Regularized Cox Cure Rate Model with R package intsurv

Regularized Cox models with a cure rate are an important tool for analyzing survival data with heaving censoring and a large number of covaraites. The R Package **intsurv** (Wang et al., 2019) provides a collection of methods for integrative survival analyses with data from multiple sources. Function `cox_cure_net.fit()` in the package is an efficient implementation for regularized Cox cure rate model with elastic-net penalty (Zou and Hastie, 2005).

The cure rate models first proposed by Berkson and Gage (1952) are commonly adopted statistical methods for survival data with a cure fraction. Consider a random sample of n subjects with right-censoring data and a cured fraction. Let $T_j = \min(V_j, C_j)$ and $\Delta_j = I(V_j > C_j)$, where V_j and C_j represents the random variable of the event time and the censoring time of subject j, respectively, $I(\cdot)$ is indicator function, $j \in \{1, \ldots, n\}$. Define $Z_j = 1$ if subject j is susceptible, and $Z_j = 0$ otherwise, with probability $p_j = \Pr(Z_j = 1)$. Notice that Z_j is observed to be 1 if $\Delta_j = 1$ and is missing otherwise. Proposed by Farewell (1982), a logistic model $p_j = 1/[1+\exp(-\gamma_0 - \boldsymbol{x}_j^\top \boldsymbol{\gamma})]$ is widely used, where $\boldsymbol{x}_j$ represents the covariate vector of subject j (excluding intercept), γ_0 is unknown coefficient of intercept and $\boldsymbol{\gamma}$ is a vector of unknown covariate coefficients. Given that $Z_j = 1$, Kuk and Chen (1992) proposed modeling the conditional survival times through a Cox proportional hazard model with the hazard function

$$h_j(t \mid Z_j = 1) = h_0(t \mid Z_j = 1) \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}),$$

where $h_0(t \mid Z_j = 1)$ is an unspecified baseline function for events, and $\boldsymbol{\beta}$ is a vector of unknown coefficients of the covariate vector $\boldsymbol{x}_j$. The conditional survival function of the event time of subject j is

$$S_j(t \mid Z_j = 1) = \exp\{-H_0(t \mid Z_j = 1) \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})\},$$

where $H_0(t \mid Z_j = 1) = \int_0^t h_0(s \mid Z_j = 1) \mathrm{d}s$. Given that subject j is cured ($Z_j = 0$), the conditional survival function satisfies $S_j(t \mid Z_j = 0) = 1$, for $t < +\infty$. The observed data likelihood function can be written as

$$\begin{aligned} L(\boldsymbol{\theta}) = \prod_{j=1}^{n} &\{p_j h_j(t_j \mid Z_j = 1) S_j(t_j \mid Z_j = 1)\}^{\delta_j} \\ &\{(1 - p_j) + p_j S_j(t_j \mid Z_j = 1)\}^{1-\delta_j}, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \gamma_0, h_0(\cdot)\}$.

An estimation procedure based on the well-known EM algorithm was proposed by Sy and Taylor (2000). Recently, a few works have been proposed to perform variable selection for cure models. For example, Scolas et al. (2016) proposed variable selection with adaptive lasso penalty (Zou, 2006) for interval-censored data in a parametric cure model, where conditional survival times follow the extended generalized gamma distribution. Masud et al. (2018) proposed variable selection methods for mixture cure model and promotion cure model through regularization by the adaptive lasso penalty. Fan et al. (2017) and Shi et al. (2019) promoted structural similarity and sign consistency of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$, respectively, with minimax concave penalty (Zhang, 2010) for variable selection. Here, we concentrate on the following regularized estimator with elastic-net penalty,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} -\frac{1}{n}\ell(\boldsymbol{\theta}) + P_1(\boldsymbol{\beta}; \alpha_1, \lambda_1) + P_2(\boldsymbol{\gamma}; \alpha_2, \lambda_2), \tag{2}$$

where $\ell(\boldsymbol{\theta})$ is the log-likelihood function under the observed data from (1) and

