Analysis of survival data with cure fraction and variable selection: A pseudo-observations approach

2022 Missing data and survival analysis workshop

Presenter:
Sy Han (Steven) Chiou

Joint work with
Drs. Chien-Lin Su, Feng-Chang Lin, Robert W. Platt.

THE UNIVERSITY OF TEXAS AT DALLAS
School of Natural Sciences and Mathematics

Table of contents

Motivation

Survival data with a cure rate fraction

- We usually assume all subjects will eventually experience the event of interest if the follow-up period is sufficiently long.
- It is possible that some subjects may never experience the event in their lifetime $\rightarrow$ cured subject.

The melanoma data

- The melanoma dataset came from the Eastern Cooperative Oncology Group phase III clinical trial e1684 [Kirkwood et al., 1996].

```
> data(e1684, package = "smcure")
> head(e1684)
  TRT FAILTIME FAILCENS         AGE SEX
1   1  1.15068        1 -11.0359437   0
2   1  0.62466        1  -5.1290437   0
3   0  1.89863        0  23.1859563   1
4   0  0.45479        1  11.1448563   1
5   0  2.09041        1 -13.3208437   0
6   1  9.38356        0   0.9421563   0
```

- The event of interest is the relapse of melanoma.
- Treatment can completely cure melanoma, especially when it has not spread extensively.

The melanoma data

- The primary objective is to determine the effectiveness of the high dose interferon alpha-2b (IFN) regimen.
- Covariates of interests are treatment (1 = IFN), gender (1 = female), and age (centered to 0).
- After excluding missing data, the overall censoring rate is 30.9% out of the 284 remaining subjects.

The dental data

- The dental dataset [Calhoun et al., 2018] contains dental records from 5336 patients between August 2007 and March 2013.

```
> data(Teeth, package = "MST")
> dim(Teeth)

[1] 65228    56
```

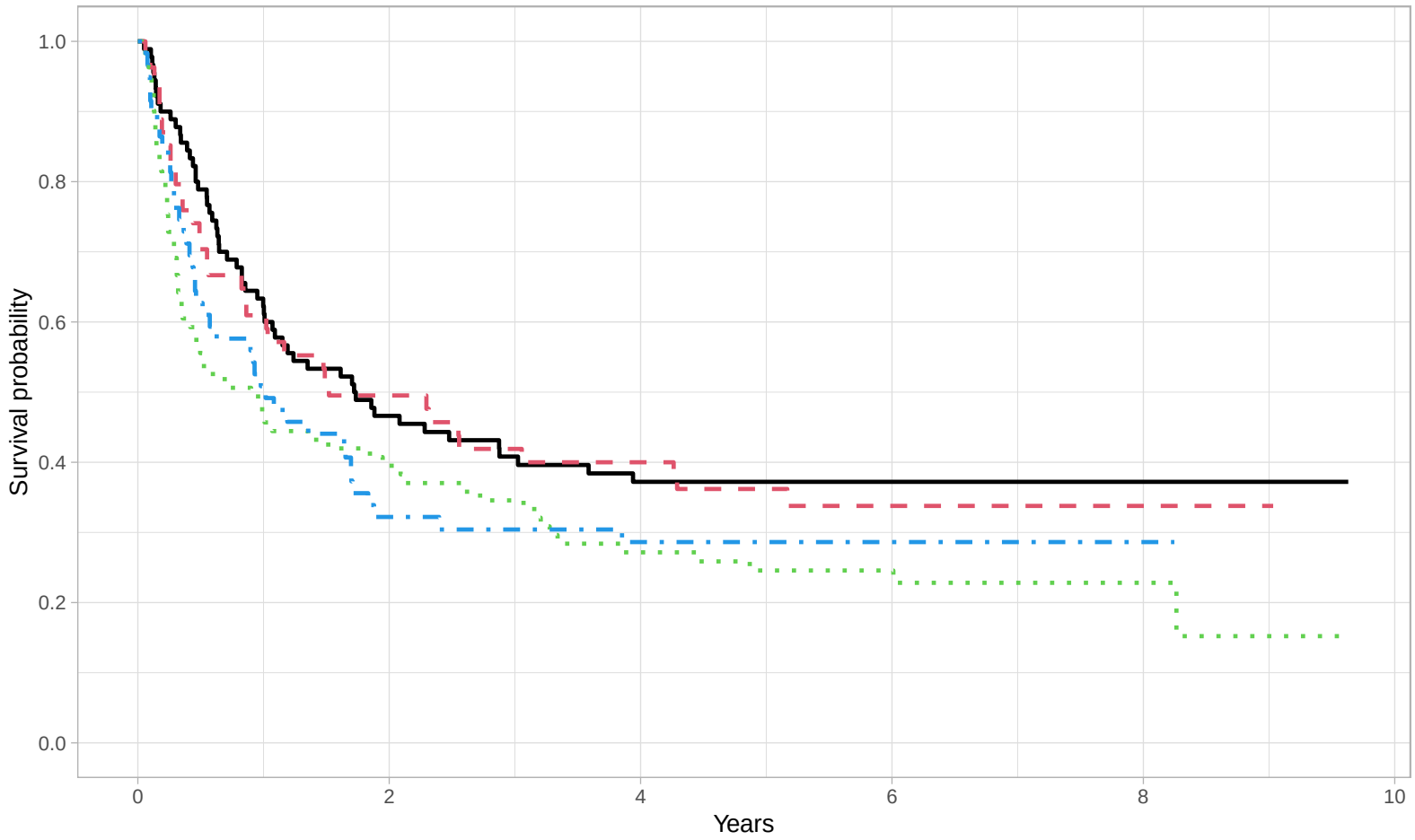
- The outcome of interest is the time to the first tooth loss.
- The overall censoring rate is 74.1%.
- There is a total of 44 risk factors, including tooth-level factors, subject-level factors, demographic factors, and health factors.

