Comparing Cross Validation Methods for Penalized Cox Regression

Abstract

Due to the semi-parametric nature of Cox Regression, conducting cross validation for Cox models has always been a challenge. While cross validation is a commonly used approach for selecting tuning parameters in penalized regression, little research has been done to study cross validation methods for penalized Cox regression. We propose two new cross-validation methods for Cox Regression, and compare them to traditional information criteria as well as a cross-validated partial likelihood approach originally proposed by Verweij et al. Our simulation studies show that, in general, cross-validation tends to be conservative (i.e., select smaller models than the ideal choice of tuning parameters) for penalized Cox regression models. However, our proposed approach of cross-validating the linear predictors generally offers the best balance of stability and performance. We also illustrate these approaches on data from studies of gene expression and progression-free survival in cancer patients.

1 Introduction

Cox propotional hazard regression is one of the most commonly used statistical model for analyzing data with survival outcomes. Despite its popularity, Cox model is limited by its semiparametric nature: it is challenging to evaluate the model's predictive accuracy.

In the Cox model [Cox, 1975], the hazard function for an observation i is given by

$$h_i(t) = h_0(t)\exp(X_i^T\beta) \tag{1}$$

where h_0 is the baseline hazard and $e^{X_i^T\beta}$ is the relative risk. Suppose there are n observations and m observed failure times, the estimation of the coefficients βs are obtained by maximizing the partial likelihood

$$L(\beta) = \prod_{j=1}^{m} \frac{\exp(X_j^T\beta)}{\sum_{k\in R(t_j)}\exp(X_k^T\beta)} \tag{2}$$

, where j indexes the observed failure times and $R(t_j)$ denotes the set of observations at risk at time t_j. As a semi-parametric model, Cox regression only gives estimations of the β coefficients, without estimating the baseline hazard. Hence the interpretation of the Cox model is only valid in a relative sense.

Penalized Cox Regression is an extension of the regular cox regression model. Estimates of the β coefficients are obtained by minimizing the objective function

$$Q(\beta|X,y) = -\frac{1}{n}\sum_{j=1}^{m}\log\left\{\frac{\exp(X_j^T\beta)}{\sum_{k\in R(t_j)}\exp(X_k^T\beta)}\right\} + P_\lambda(\beta). \tag{3}$$

[Simon et al., 2011]. The first part of the objective function is based on the partial likelihood. The second part is a penalty term that depends on some regularization parameters. In this paper, only the LASSO penalty $P_\lambda(\beta) = \lambda(\sum_j |\beta_j|)$ was considered. The penalized cox regression can be used for analyzing high dimensional data, where the number of covariates p $\gg$ n. A common application would be to predict the overall survival time of a particular cancer by certain gene expressions.

Selecting the tuning parameter λ is crutial to getting LASSO estimations for β. As λ increases, more coefficients would be set to 0 and fewer variables will be selected by the model. When λ is too small, the

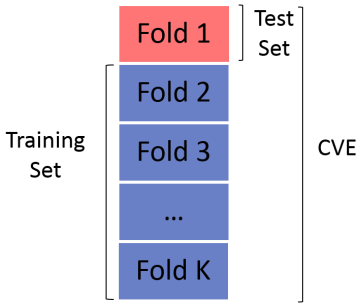

Figure 1: K-fold Cross Validation

model tends to be redundant. When λ is too large, the model would be too conservative to select any variables.

One common approach for selecting the tuning parameter λ for linear regression and logistic regression is via cross validation. The idea of K - fold cross validation is illustrated in Figure 1. The data set would be split into K folds. One fold would be treated as the test set and the other K - 1 folds as the training set. The model would first be built on the training set, then fitted to the test set to obtain cross validation error (CVE). CVE would be calculated for each candidate λ, then the one that minimizes cross validated error would be selected.

