THE OHIO STATE UNIVERSITY

Identifying latent subgroups using regularized mixture cure models to develop risk stratification systems in cancer: Illustration using the hdcuremodels R package

Kellie J. Archer, Han Fu

archer.43@osu.edu

Division of Biostatistics
College of Public Health
The Ohio State University

THE OHIO STATE UNIVERSITY
COLLEGE OF PUBLIC HEALTH

Survival models

- The Cox proportional hazards (PH) model is the most frequently used method for assessing the effect of a covariate on a time-to-event outcome and assumes:
    - independence of time-to-event;
    - failure and censoring times are independent given covariates (non-informative censoring);
    - hazard ratio is constant over time; and
    - all subjects will experience the event of interest $[S(t = \infty) = 0]$.
- When modeling time to relapse, some patients may:
    - experience "cure" and therefore are "immune" to the event
    - whereas others are susceptible to the event.

Survival models

- The Cox model is known to yield inaccurate estimates of the hazard and survival when cured subjects are in the risk set [1].
- Recall, the Cox PH model assumes the same risk or hazard of an event throughout the entire follow-up period for all subjects. This is violated when cured subjects are in the dataset.
- Moreover, when a cured fraction exists in the sample being modeled but is ignored, the survival function $S(t)$ is no longer proper because $S(t = \infty) \neq 0$.
- Although we don't observe "cure," various methods have been proposed for estimating the cure fraction [2].
- Methods that account for the proportion cured separately from the latency of susceptibles (AKA, time-to-event among those susceptible - that is, those who will experience the event) are called mixture cure models.

Notation

- We let $i = 1, \ldots, N$ index subjects in our dataset.
- $T_i^\star$ is the time until the event of interest for subject i.
- If $T^\star$ is subject to right censoring, we observe $T_i = \min(T_i^\star, C_i)$ where C_i is the censoring time.
- Thus

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{otherwise.} \end{cases}$$

- Therefore we only observe T_i and δ_i, the time-to-event or time to last follow-up, and whether or not the event of interest occurred.

Notation

- If $\delta_i = 1$ we know the subject was uncured or susceptible.
- If $\delta_i = 0$ the subject either was cured or would have experienced the event if followed longer.
- Thus, we do not observe who is cured but they are usually censored observations, and those having long-term survival for which $P(T = \infty) = 1 - p$.
- Provided sufficiently long follow-up, the cure fraction, or probability of a cure, has been taken to be the Kaplan-Meier estimate of survival beyond the last observed event [3] or the limit of the cancer-specific survival distribution [4].

Mixture Models

- Consider that the population consists of K independent subgroups.
- Let $f(t)$ be a probability density function of the survival time T for the population.
- The finite mixture form is

$$f(t) = \sum_{k=1}^{K} p_k f_k(t)$$

where $p_1, \ldots, p_K$ represents the proportion of subjects in sub-population K where $\sum_{k=1}^{K} p_k = 1$ and $f_1(t), \ldots, f_K(t)$ are the K component densities.

Mixture Cure Models

- Let p_c = proportion "immunes" or cured subjects ($Y = 0$)
- Let p_u = proportion "susceptibles" or uncured subjects ($Y = 1$)
- $K = 2$ so $p_c + p_u = 1$.
- Let $S_u(t) = P(T > t|Y = 1)$ and $S_c(t) = P(T > t|Y = 0)$ then

$$S(t) = p_c S_c(t) + p_u S_u(t)$$

Notice $S_c(t) = P(T > t|Y = 0) = 1$

- Let $p = P(Y = 1)$ so that $(1 - p) = P(Y = 0)$ such that

$$S(t) = (1 - p) + pS_u(t)$$

- The latency or time-to-event $S_u(t)$ for susceptibles can be modeled using either
  - parametric [5, 6, 7];
  - non-parametric [8]; or
  - semi-parametric methods [9].