Survival data with a cure rate fraction

- Models do not account for the cure fraction and could lead to biased estimates of the survival of the uncured subjects [Peng and Yu, 2021].
- In general, it is difficult to identify the cured subjects, but their presence is signaled by a leveling of the Kaplan-Meier (KM) survival curve at the end of the follow-up.

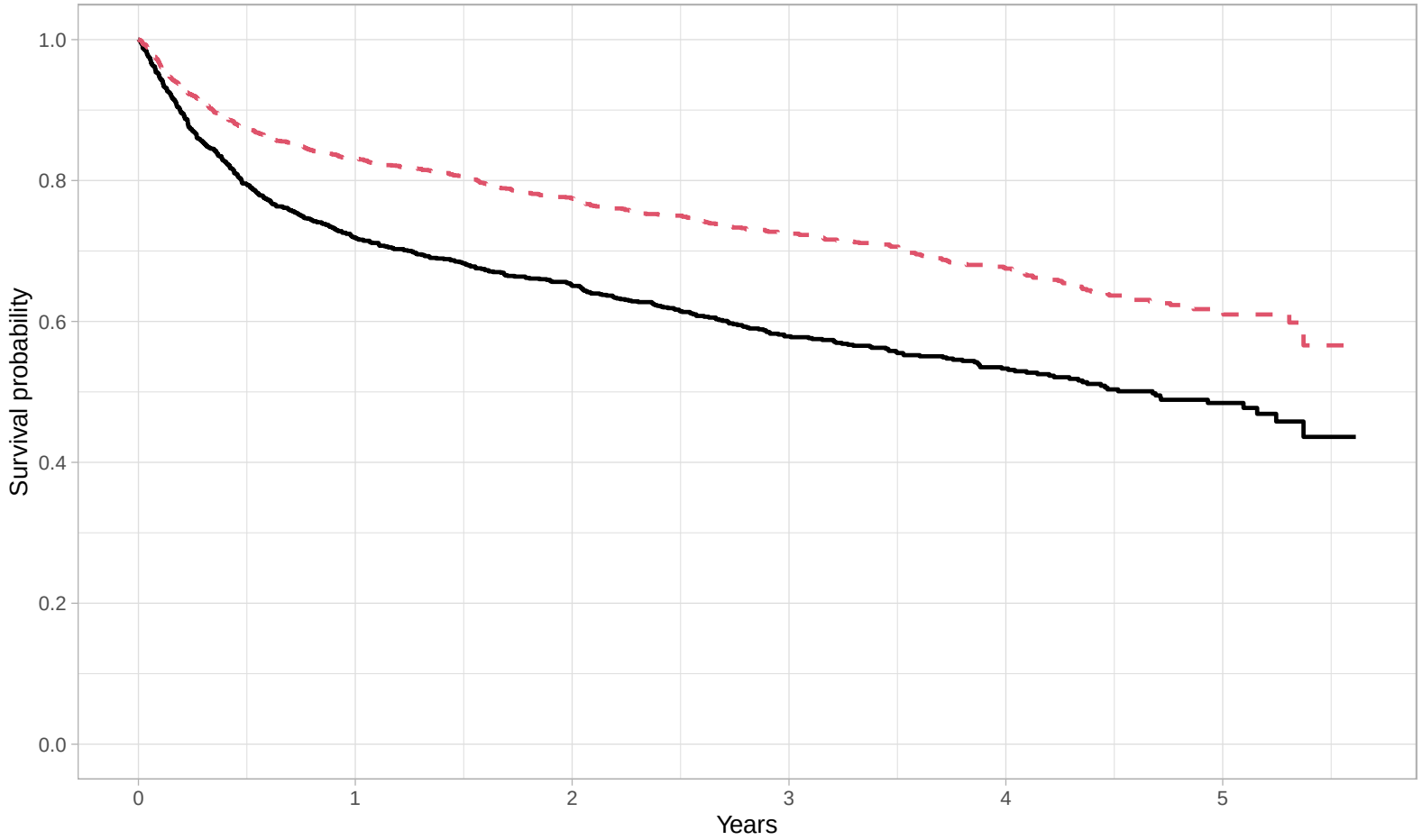
Survival data with a cure rate fraction

- The KM survival curve for the melanoma data.
- —— Treatment/Male; - - - Treatment/Female; · · · Placebo/Male; · · · Placebo/Female.



Survival data with a cure rate fraction

- The KM survival curve for the tooth data.
- —— Non-molar; - - Molar.



Survival data with a cure rate fraction

- Maller and Zhou [1992] proposed a nonparametric test to assess the existence of a cure fraction.
- The test is implemented in the `npcure::testmz()` function from the **npcure** package [de Ullibarri et al., 2020].
- The Maller-Zhou test confirms the observations on the KM plot with p-values < 0.001.

Challenges

- Two types of cure models have been popular in the literature;
  - Mixture cure model, e.g., Farewell [1982], Peng and Dear [2000], Li and Taylor [2002].
  - Bounded cumulative hazard (BCH) model (also known as the promotion time cure model), e.g., Yakovlev et al. [1993], Tsodikov [2002], Sposto [2002].
- EM-based approaches are usually computationally demanding.
- Even if the cure fraction is accounted for, the dental study posts additional challenges on high-dimensionality.