$$P_1(\boldsymbol{\beta}; \alpha_1, \lambda_1) = \lambda_1 \left(\alpha_1 \sum_{k=1}^{p} \omega_k |\beta_k| + \frac{1-\alpha_1}{2} \sum_{k=1}^{p} \beta_k^2 \right),$$

$$P_2(\boldsymbol{\gamma}; \alpha_2, \lambda_2) = \lambda_2 \left(\alpha_2 \sum_{k=1}^{p} \nu_k |\gamma_k| + \frac{1-\alpha_2}{2} \sum_{k=1}^{p} \gamma_k^2 \right),$$

where ω_k and ν_k represent non-negative weights (Zou, 2006), $0 \le \alpha_1 \le 1$, $0 \le \alpha_2 \le 1$, $\lambda_1 \ge 0$, and $\lambda_2 \ge 0$ are tuning parameters. The coordinate descent algorithm (Friedman et al., 2007) or local quadratic approximations (Fan and Li, 2001) may be utilized in the M-steps of the EM algorithm to obtain the regularized estimator. Under the hood, `cox_cure_net.fit()` utilizes the coordinate-majorization-descent (CMD) algorithm proposed by Yang and Zou (2013) in the M-steps due to its descent property.

To demonstrate the usage of `cox_cure_net.fit()`, we may simulate a dataset of sample size 200 as follows. 100 covariates are simulated from multivariate normal distribution with means zero and variances one. The correlation between x_k and x_l, $k \ne l$, was set to be $\rho^{|k-l|}$, where $\rho = 0.5$. For each model part, only five covariates actually have non-zero coefficients. The true non-zero coefficients are simulated from $\mathrm{Unif}(0.6, 1)$ independently. For susceptible subjects, the event times were generated from Weibull-Cox model with baseline hazard function $h_0(t; \boldsymbol{x}) = 0.2t \exp(\boldsymbol{x}^\top \boldsymbol{\beta})$. For cured subjects, the event times were set to be infinity. The censoring times were generated independently with the event times from exponential distribution with rate 0.01 and truncated at 10. The generation of event times and censoring times takes advantage of function `intsurv::simData4cure()`.

```r
library(intsurv)
set.seed(123)
p <- 100; n <- 200; rho <- 0.5
beta0 <- gamma0 <- rep(0, p)
beta0[c(1, 2, 4, 6, 8)] <- runif(5, 0.6, 1)
gamma0[c(1, 3, 5, 7, 9)] <- runif(5, 0.6, 1)
ij_mat <- expand.grid(i = seq_len(p), j = seq_len(p))
Sigma <- matrix(mapply(function(i, j) {
    rho^abs(i - j)
}, ij_mat$i, ij_mat$j), nrow = p)
x_mat <- MASS::mvrnorm(n, mu = rep(0, p), Sigma)
colnames(x_mat) <- paste0("x", seq_len(p))
dat <- simData4cure(
    n, survMat = x_mat, survCoef = beta0,
    cureCoef = gamma0, b0 = 1, lambda_censor = 0.01,
    max_censor = 10, p1 = 1, p2 = 1, p3 = 1
)
```

Similar to function `glmnet::glmnet()` for regularized generalized linear models, `cox_cure_net.fit()` fits the regularized Cox cure rate model over a specified grid of tuning parameter λ_1 and λ_2 with fixed α_1 and α_2. Instead, the desired length of each λ sequence can be specified and an equally-spaced (in logarithm scale) sequence will be generated from the smallest "large enough" $\lambda_{\max}$ that results in all zero coefficient estimates to a specified "small enough" $\lambda_{\min}$. By default, $\lambda_{\min} = 0.1\lambda_{\max}$ is set for both model parts in `cox_cure_net.fit()`. Here we set $\alpha_1 = \alpha_2 = 0.5$ and specify a 10 by 10 grid for λ_1 and λ_2.

```
system.time({
    fit1 <- cox_cure_net.fit(
        surv_x = x_mat, cure_x = x_mat,
        time = dat$obs_time, event = dat$obs_event,
        surv_nlambda = 10, cure_nlambda = 10,
        surv_alpha = 0.5, cure_alpha = 0.5
    )
})
```

```
##    user  system elapsed
##   5.437   0.006   5.455
```