Mixture cure models (MCM)

- The mixture structure allows one to investigate the effect of covariates on two components of the model: incidence (susceptible versus cured) and latency (time-to-event for susceptibles).

$$S(t|\mathbf{x},\mathbf{w}) = (1 - p(\mathbf{x})) + p(\mathbf{x})S_u(t,\mathbf{w}|Y = 1)$$

- Example: The density of a Weibull distributed variable is $f(t) = \lambda\alpha(\lambda t)^{\alpha-1}\exp[-(\lambda t)^{\alpha}]$ where the shape α and scale λ parameters are both > 0.
- Letting the hazard depend on covariates, we replace λ^{α} with $\lambda^{\alpha}\exp(\boldsymbol{\beta}^{\top}\mathbf{w})$ such that

$$f(t|\mathbf{w}) = \lambda\alpha(\lambda t)^{\alpha-1}\exp(\boldsymbol{\beta}^{\top}\mathbf{w})\exp\left(-(\lambda t)^{\alpha}\exp(\boldsymbol{\beta}^{\top}\mathbf{w})\right).$$

Mixture cure models (MCM)

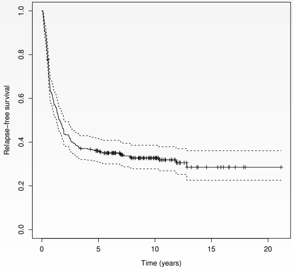
- The incidence component, $p(\mathbf{x})$, is usually modeled using logistic regression.
- We replace $p(\mathbf{x})$ with $\exp(b_0 + \mathbf{b}^\top \mathbf{x})/(1 + \exp(b_0 + \mathbf{b}^\top \mathbf{x}))$.
- The Weibull MCM likelihood consists of $p(\mathbf{x})$ and $f(t|\mathbf{w})$.
- The set of parameters for the Weibull MCM is $\boldsymbol{\theta} = (b_0, \mathbf{b}, \boldsymbol{\beta}, \alpha, \lambda)$.
- The exponential mixture cure model is a special case of the Weibull where $\alpha = 1$.
- $\boldsymbol{\theta} = (b_0, \mathbf{b}, \boldsymbol{\beta})$ for the semi-parametric mixture cure model.

Mixture cure models

- Mixture cure models assume:
  - independence of time-to-event;
  - failure and censoring times are independent given covariates (non-informative censoring);
  - some subjects will not experience the event of interest; and
  - there is sufficient follow-up time.
- Farewell (1982) [5] cautions that there should be strong scientific evidence that two sub-populations are present.
- Maller and Zhou (1996) [7] provided
  - an inferential procedure for assessing whether the proportion immunes > 0.
  - an inferential procedure for assessing whether there is sufficient follow-up.

hdcuremodels: Data example

```
> library(hdcuremodels)
> library(survival)
> dim(amltrain)
[1] 306 322
> dim(amltest)
[1]  40 322
> km <- survfit(Surv(cryr, relapse.death) ˜ 1, data = amltrain)
> plot(km, mark.time = TRUE, xlab = "Time (years)",
+  ylab = "Relapse-free survival")
```



hdcuremodels: Assessing MCM assumptions

```
> cure_estimate(km)
[1] 0.2853081
> nonzerocure_test(km)
$proportion_susceptible
[1] 0.7146919

$proportion_cured
[1] 0.2853081

$p.value
[1] "< 0.001"

$time_95_percent_of_events
[1] 5.294299
> sufficient_fu_test(km)
       p.value Nn   N
1 4.825325e-06 12 306
```

Penalized Mixture Cure Models*

```
> args(curegmifs)
function (formula, data, subset, x.latency = NULL, model = "weibull",
    penalty.factor.inc = NULL, penalty.factor.lat = NULL, epsilon = 0.001,
    thresh = 1e-05, scale = TRUE, maxit = 10000, inits = NULL,
    verbose = TRUE, ...)

> args(cureem)
function (formula, data, subset, x.latency = NULL, model = "cox",
    penalty = "lasso", penalty.factor.inc = NULL,
    penalty.factor.lat = NULL,
    thresh = 0.001, scale = TRUE, maxit = NULL, inits = NULL,
    lambda.inc = 0.1, lambda.lat = 0.1, gamma.inc = 3, gamma.lat = 3,
    ...)
```

