsymbol{A}_i(t_q)\}^{(j,k)}$.

- The key fact is that the transition probability $\widetilde{\boldsymbol{P}}_i(u,v)^{(s_0,s_r)}$ is equal to the probability of the event

$$\bigcup_{\text{traj}(s_0,s_1,\ldots,s_r)} \Big\{ W_{ijkq} > 0 \text{ if there's a transition } j \to k \text{ at time } t_q \text{ and } W_{ijkq} = 0 \text{ otherwise} \Big\}.$$

- Thus, maximizing the original likelihood is tantamount to maximizing the likelihood arising from the events of W's.

EM Algorithm

- We can treat W_{ijkq}'s and $\boldsymbol{b}_i$'s as missing data and apply the EM algorithm for maximizing the likelihood function.
- E-step involves the conditional expectations of W_{ijkq}'s and functions of $\boldsymbol{b}_i$'s given the observed data, which all have explicit expressions.
- In the M-step, $\boldsymbol{\beta}_{jk}$'s can be updated using the one-step Newton-Raphson approach, and λ_{jkq}'s can be updated explicitly.
- Therefore, the EM algorithm is immune to the high-dimensional parameters in NPMLE.

Asymptotic Properties

- (Consistency) Under some regularity conditions,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \sum_{(j,k)\in\mathcal{D}} \|\widehat{\Lambda}_{jk} - \Lambda_{0jk}\|_\infty \xrightarrow{a.s.} 0.$$

- (Asymptotic normality) $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a multivariate normal vector with mean zero and a covariance matrix that attains the semiparametric efficiency bound.

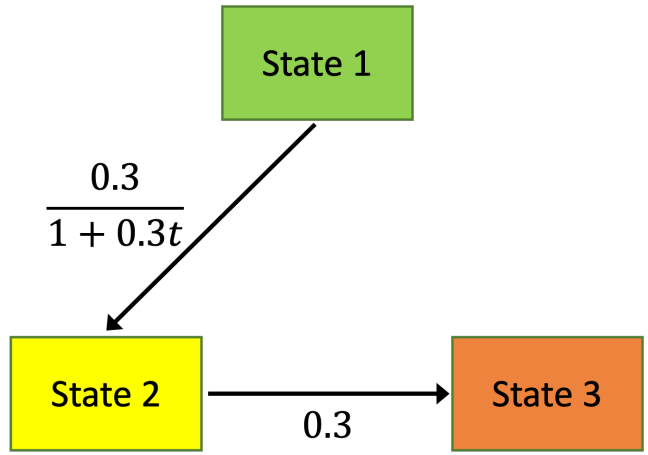
- (Variance estimation) The limiting covariance matrix of $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ can be consistently estimated by the inverse of

$$n^{-1} \sum_{i=1}^{n} \left\{ \nabla pl_i(\widehat{\boldsymbol{\theta}}) \right\}^{\otimes 2}.$$

Outline

Simulation Setting

- $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim \text{Unif}(0, 1)$
- $b \sim N(0, 0.8)$
- Six potential examination times separated by $0.05 + \text{Unif}(0, 1)$; no examinations beyond $\tau = 3$.