The tuning parameters may be selected based on BIC and a `coef()` method for `cox_cure_net` objects can be used to return the coefficient estimates from the selected model. We may quickly check the true positive rate and false positive rate in terms of variable selection as follows:

```
eval_vs <- function(x, beta0, gamma0) {
    foo <- function(b, b0) {
        c("% True Positive" = mean(b[b0 != 0] != 0),
          "% False Positive" = mean(b[b0 == 0] != 0))
    }
    rbind(beta = foo(coef(fit1)$surv, beta0),
          gamma = foo(coef(fit1)$cure, gamma0))
}
eval_vs(fit1, beta0, gamma0)
```

```
##       % True Positive % False Positive
## beta        1.0000000       0.09473684
## gamma       0.8333333       0.07368421
```

To reduce computational burden, the generalized EM algorithm may be used by setting one-step CMD update as follows. In this example, we are able to substantially decrease the computation time and obtain the same variable selection results.

```
system.time({
    fit2 <- cox_cure_net.fit(
        surv_x = x_mat, cure_x = x_mat,
        time = dat$obs_time, event = dat$obs_event,
        surv_nlambda = 10, cure_nlambda = 10,
        surv_alpha = 0.5, cure_alpha = 0.5,
        surv_max_iter = 1, cure_max_iter = 1
    )
})
```

```
##    user  system elapsed
##   2.319   0.002   2.322
```

```
eval_vs(fit2, beta0, gamma0)
```

```
##       % True Positive % False Positive
## beta        1.0000000       0.09473684
## gamma       0.8333333       0.07368421
```

After variable selection, a regular Cox cure rate model may be fitted by `intsurv::cox_cure()`. See `https://wenjie-stat.me/intsurv/` for the full package documents.

Reference

Berkson, J. and Gage, R. P. (1952), "Survival Curve for Cancer Patients Following Treatment," *Journal of the American Statistical Association*, 47, 501–515.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American statistical Association*, 96, 1348–1360.

Fan, X., Liu, M., Fang, K., Huang, Y., and Ma, S. (2017), "Promoting Structural Effects of Covariates in the Cure Rate Model with Penalization," *Statistical Methods in Medical Research*, 26, 2078–2092.

Farewell, V. T. (1982), "The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors," *Biometrics*, 1041–1046.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332.

Kuk, A. Y. C. and Chen, C.-H. (1992), "A Mixture Model Combining Logistic Regression with Proportional Hazards Regression," *Biometrika*, 79, 531–541.

Masud, A., Tu, W., and Yu, Z. (2018), "Variable Selection for Mixture and Promotion Time Cure Rate Models," *Statistical methods in medical research*, 27, 2185–2199.

Scolas, S., El Ghouch, A., Legrand, C., and Oulhaj, A. (2016), "Variable Selection in A Flexible Parametric Mixture Cure Model with Interval-Censored Data," *Statistics in Medicine*, 35, 1210–1225.

Shi, X., Ma, S., and Huang, Y. (2019), "Promoting Sign Consistency in the Cure Model Estimation and Selection," *Statistical methods in medical research*, 1–14.

Sy, J. P. and Taylor, J. M. G. (2000), "Estimation in a Cox Proportional Hazards Cure Model," *Biometrics*, 56, 227–236.

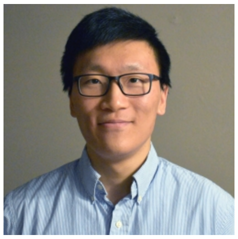
Wang, W., Chen, K., and Yan, J. (2019), *intsurv: Integrative Survival Models*, R package version 0.2.1.

Yang, Y. and Zou, H. (2013), "A Cocktail Algorithm for Solving the Elastic Net Penalized Coxs Regression in High Dimensions," *Statistics and its Interface*, 6, 167–173.

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of statistics*, 38, 894–942.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American statistical association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.



Wenjie Wang
Research Scientist
Machine Learning, Artificial
Intelligence, and Connected Care
Advanced Analytics and Data Sciences
Eli Lilly and Company
Email: `wang_wenjie@lilly.com`