- We will focus on the mixture cure model in this presentation, but extension to the BCH model is available in our manuscript.

Notations and cure models

Notations

- Let Y denote the cure status; $Y = 1$ if the subject eventually experiences an event (uncured).
- The survival time of a subject can be expressed as

$$T = YT^* + (1 - Y) \times \infty,$$

  where $T^* < \infty$ is the failure time if the subject is uncured.
- Let C be the censoring time, $\delta = I(T \leq C)$ be the censoring status, and $\widetilde{T} = \min(T, C)$ be the observed survival time.
- We note that $\delta = 1 \rightarrow Y = 1$ but Y is not observed when $\delta = 0$.

The mixture cure model

- The mixture cure (MC) model expresses the population survival function as

$$S(t) = (1 - \pi) + \pi S_u(t),$$

  where $\pi = P(Y = 1)$ is the uncured rate, and $S_u(t)$ is the conditional survival function of T given $Y = 1$.
- The MC model consists of two components;
  - an incidence component that models π.
  - a latency component that models $S_u(t)$.

The mixture cure model

- The incidence component, π, is usually assumed to follow a logistic regression model

$$\pi(X) = P(Y = 1|\boldsymbol{X}) = \frac{\exp(\alpha_0 + \boldsymbol{X}^\top \boldsymbol{\alpha})}{1 + \exp(\alpha_0 + \boldsymbol{X}^\top \boldsymbol{\alpha})},$$

  where α_0 is a scalar and $\boldsymbol{\alpha}$ is a p-dimensional vector.
- Other link functions can also be applied to the incidence part.
  - log-log link: $\log[-\log\{1 - \pi(X)\}] = \boldsymbol{X}^\top \boldsymbol{\alpha}$.
  - probit link: $\Phi^{-1}\{\pi(X)\} = \boldsymbol{X}^\top \boldsymbol{\alpha}$.