Simulation Results

Table 1: Estimation of regression parameters

		Proposed methods				**msm** package			
	Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$n = 400$	$\beta_{121} = 0.5$	0.014	0.265	0.259	95.0	−0.091	0.209	0.207	92.4
	$\beta_{122} = -0.5$	−0.021	0.458	0.448	94.7	0.087	0.363	0.356	94.0
	$\beta_{231} = 0.4$	0.013	0.206	0.198	94.5	−0.078	0.156	0.147	90.4
	$\beta_{232} = 0.2$	0.005	0.350	0.339	94.5	−0.053	0.268	0.254	92.8
	$\sigma^2 = 0.8$	0.060	0.422	0.396	95.1				
$n = 800$	$\beta_{121} = 0.5$	0.010	0.181	0.181	95.4	−0.092	0.145	0.146	90.4
	$\beta_{122} = -0.5$	−0.008	0.315	0.311	95.1	0.095	0.253	0.251	93.1
	$\beta_{231} = 0.4$	0.007	0.139	0.138	95.3	−0.079	0.107	0.104	87.4
	$\beta_{232} = 0.2$	0.006	0.240	0.236	94.6	−0.053	0.187	0.179	92.8
	$\sigma^2 = 0.8$	0.024	0.270	0.263	95.5				
$n = 1600$	$\beta_{121} = 0.5$	0.002	0.127	0.126	94.8	-0.096	0.103	0.103	84.8
	$\beta_{122} = -0.5$	−0.000	0.217	0.216	95.0	0.100	0.176	0.177	91.2
	$\beta_{231} = 0.4$	0.000	0.098	0.096	94.9	−0.080	0.076	0.073	79.7
	$\beta_{232} = 0.2$	−0.002	0.168	0.164	94.7	−0.057	0.132	0.126	91.3
	$\sigma^2 = 0.8$	−0.004	0.181	0.178	95.6				