However, for penalized cox regression, it is challenging to carry out cross validation. Since the partial likelihood of the cox model does not give estimates for the baseline, it is not meaningful to compare the coefficients if they are not estimated from different models and different observations. Only if the baseline is the same, then the coefficients can be interpreted as relative risks. This is against the idea of cross validation, which splits the data and builds models on different subsets.

The most intuitive approach to conduct cross validation in Cox regression is i) fit the model to training set then ii) calculate cross-validated partial likelihood based on the observations from the test set as a measure for the model's predictive accuracy. As will be discussed in Section 3, this approach is unstable when number of events gets small. Verweij elt. [Verweij and Van Houwelingen, 1993] proposed a stablized version of cross-validated log likehood, which has been widely used and implemented in R packages such as `glmnet`. In this paper, we proposed two alternative ways to carry out cross validation. Instead of cross-validating over partial likelihood, we propose to cross validate over the linear predictors of the regression model and cross validate over the deviance residual [Therneau et al., 1990]of the Cox model.

We conducted simulation studies to compare how those methods perform in both low dimensional and high dimensional settings with LASSO penalty. We compared the two proposed methods with the cross-validated partial likelihood approaches and traditional information criteria. We showed that the linear predictor approach outperforms other cross validation approaches in various scenarios. We found out that using cross validation to select tuning parameters for LASSO penalized cox regression tends to be a conservative approach in general. Finally, we apply those methods to two high dimensional data sets with time-to-event outcomes.

2 Methods

2.1 Cross Validated Likelihood

Suppose a data set of n observations is split into K folds and the ith fold is left out, then a partial likelihood can be built over the K -1 folds and yields maximum likelihood estimates of the coefficients, denoted by $\hat{\beta}_{-i}$. An intuitive way to carry out cross validation is to use the partial likelihood on the test data as the cross validated error:

$$cvl = \sum_{i=1}^{K} l_i(\hat{\beta}_{-i}) \tag{4}$$

This is implemented in the `glmnet` package as the `ungrouped` option.

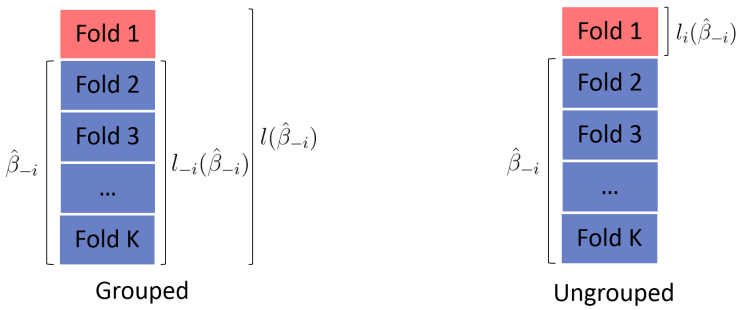

Figure 2: Grouped vs Ungrouped Cross Validated Likelihood

An alternative way to calculate cross-validated log likelihood was proposed by Verweij elt. They defined cross-validated log likelihood of leaving the ith fold out to be

$$cvl = \sum_{i=1}^{K} \{l(\hat{\beta}_{-i}) - l_{-i}(\hat{\beta}_{-i})\}. \tag{5}$$

When the ith fold is left out, $l(\hat{\beta}_{-i})$ is the log partial likelihood evaluated at $\hat{\beta}_{-i}$ with all k folds of observations. $l_{-i}(\hat{\beta}_{-i})$ is the log partial likelihood evaluated at $\hat{\beta}_{-i}$ with observations from the other K-1 folds.

We refer to the first definition as ungrouped cross-validated partial likelihood (CVL) and the second definition as grouped cross-validated partial likelihood. Figure 2 illustrates the ideas of the two methods. The grouped CVL uses observations more efficiently than the ungrouped CVL. Since only the left-out fold would be used, the ungrouped cvl is unstable in practice when there are only very few observations in some folder.