* `cureem` uses the E-M algorithm [10]; curegmifs uses the GMIFS algorithm [11]

Penalized Mixture Cure Models

```
> coxem <-cureem(Surv(cryr, relapse.death) ~ ., data = amltrain,
+               x.latency = amltrain, model = "cox",
+               lambda.inc=0.009993, lambda.lat=0.02655)

> fitgmifs <- curegmifs(Surv(cryr, relapse.death) ~ ., data = amltrain,
+               x.latency = amltrain, model = "weibull", maxit = 20000)
```

Penalized Mixture Cure Models

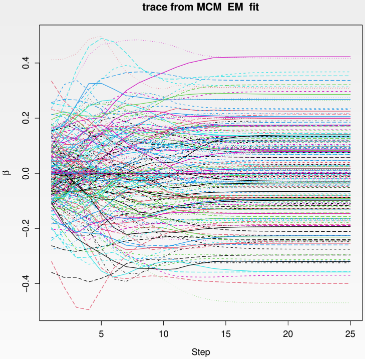
```
> args(cv_cureem)
function (formula, data, subset, x.latency = NULL, model = "cox",
    penalty = "lasso", penalty.factor.inc = NULL, penalty.factor.lat = NULL
    fdr.control = FALSE, fdr = 0.2, grid.tuning = FALSE, thresh = 0.001,
    scale = TRUE, maxit = NULL, inits = NULL, lambda.inc.list = NULL,
    lambda.lat.list = NULL, nlambda.inc = NULL, nlambda.lat = NULL,
    gamma.inc = 3, gamma.lat = 3, lambda.min.ratio.inc = 0.1,
    lambda.min.ratio.lat = 0.1, n_folds = 5, measure.inc = "c",
    one.se = FALSE, cure_cutoff = 5, parallel = FALSE, seed = NULL,
    verbose = TRUE, ...)

> args(cv_curegmifs)
function (formula, data, subset, x.latency = NULL, model = "weibull",
    penalty.factor.inc = NULL, penalty.factor.lat = NULL, fdr.control = FAL
    fdr = 0.2, epsilon = 0.001, thresh = 1e-05, scale = TRUE,
    maxit = 10000, inits = NULL, n_folds = 5, measure.inc = "c",
    one.se = FALSE, cure_cutoff = 5, parallel = FALSE, seed = NULL,
    verbose = TRUE, ...)
```

hdcuremodels: Generic functions

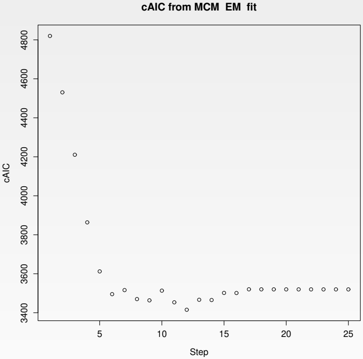
```
> print(coxem)
 [1] "b_path"      "beta_path"   "b0_path"     "logLik.inc"
 [5] "logLik.lat"  "x.incidence" "x.latency"   "y"
 [9] "model"       "scale"       "method"      "call"
[13] "cv"
> summary(coxem)
Mixture cure model fit using the EM algorithm
at step    =  25 logLik     =  -1113.552
at step    =  12 AIC        =  2634.476
at step    =  12 mAIC        =  5510.632
at step    =  12 cAIC        =  3415.284
at step    =  12 BIC        =  3382.917
at step    =  12 mBIC        =  5423.137
at step    =  12 EBIC        =  3777.815
```

hdcuremodels: Generic functions

```
> plot(coxem)
```



```
> plot(coxem, type = "cAIC")
```