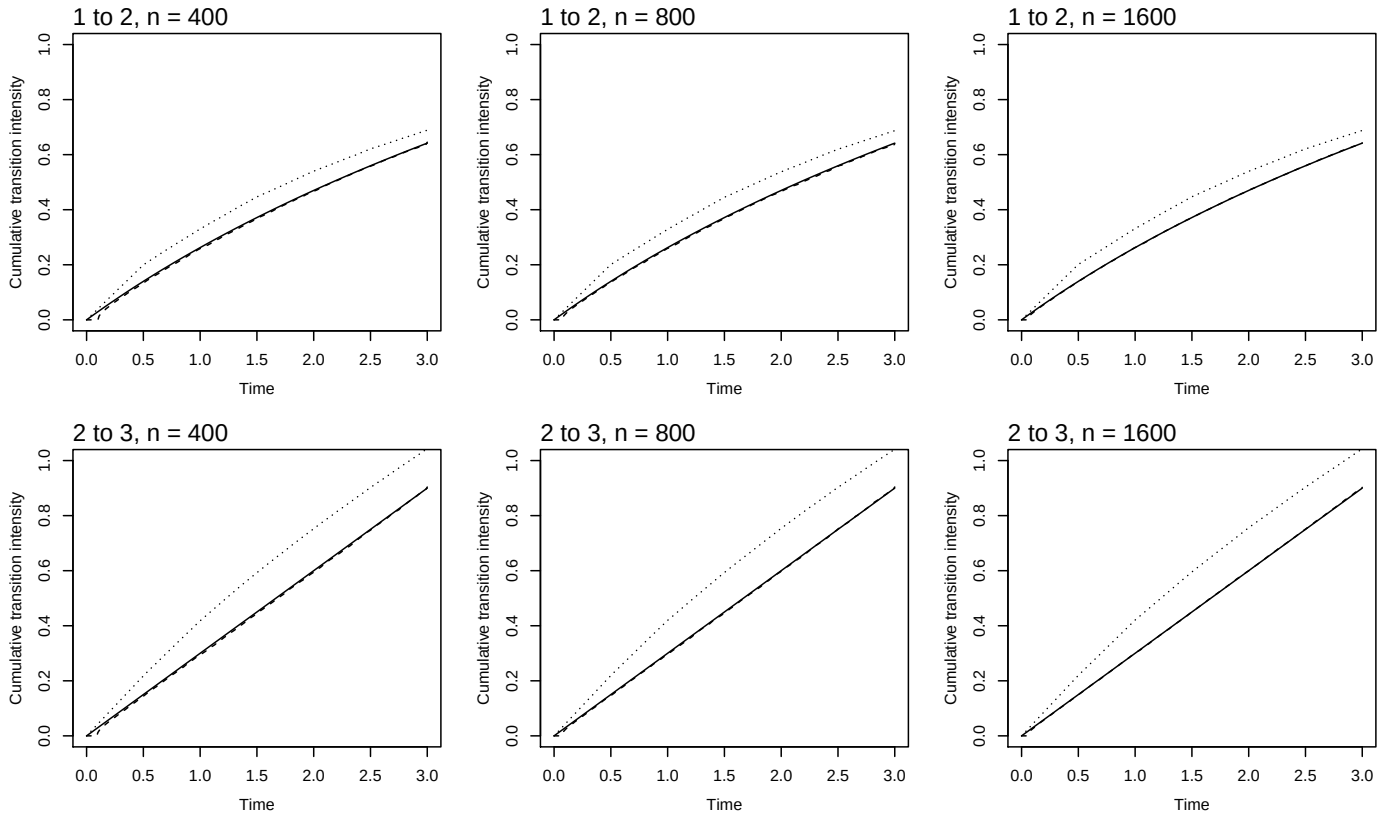


Figure 1: Estimation of cumulative baseline transition intensity functions.

Outline

ARIC Study

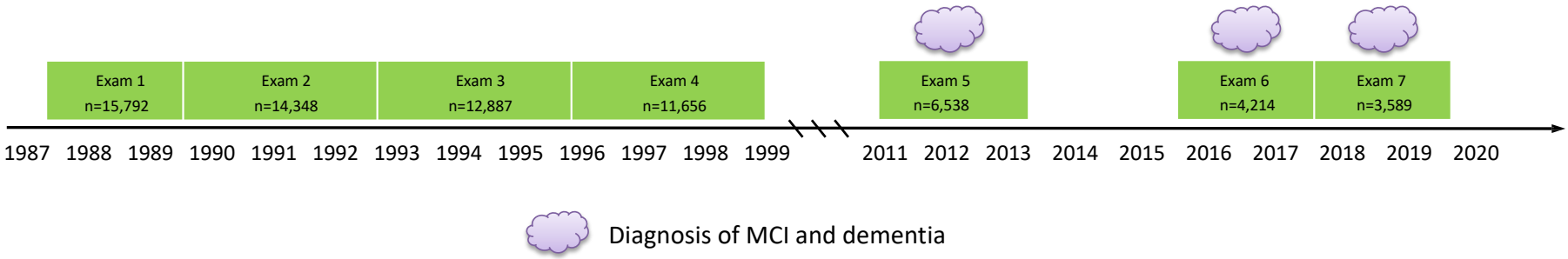


Figure 2: Timeline of the ARIC study.

- 6,407 participants remained after data cleaning
- Median follow-up time: 28.8 years since baseline examination

Analysis Results

Table 2: Estimation results on the regression parameters in the ARIC study.

	Normal to MCI			MCI to dementia		
Covariate	Estimate	St error	p-value	Estimate	St error	p-value
Age (years)	0.089	0.003	<0.001	0.111	0.006	<0.001
Male	0.319	0.052	<0.001	0.164	0.102	0.108
Advanced education	−0.100	0.053	0.056	−0.616	0.110	<0.001
Diabetes	0.559	0.099	<0.001	0.396	0.165	0.016
Smoker	0.155	0.066	0.019	0.201	0.136	0.138
Body mass index (kg/m^2)	0.021	0.005	<0.001	0.016	0.009	0.068
Systolic blood pressure (mmHg)	0.005	0.002	0.001	0.006	0.003	0.038
Black, Jackson	−0.001	0.079	0.992	1.469	0.161	<0.001
White, Minneapolis	−0.205	0.072	0.004	0.485	0.162	0.003
White, Washington County	−0.083	0.072	0.252	0.522	0.159	0.001

Key Findings

- Older people have significantly higher risk of developing both MCI and dementia.
- Advanced education can significantly reduce the risk of progression from MCI to dementia.
- Baseline diabetes occurrence, BMI, and systolic blood pressure are all positively associated with the risk of MCI.
- The variance of the random effect is estimated at 0.928 with a standard error estimator of 0.146, suggesting strong dependence among transitions.

Extension

- Our methods can be extended to allow a terminal event (e.g., death) that can be exactly observed or right-censored.
- Joint modelling can be used when time-dependent covariates are measured only at a finite number of time points.
- Joint analysis of multiple multi-state disease processes.

Thank you!