2.2 Cross Validated Linear Predictors

Besides cross validating over the partial likelihood, an alternative approach is to cross validate over the linear predictors. The data would still be split into K folds. Suppose ith fold is left out and the other K-1 folds are used as the training set to get the estimates $\hat{\beta}_{-i}$. Then the cross-validated linear predictors would be calculated based on the observations in the test set:

$$\hat{\eta}_{-i} = X_i \hat{\beta}_{-i}. \tag{6}$$

After repeating this for all K folds, a whole set of linear predictors $\hat{\eta}_{-} = (\hat{\eta}_{-1}, \hat{\eta}_{-2}, ...\hat{\eta}_{-K})$ can be obtained. A partial likelihood can be built over this set of linear predictors:

$$L(\hat{\eta}_{-}) = \prod_{j=1}^{m} \frac{exp(\hat{\eta}_{-j})}{\sum_{k \in R(t_j)} exp(\hat{\eta}_{-k})} \tag{7}$$

.

Cross validated error is defined as the log of $L(\hat{\eta}_{-})$. This idea of cross validating over linear predictors is implemented in the package ncvreg. For all the approaches introduced so far, they are all equivalent to each other in linear regressions. But since they are building over different baselines, they are different in cox regression.

2.3 Cross Validated Deviance Residuals

Since the biggest challenge of conducting cross validation for cox regression is due to the missingness of baseline hazard, we propose a second approach which involves estimating the actual baseline hazard $\hat{\Lambda}_0$. The sum of squares of Martingale residuals is a natural candicate for evaluating the model performance

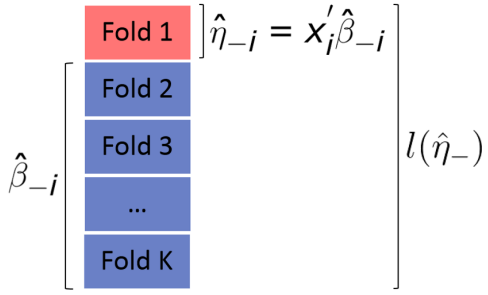

Figure 3: CV Linear Predictors

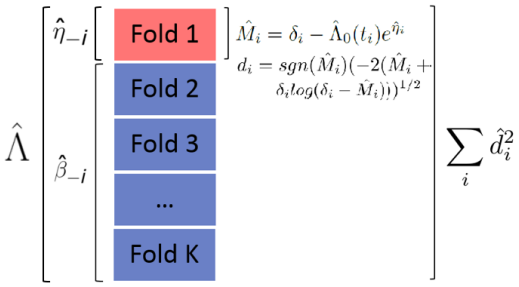

Figure 4: CV Deviance Residuals

using baseline hazard. Martingale residuals and its normalized form, deviance residuals, were first proposed by [Therneau et al., 1990] for Cox regression model diagnostics.

To estimate the baseline hazard, a set of linear predictors $\hat{\eta}_- = (\hat{\eta}_{-1}, \hat{\eta}_{-2}, ...\hat{\eta}_{-K})$ would be obtained in the same way as the cross validated linear predictor approach. $\hat{\eta}_{-i}$ is computed based on $\hat{\beta}_{-i}$s estimated from the training set and observations from the ith test set. With $\hat{\eta}_-$, a baseline hazard $\hat{\Lambda}_0$ can be estimated via Kalbfleisch and Prentice's method [Kalbfleisch and Prentice, 2011]. Then for each observations, the Martingale Residual can be calculated:

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(t_i)e^{\hat{\eta}_i}, \tag{8}$$

where δ_j is the status of the jth observation and t_j is the time component of the jth observation. The deviance residual can be calcualted from the Martingale Residuals:

$$d_i = sgn(\hat{M}_i)(-2(\hat{M}_i + \delta_i log(\delta_i - \hat{M}_i)))^{1/2} \tag{9}$$

The sum of squares of the deviance residuals, $\sum_i \hat{d}_i^2$, would be used as the cross validated error.