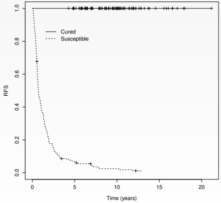


hdcuremodels: Generic functions

```
> coef.cAIC <- coef(coxem, model.select = "cAIC")
> coef.12 <- coef(coxem, model.select = 12)
> names(coef.cAIC)
[1] "b0"       "beta_inc" "beta_lat"
> coef.cAIC$b0
[1] 1.638612
> sum(coef.cAIC$beta_inc != 0)
[1] 112
> sum(coef.cAIC$beta_lat != 0)
[1] 88
```

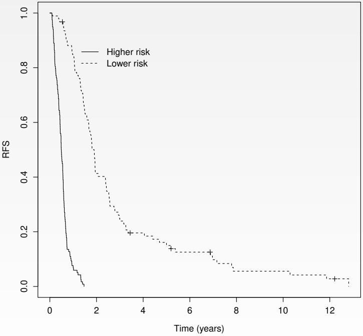
hdcuremodels: Generic functions

```
> train.predict <- predict(coxem, model.select = "cAIC")
> names(train.predict)
[1] "p.uncured"      "linear.latency" "latency.risk"
> head(train.predict$p.uncured)
[1] 0.95394997 0.89003059 0.94182718 0.76070954 0.18856553 0.04190272
> head(train.predict$linear.latency)
[1] -0.8706797 -0.4249528 -1.3281981 -1.3928568  0.9210771 -2.8833691
> head(train.predict$latency.risk)
[1] "low risk"  "low risk"  "low risk"  "low risk"  "high risk" "low risk"
> p_group <- ifelse(train.predict$p.uncured < 0.50, "Cured", "Susceptible")
> km.cured <- survfit(Surv(cryr, relapse.death) ~ p_group, data = amltrain)
> plot(km.cured, mark.time = TRUE, lty = c(1,2), xlab="Time (years)", ylab = "RFS")
> legend(c(.9, .1), legend = c("Cured", "Susceptible"), lty = c(1, 2), bty = "n")
```



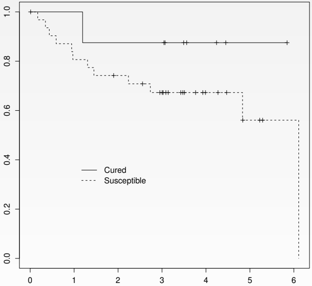
hdcuremodels: Generic functions

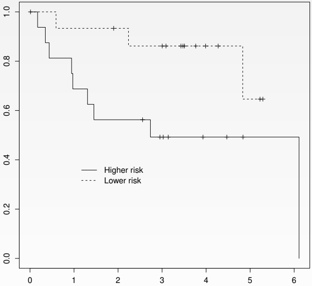
```
> km.suscept <- survfit(Surv(cryr, relapse.death) ~ train.predict$latency.risk,
+  data = amltrain, subset = (p_group == "Susceptible"))
> plot(km.suscept, mark.time = TRUE, lty = c(1,2), xlab = "Time (years)",
+ ylab="RFS")
> legend(c(.9, .1), legend = c("Higher risk", "Lower risk"), lty = c(1,2),
+ bty = "n")
```



hdcuremodels: Generic functions

```
> test.predict <- predict(coxem, newdata = amltest, model.select = "cAIC")
> test_p_group <- ifelse(test.predict$p.uncured < 0.50, "Cured", "Susceptible")
> km.cured.test <- survfit(Surv(cryr, relapse.death) ~ test_p_group, data = amltest)
> plot(km.cured.test, mark.time = TRUE, lty = c(1, 2))
> legend(c(.4, .1), legend = c("Cured", "Susceptible"), lty = c(1,2), bty = "n")
> km.suscept.test <- survfit(Surv(cryr, relapse.death) ~ test.predict$latency.risk, data = amltest,
+ subset = (test_p_group == "Susceptible"))
> plot(km.suscept.test, mark.time = TRUE, lty = c(1,2))
> legend(c(.4, .1), legend = c("Higher risk", "Lower risk"), lty = c(1, 2), bty = "n")
```





hdcuremodels: Generic functions

- The outcome for `cutoff` $= \tau$ is defined as

$$Y_i = \begin{cases} 0 \text{ if } T_i > \tau \\ 1 \text{ if } T_i \leq \tau \text{ and } \delta_i = 1 \\ \text{missing if } T_i \leq \tau \text{ and } \delta_i = 0 \end{cases} .$$