The mixture cure model

- The latency component, $S_u(t)$, can be modeled with a Cox model

$$h(t|Z) = h_0(t)e^{\boldsymbol{Z}^\top \beta}$$

- or an AFT model.

$$h(t|Z) = h_0(te^{\boldsymbol{Z}^\top \beta})e^{\boldsymbol{Z}^\top \beta},$$

where β is a q-dimensional vector.

Pseudo-observations and cure models

Pseudo-observations

- Pseudo-observations was first proposed by Andersen et al. [2003] to model the transition probabilities in multi-state models.
- Since then, the pseudo-observations approach has been applied to many settings in survival analysis;
  - survival estimates [Klein et al., 2007],
  - restricted mean survival times [Andrei and Murray, 2007],
  - cumulative incidence function [Nicolaie et al., 2013],
  - relative survival function [Pavlič and Perme, 2019],
  - illness-death model with interval-censored data [Sabathé et al., 2020],
  - causal inference for recurrent event data [Su et al., 2020].
- However, the pseudo-observations approach has not been applied to the analysis of survival data with a cure fraction.

Pseudo-observations

- The concept is to create pseudo values for the quantities of interest using the analogy of leave-one-out cross-validation.
- These pseudo values are then treated as complete data where standard methods can be conveniently applied.

Pseudo-observations

- To illustrate the idea, let V_i be independent and identically distributed random variables, $\boldsymbol{X}_i$ be a vector of covariates, and suppose the interest lies on modeling $\mathrm{E}\{f(V_i)|\boldsymbol{X}_i\}$.
- Suppose not all of $f(V_i)$ are observed, the pseudo-observation of $f(V_i)$ can be constructed by

$$\widehat{\nu}^i = n\widehat{\nu} - (n-1)\widehat{\nu}^{(-i)},$$

  where $\widehat{\nu}$ is a consistent and unbiased estimator of $f(V_i)$ and $\widehat{\nu}^{(-i)}$ is the estimator without the ith subject.
- Large sample properties have also been investigated [e.g., Jacobsen and Martinussen, 2016, Overgaard et al., 2017].

Pseudo-observations for mixture cure model

- The MC model has two components; π and $S_u(t)$.
- Motivated by Maller and Zhou [1992], the uncured rate π can be estimated by $\widehat{\pi} = 1 - \widehat{S}(t_{\max})$ implying the pseudo observations

$$\widehat{\pi}^i = n\widehat{\pi} - (n-1)\widehat{\pi}^{(-i)},$$

where $\widehat{S}(\cdot)$ is the KM estimator and $t_{\max}$ is the maximum observed event time.

Pseudo-observations for mixture cure model

- The MC model implies that an consistent estimator for $S_u(t)$ is

$$\widehat{S}_u(t) = \frac{\widehat{S}(t) - \widehat{S}(t_{\max})}{1 - \widehat{S}(t_{\max})}.$$

- This implies the pseudo-observations for $S_u(t)$,

$$\widehat{S}_u^i(t) = n\widehat{S}_u(t) - (n-1)\widehat{S}_u^{(-i)}(t),$$

where $\widehat{S}_u^{(-i)}(t) = \{\widehat{S}^{(-i)}(t) - \widehat{S}^{(-i)}(t_{\max})\}/\{1 - \widehat{S}^{(-i)}(t_{\max})\}$ is the estimator for $S_u(t)$ without the ith sample.

Pseudo observations for mixture cure model

- To estimate the incidence parameters, we consider the following generalized linear model (GLM),

$$g_1\Big(E[Y_i|\boldsymbol{X}_i]\Big) = \alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{X}_i,$$

where $g_1(\cdot)$ is a link function, e.g, $g_1(x) = \log\{x/(1-x)\}$ corresponds to the logit link.

Pseudo observations for mixture cure model

- To estimate β, the pseudo observations for $S_u(t)$ are evaluated at $\boldsymbol{t} = \{t_1, \ldots, t_H\}$, a set of distinct times between 0 and $t_{\max}$.
- We can assume the GLM

$$g_2\Big(E[I(T_i^* > t_h)|\boldsymbol{Z}_i]\Big) = \xi_{t_h} + \beta^\top \boldsymbol{Z}_i,$$

  where ξ_{t_h} is the intercept at time t_h, β is the regression parameters, and g_2 is a link function.
- We set $g_2(x) = \log\{-\log(x)\}$ and $\xi_{t_h} = \log \Lambda_0(t_h)$, so the GLM model for β becomes the Cox model.