As an alternative, the baseline hazard can also be estimated only based on the training set, leaving the test set as out-of-sample observations for calculating the residuals. There are two major issues with this approach. First, it would be common to run into situations where events in the test set occurred after the last event occured in the baseline, where the survival function dropped to 0. Unless those observations are treated as censored, this approach would not work. Second, there would be an extra numeric challenge for deviance residuals. Since the baseline hazard function $\hat{\Lambda}_0$ is a step function from 0, the cumulative hazard at the first time point from 0 is always 0. Since deviance residual contains a term $log(\delta_i - \hat{M}_i)$, where $\delta_i - \hat{M}_i = \hat{\Lambda}_0$, this term would be negative infinity for the first time point and lead to a numeric issue for whichever fold that contains the first time point. Hence extra smoothing for the step function would be needed for this approach to work.

3 Simulation Studies

Simulation studies were conducted to compare how those methods behave relative to each other. We generate data with pre-specified baseline hazard, covariate matrix X, coefficient β and censoring mechanism. Both low dimensional and high dimentional scenarios were examined. All cross-validation methods mentioned in Section 2 were applied to the data to select the tuning parameter λ and produce $\hat{\beta}$ estimates. Cross-validation were also compared to model selection criteria AIC and BIC.

Survival times were generated from exponential distribution $h(t) = h_0 exp(X\beta)$, conditioned on covariates. The entries in the covariate matrix $X_{n\times p}$ were independently generated from Normal (0, 1). True coefficients β were assumed to have sparsity: $\beta_{p\times 1} = (\beta_1, \beta_2, ..., \beta_{10}, 0, ..., 0)^T$. Censoring status were generated based on binomial distribution.

Suppose $\hat{\beta}$ is the coefficients estimated by a fitted cox model. We measure the distance between the fitted model and the true model by mean squared error $MSE = E(\hat{\beta} - \beta)^2$. For each generated data set, the λ that has the minimal MSE is chosen as the optimal λ. Then the λs that selected by the cross validation or information criteria would be compared to this optimal λ. If the λ chosen by cross validation is smaller

than the optimal λ, then the cross validation method would be considered liberal. If the λ chosen by cross validation is larger than the optimal λ, then the cross validation method would be considered conservative.

3.1 Simulations Comparing Cross-validation Approaches

The simulation experiments were first conducted under lower dimension settings. Number of observations n is set to be 100. Dimension of the data p is set to be 100. 200 replications were used. Censoring percentage, number of folds used in cross-validation, number of non-zero βs and the magnitude of the βs were varied in several different simulation scenarios. Patterns and results shown in various simulation scenarios are consistent with one another. Figure 5 and Figure 6 illustrates one simulation scenario, where about 10% of which were set to be censored and 10 of the βs were set to be non-zero and 10 folds were used for cross validation. The x - axis of the plots are the magnitude of the non-zero β coefficients.