- The mean score imputation AUC lets $Y_i = 1 - \hat{p}(\mathbf{x}_i)$ for those subjects with a missing outcome [12].
- The C-statistic for MCMs was adapted to weight patients by their outcome (cured, susceptible, censored) [13].

```
> AUC(coxem, model.select = "cAIC")
[1] 0.9690409
> AUC(coxem, newdata = amltest, model.select = "cAIC")
[1] 0.8049214
> concordance_mcm(coxem, model.select = "cAIC")
[1] 0.8546535
> concordance_mcm(coxem, newdata = amltest, model.select = "cAIC")
[1] 0.6987875
```

Contributions⋆

- Our `hdcuremodels` R package can be used to model a censored time-to-event outcome in the presence of a cure fraction
- Our `hdcuremodels` R package can handle a high-dimensional covariate space
- Our `hdcuremodels` R package does not require the same variables to be included in the incidence and latency portions of the model
- Our `hdcuremodels` R package includes relevant functions for testing MCM assumptions
- As previously demonstrated [11], our GMIFS and E-M algorithms outperformed existing methods in terms of variable selection and prediction

⋆Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01 LM013879. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References I

Price D.L. and Manatunga A.K.
Modelling survival data with a cured fraction using frailty models
Statistics in Medicine, **20**:1515–1527, 2001.

Boag J.W.
Maximum likelihood estimates of the proportion of patients cured by cancer therapy
Journal of the Royal Statistical Society. Series B (Methodological), **11**:15–53, 1949.

Laska E.M. and Meisner M.J.
Nonparametric estimation and testing in a cure model
Biometrics, **48**:1223–1234, 1992.

Gamel J.W., McLean I.W., Rosenberg S.H.
Proportion cured and mean log survival time as functions of tumour size
Statistics in Medicine, **9**:999–1006, 1990.

References II

Farewell V.T.

The use of mixture models for the analysis of survival data with long-term survivors.

Biometrics **4**: 1041–1046, 1982.

Farewell V.T.

Mixture models in survival analysis: Are the worth the risk?

The Canadian Journal of Statistics **14**: 257–262, 1986.

Maller R.A. and Zhou X.

Survival Analysis with Long-Term Survivors, John Wiley & Sons, West Sussex, England, 1996.

Sposto R., Sather H.N., Baker S.A.

A comparison of tests of the difference in the proportion of patients who are cured.

Biometrics **48**: 87–99, 1992.

References III

- Kuk A.Y.C. and Chen C.-H.
  A mixture model combining logistic regression with proportional hazards regression.
  Biometrika **79**: 531–541, 1992.

- Archer K.J., Fu H., Mrózek K., Nicolet D., Mims A.S., Uy G.L., Stock W., Byrd J.C., Hiddemann W., Braess J., Spiekermann K., Metzeler K.H., Herold T., Eisfeld A.-K.
  Identifying long-term survivors and those at higher or lower risk of relapse among patients with cytogenetically normal acute myeloid leukemia using a high-dimensional mixture cure model.
  Journal of Hematology and Oncology **17**(1):28, 2024.

- Fu H., Nicolet D., Mrózek K., Stone R.M., Eisfeld A.-K., Byrd J.C., Archer K.J.
  Controlled variable selection in Weibull mixture cure models for high-dimensional data.
  Statistics in Medicine **41**(22):4340-4366, 2022.

References IV

Asano J., Hirakawa H., Hamada C.

Assessing the prediction accuracy of cure in the Cox proportional hazards cure model: an application to breast cancer data.

Pharmaceutical Statistics **13**: 357–363, 2014.

Asano J. and Hirakawa H.

Assessing the prediction accuracy of a cure model for censored survival data with long-term survivors: Application to breast cancer data.

Journal of Biopharmaceutical Statistics **27**: 918–932, 2017.