Pseudo observations for mixture cure model

- GEE approach is used to estimate the regression parameters in the two components.
- The GEE for the two components are:

$$\boldsymbol{U}(\alpha_0, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{\partial g_1^{-1}(\alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{X}_i)}{\partial(\alpha_0, \boldsymbol{\alpha})} \left\{ \widehat{\pi}^i - g_1^{-1}(\alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{X}_i) \right\} = 0,$$

$$\boldsymbol{S}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \frac{\partial \boldsymbol{g}_2^{-1}(\boldsymbol{t}, \boldsymbol{\psi}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\psi}} \boldsymbol{V}_i^{-1} \left\{ \hat{U}_u^i(\boldsymbol{t}) - \boldsymbol{g}_2^{-1}(\boldsymbol{t}, \boldsymbol{\psi}; \boldsymbol{Z}_i) \right\} = 0,$$

where $\boldsymbol{\psi} = (\xi_{t_1}, \ldots, \xi_{t_h}, \boldsymbol{\beta})$, $\hat{S}_u^i(\boldsymbol{t}) = \{\hat{S}_u^i(t_1), \ldots, \hat{S}_u^i(t_H)\}^\top$, V_i is the working variance.

Variance estimation and model diagnosis

- Solving GEE by the `geese()` function in *geepack* [Halekoh et al., 2006], which utilizes the jackknife variance estimates .
- We follow the idea of pseudo-residuals [Andersen and Perme, 2010] to assess the goodness-of-fit.
- The pseudo- residuals are defined as $\widehat{\pi}^i - g_1^{-1}(\widehat{\alpha}_0 + \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i)$ and $\widehat{S}_u^i(t) - g_2^{-1}(\widehat{\xi}_t + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_i)$ calculated at a given time $t \in \boldsymbol{t}$.
- If the model fits the data well, no trend should be perceptible when plotting residuals against a covariate.

Variable selection

- The proposed pseudo-observations approach allows variable selection and parameter estimation to be simultaneously implemented in each component.
- We consider penalizing the GEEs [Wang et al., 2012].

$$\boldsymbol{S}(\alpha_0, \boldsymbol{\alpha}) = \boldsymbol{U}(\alpha_0, \boldsymbol{\alpha}) - \boldsymbol{q}_\lambda(|\boldsymbol{\alpha}|) \circ \text{sign}(\boldsymbol{\alpha}),$$
$$\boldsymbol{S}(\boldsymbol{\psi}) = \boldsymbol{U}(\boldsymbol{\psi}) - \boldsymbol{q}_\lambda(|\boldsymbol{\beta}|) \circ \text{sign}(\boldsymbol{\beta}),$$

where $\boldsymbol{q}_\lambda(\cdot)$ is a vector of penalty functions for some tuning parameter λ and $\boldsymbol{u} \circ \boldsymbol{v}$ is the element-wise product of u and v.

Variable selection

- We select the tuning parameters via cross-validation.
- For a given tuning parameter λ, we calculate the overall cross-validated prediction error based on the pseudo-residuals.
- We use the estimates obtained from the unpenalized GEEs as the initial values.

Simulation

Simulation 1

- We generate the cure status from a logistic model

$$P(Y = 1|X) = \frac{\exp(\alpha_0 - X)}{1 - \exp(\alpha_0 - X)},$$

  where X is a Bernoulli(0.5) random variable.
- The intercept was chosen to be $\alpha_0 = 2.8, 2$ or 0.9 to achieve the average cure rates of 10%, 20%, and 40%.

Simulation 1

- The event times were generated from the proportional Cox model:

$$\lambda(t|Z) = \lambda_0(t)\exp(Z_1 + 0.5Z_2),$$

  where $\lambda_0(t) = 1/3$, Z_1 is a Bernoulli(0.5) random variable, and Z_2 is a uniform(0, 1) random variable.
- The censoring time is generated from Uniform(0, c), where c is chosen so the overall censoring rate is 10% more than the cure rate.
- The pseudo-observations for $S_u(t)$ are calculated at $0 < t_{10} < \ldots < t_{90} < t_{\max}$, where t_p is the pth quantile of the observed event times.
- We used exchangeable variance structure in the latency component.