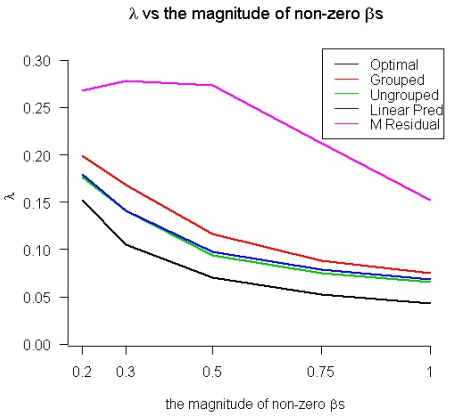

Figure 5

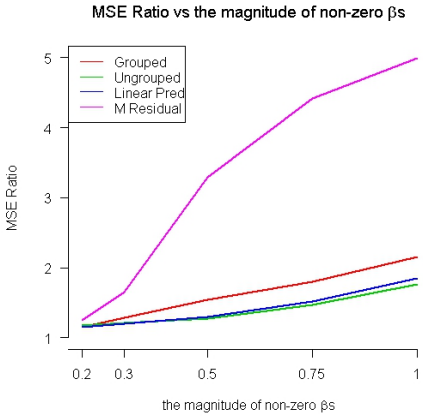

Figure 6

λs chosen by different cross-validation methods are compared in Figure 5. All values are from taking the mean of 200 replications. The black line represents the optimal λ. All four cross validation methods consistently select λs larger than the optimal one. Since larger λ leads larger penalties and fewer variables are selected, all four cross validation methods seem to be conservative in terms of selecting variables. Ungrouped cross-validated log likelihood and the cross-validated linear predictor seem to be more liberal than the other two methods. The cross-validated Deviance residual is the most conservative one and its performance is quite off from the other three methods.

Mean squared errors are compared in Figure 6. The MSEs of the cross-validation selected models are compared with the MSE of the model given by the optimal λ. The y-axis represents the ratio of the two. The ungrouped method and the linear predictor approach are outperforming the other methods and are closest to the minimal MSE.

Simulation studies were also conducted in high dimension scenarios, where the number of observations and dimension of the covariate matrix is more similar to real genetics research. The results are shown in Table 1. Conclusions are close to what is shown in the low dimensional setting. The M- residual approach performs most conservatively. Linear predictor approach and ungrouped approach are relatively more liberal than the other approaches but still slightly more conservative than the optimal λ.

3.2 Stability

While the ungrouped cross-validated likelihood is an intuitive way to carry out cross-validation, it is quite unstable when it runs into cases where the number of observed events in some fold is really small. As is illustrated in Figure 5, cases of undefined cross-validation error would increase when censoring increases or when number of folds increases.

CV Methods	λ (SD)	MSE Ratio
Optimal	0.0606 (0.0019)	/
Grouped	0.0895 (0.0043)	1.598
Ungrouped	0.0754 (0.0060)	1.225
Linear Pred	0.0759 (0.0058)	1.234
M Residual	0.1283 (0.0047)	2.959

Table 1: Simulation Results in High Dimensional Setting

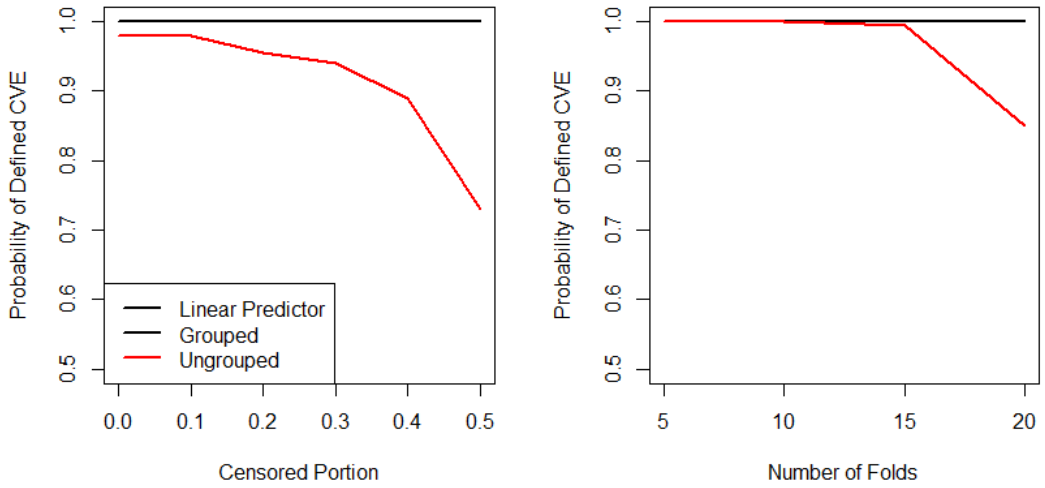

Figure 7: Comparing Stability

3.3 Comparisons with Information Criteria

We also compared cross-validation with information criteria AIC and BIC. Simulation studies were conducted in lower dimension and higher dimension respectively. As is illustrated in Figure 7, in lower dimension, AIC tends to perform more liberal and BIC tends to perform more conservative than the optimal λ. While the cross-validation methods are also conservative, most of them are less conservative than BIC. In terms of MSE, cross-validation methods yield smaller MSE than the information criteria. In higher dimension, both BIC and AIC tend to be more liberal than the optimal λ and yield larger MSE.