Simulation 1

Table 1: Simulation results under simulation 1; ESE is the empirical standard error; ASE is the average standard error; CP is the coverage probability.

n		Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
		20% cure rate				40% cure rate			
200	α_0	0.040	0.431	0.453	95.4	-0.025	0.335	0.312	92.6
	α_1	-0.034	0.503	0.506	96.3	-0.001	0.425	0.411	95.6
	β_1	0.011	0.226	0.232	96.4	0.011	0.311	0.328	95.4
	β_2	0.005	0.379	0.381	96.5	0.005	0.558	0.556	94.6
400	α_0	0.019	0.327	0.324	95.6	-0.017	0.290	0.276	93.8
	α_1	-0.027	0.377	0.388	95.8	-0.017	0.339	0.331	95.4
	β_1	0.016	0.159	0.162	96.0	-0.010	0.228	0.239	94.5
	β_2	0.022	0.277	0.272	96.2	-0.014	0.492	0.442	94.2
1000	α_0	-0.010	0.216	0.211	94.9	-0.015	0.129	0.128	94.4
	α_1	0.007	0.201	0.199	95.2	0.016	0.168	0.169	94.8
	β_1	-0.001	0.093	0.104	96.4	-0.024	0.119	0.124	95.8
	β_2	-0.004	0.172	0.173	94.6	-0.023	0.234	0.229	94.8

Simulation 1

Table 2: Run time (in seconds) comparison with EM-based method implemented in the `smcure` package [Cai et al., 2012].

		Proposed method		
n	smcure	Incidence	Latency	Total
200	7.47	0.17	0.20	0.37
400	11.58	0.44	0.46	0.90
600	16.16	0.78	0.76	1.54
1000	30.03	1.75	1.64	3.39

Simulation 1

- Overall virtually unbiased, ESE$\approx$ASE, and reasonable CP.
- Discrete approximation to the baseline hazard function with a continuous time scale, e.g., number of time points in constructing pseudo-observations in $S_u(t)$.
- Our method is much faster than existing ones.

Simulation 2

- We expand Simulation 1 by considering $\boldsymbol{X} = (X_1, ..., X_{20})$.
  - X_1, X_2 are independently generated from Uniform(0,1)
  - X_3 and X_4 are independently generated from Bernoulli(0.5)
  - $X_5, \ldots, X_{20}$ are generated from a multivariate normal distribution with $\mathrm{E}(X_i) = 0$ and $\mathrm{Cov}(X_i, X_j) = 0.5^{|i-j|}$, for $i, j = 5, \ldots, 20$.
- The regression parameters were set at $\alpha_0 = 1.1$, $\boldsymbol{\alpha} = (0, 1, -1.2, 0, 0, -0.9, 0.8, 0, 0, ..., 0)^\top$, and $\boldsymbol{\beta} = (-0.7, 0, 1, 0, -0.5, 0.8, 0, 0, 0, ..., 0)^\top$.
- Those configurations yield a cure rate of 30%.

Simulation 2

Table 3: Simulation results under simulation 2; MSE is the mean squared error; TP is the true positive; FP is the false positive.

		Incidence			Latency		
n		MSE	TP	FP	MSE	TP	FP
200	Oracle	1.64	4	0	0.41	4	0
	SCAD	5.18	2.77	5.59	1.08	2.98	0.86
400	Oracle	0.75	4	0	0.23	4	0
	SCAD	1.83	3.42	0.55	0.40	3.69	0.47
1000	Oracle	0.28	4	0	0.08	4	0
	SCAD	0.44	3.93	0.30	0.22	3.88	0.21

Data analysis

The melanoma data

- The event of interest is the relapse of melanoma.
- $n = 284$, overall censoring rate = 30.9%, and covariates are treatment (1 = IFN), gender (1 = female), and age.

Table 4: The melanoma data; PE is the point estimator; SE is the standard error; p is the p-value.

	Incidence			Latency		
	PE	SE	p	PE	SE	p
Intercept	1.737	0.584	0.003	-	-	-
Treatment	-1.294	0.633	0.040	-0.009	0.299	0.975
Gender	-0.628	0.635	0.320	0.151	0.310	0.629
Age	0.024	0.013	0.066	-0.009	0.007	0.206
Treatment:Gender	0.903	0.755	0.228	-0.151	0.383	0.695

The melanoma data

- The treatment improves the cured rate,for males.
- The positive age effect indicates that older patients tend to have a higher relapse rate of melanoma.
- In the latency component, none of the four covariates are significantly associated with the failure time if patients are uncured.

The dental data

- The event of interest is time to the first tooth loss.
- $n = 5336$, overall censoring rate is 74.1%, and 44 risk factors.
- We considered the following penalties:
  - least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996]
  - adaptive LASSO (aLASSO) [Zou, 2006]
  - smoothly clipped absolute deviation (SCAD) penalty [Fan and Li, 2001].
- The tuning parameters are selected via five-fold cross-validation.

The dental data

Table 5: The dental data based on penalized GEE. Only the selected variables and their estimated coefficients are presented.

	LASSO		aLASSO		SCAD	
	Incidence	Latency	Incidence	Latency	Incidence	Latency
Molar	-0.327	-0.268	-0.392	-0.224	-0.618	-0.027
mobil	0.541	0.328	0.718	-	1.125	0.787
bleed	0.003	-	0.002	-	0.002	-
pocket_mean	0.428	-	0.417	-	0.746	-
cal_max	0.063	-	0.066	-	-	-
filled	-	-	-	-	1.372	-
decay_new	0.063	0.111	0.119	-	1.591	0.826
decay_rec	0.194	-	0.165	-	1.619	0.641
endo	1.091	-	1.088		2.074	-
filled tooth	-0.431	-0.066	-0.450	-	-0.690	-0.967
decayed tooth	0.524	-	0.577	-	-	-1.447

The dental data

Table 6: The dental data based on penalized GEE (cont'd).

	LASSO		aLASSO		SCAD	
	Incidence	Latency	Incidence	Latency	Incidence	Latency
bleed_ave	-	-	-	-	-	0.007
plaque_ave	-	-	-	-	-	-0.008
pocket_mean_ave	0.013	-	0.021	-	-	-
pocket_max_ave	0.017	-	0.023	-	-	-
decay_new_ave	0.788	-	-	0.493	1.286	1.457
decay_new_sum	0.097	-	0.087	-	0.427	-0.006
decayed_tooth_sum	0.044	-	0.049	-	-	-0.001
decayed_tooth_ave	-	-	-	-	-0.981	-
missing_tooth_ave	-	-	-	-	-1.683	0.771
total_tooth	0.002	-	-	-	-	-
age	0.013	-	0.013	-	-	0.004
gender	-0.051	-	-0.032	-	-	
diabetes	-	-	-	-	-	0.691
Tobacco use	0.171	-	0.159	-	-	0.422

The dental data

- The LASSO and aLASSO selected more variables in the incidence component while SCAD selected more variables in the latency component.
- Commonly selected incidence factors are molar, mobility score (mobil), bleeding on probing (bleed), periodontal probing depth (pocket mean), decayed surfaces new (decayed new), decayed surfaces recurrent (decayed recurrent), and endodontic therapy (endo), filled tooth (filled tooth)

Extension to the BCH model

Extension to the BCH model

- Suppose $\Lambda(t)$ is the cumulative hazard function of T^* such that $\Lambda(\infty) = \theta > 0$.
- Under the BCH model, the population survival function is

$$S(t) = \exp\{-\theta F(t)\},$$

  where $F(t) = \Lambda(t)/\theta$ is a proper cumulative distribution function of a nonnegative random variable with $F(0) = 0$ and $F(\infty) = 1$.
- The cure rate is indicated by $\lim_{t\to\infty} S(t) = e^{-\theta}$.
- The BCH model consists of two parts;
  - a long-term effect that models θ.
  - a short-term effect that models $F(t)$.

Extension to the BCH model

- To incorporate covariates effect in the BCH model, Tsodikov et al. [2003] proposed a PHPH model

$$S(t) = \exp[-\theta(\boldsymbol{X})\{1 - \bar{F}(t)^{\eta(\boldsymbol{Z})}\}],$$

  where $\theta(X) = \exp(\gamma_0 + \boldsymbol{\gamma}^\top X)$, $\eta(\boldsymbol{Z}) = \exp(\phi^\top Z)$, and ϕ is a $q \times 1$ vector of regression coefficients.
- We assume ϕ does not contain an intercept term to avoid overparameterization.

Extension to the BCH model

- Since the cure rate $\lim_{t\to\infty} S(t) = \exp(-\theta)$ can be nonparametrically estimated by $\widehat{S}(t_{\max})$, θ can be estimated by $\theta = -\log \widehat{S}(t_{\max})$.
- The pseudo-observations for θ is

$$\widehat{\theta}^i = n\widehat{\theta} - (n-1)\widehat{\theta}^{(-i)},$$

  where $\widehat{\theta} = -\log \widehat{S}(t_{\max})$.
- The pseudo-observations for $F(t)$ is

$$\widehat{F}^i(t) = n\widehat{F}(t) - (n-1)\widehat{F}^{(-i)}(t),$$

  where $\widehat{F}^{(-i)} = \log\{\widehat{S}(t)\}/\log\{\widehat{S}(t_{\max})\}$ [Tsodikov, 2002].

Conclusion

Conclusion

- We extended the pseudo-observation to modeling cure models.
- Advantages:
  - Faster than the usual EM-based approaches
  - Fits two components separately.
  - Variable selection.
- Future investigations
  - AFT model in the latency component (short term effect).
  - Post-selection inference.
  - An R package.

References

References

P. Andersen and M. P. Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99, 2010.

P. Andersen, J. Klein, and S. Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

A.-C. Andrei and S. Murray. Regression models for the mean of the quality-of-life-adjusted restricted survival time using pseudo-observations. *Biometrics*, 63(2):398–404, 2007.

C. Cai, Y. Zou, Y. Peng, and J. Zhang. smcure: An R-package for estimating semiparametric mixture cure models. *Comput Methods Programs Biomed*, 108(3):1255–1260, 2012.

P. Calhoun, X. Su, M. Nunn, and J. Fan. Constructing multivariate survival trees: The MST package for R. *Journal of Statistical Software*, 83(12):1–21, 2018. doi: 10.18637/jss.v083.i12.

I. L. de Ullibarri, A. López-Cheda, and M. A. Jácome. *npcure: Nonparametric Estimation in Mixture Cure Models*, 2020. URL `https://CRAN.R-project.org/package=npcure`. R package version 0.1-5.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

U. Halekoh, S. HÃ¸jsgaard, and J. Yan. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15/2:1–11, 2006.

M. Jacobsen and T. Martinussen. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862, 2016.

J. Kirkwood, M. Strawderman, M. Ernstoff, T. Smith, E. Borden, and R. Blum. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: The eastern cooperative oncology group trial est 1684. *J. Clin. Oncol.*, 14(1):7–17, 1996.

J. Klein, B. Logan, M. Harhoff, and P. Andersen. Analyzing survival curves at a fixed point in time. *Stat Med*, 26(24):4505–4519, 2007.

C.-S. Li and J. M. G. Taylor. A semi-parametric accelerated failure time cure model. *Stat Med*, 21 (21):3235–3247, 2002.

R. A. Maller and S. Zhou. Estimating the proportion of immunes in a censored sample. *Biometrika*, 79(4):731–739, 1992.

M. A. Nicolaie, J. C. van Houwelingen, T. M. de Witte, and H. Putter. Dynamic pseudo-observations: A robust approach to dynamic prediction in competing risks. *Biometrics*, 69(4):1043–1052, 2013.

M. Overgaard, E. Parner, and J. Pedersen. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *Ann Stat*, 45(5):1988–2015, 2017.

K. Pavlič and M. P. Perme. Using pseudo-observations for estimation in relative survival. *Biostatistics*, 20(3):384–399, 2019.

Y. Peng and K. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1): 237–243, 2000.

Y. Peng and B. Yu. *Cure Models: Methods, Applications, and Implementation*. CRC Press, 2021.

C. Sabathé, P. Andersen, C. Helmer, T. Gerds, H. Jacqmin-Gadda, and P. Joly. Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. *Statistical Methods in Medical Research*, 29(3):752–764, 2020.

R. Sposto. Cure model analysis in cancer: An application to data from the Children's Cancer Group. *Stat Med*, 21(2):293–312, 2002.

C.-L. Su, R. Platt, and J.-F. Plante. Causal inference for recurrent event data using pseudo-observations. *Biostatistics*, 2020.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A. Tsodikov. Semi-parametric models of long-and short-term survival: An application to the analysis of breast cancer survival in Utah by age and stage. *Stat Med*, 21(6):895–920, 2002.

A. Tsodikov, J. G. Ibrahim, and A. Y. Yakovlev. Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association*, 98 (464):1063–1078, 2003.

L. Wang, J. Zhou, and A. Qu. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360, 2012.

A. Yakovlev, B. Asselain, V. Bardou, A. Fourquet, T. Hoang, A. Rochefediere, and A. Tsodikov. A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et Analyse de Donnees SpatioTemporelles*, 12:66–82, 1993.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.