3.4 Issues with Deviance Residual

4 Application to Real Data

We also compared those cross validation methods when they are applied to real data sets. Unlike simulated set, we do not know the optimal λ in the real data sets, but we can still see how those methods perform relative to each other. The first data set is a study on ovariance cancer from The Cancer Genome Atlas. The second data set is a study on lung cancer (Shedden et al 2004). The outcome of both data sets are time-to-event data.

There are 460 patients and 236 events for the ovarian cancer data set. About 40% of the observations are censored. The covariate matrix records whether or not there is mutation at the location and has dimension 12376. When we fit the penalized cox model, we first adjusted three clinical variables which is known to have large impact on survival. We first fitted these three covariates into regular cox regression and keep the linear predictors. The selection for the penalty term does not affect the estimation of those clinical variables. The

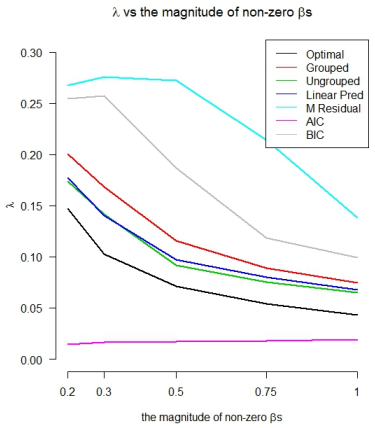

Figure 8: Comparison to Information Criteria in Lower Dimension

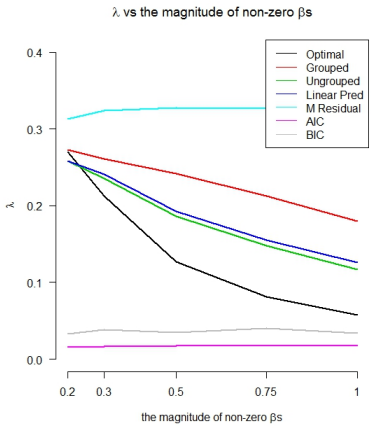

Figure 9: Comparison to Information Criteria in Higher Dimension

results are illustrated in Table 2. The linear predictor approach has the best predictive accuracy according to AUC.

CV Methods	λ	$log(\lambda)$	AUC
Grouped	0.117	-2.14	0.595
Ungrouped	0.112	-2.20	0.597
Linear Pred	0.085	-2.46	0.607
M Residual	0.171	-1.76	0.529

Table 2: Tuning Parameter Selected for Ovarian Cancer Data

In Figure 9, the Cross Validated Error is rescaled and plotted for all four methods. A λ will be selected when CVE curve reaches its lower point. The blue line, which represents the linear predictor approach, it has more curvature near its lowest point. It is easier to pick out the minimum point for this blue curve. Hence this approach is better at picking out signals than the other three approaches. If we look at the curve for the cross-validated Deviance Residual, there's almost no signal there.

In the lung cancer data set that we applied our methods to, there are 442 patients and 236 events. About 50% of the patients are censored. Again we adjusted three clinical variable before we fit the penalized regression. The results of the analysis are listed in Table 3.

CV Methods	λ	$log(\lambda)$	AUC
Grouped	0.119	-2.12	0.608
Ungrouped	0.119	-2.12	0.608
Linear Pred	0.090	-2.41	0.633
M Residual	0.203	-1.59	0.564

Table 3: Tuning Parameter Selected for Lung Cancer Data

If we look at the CVE curve for all four of those methods, we still see that blue curve has more curvature near its lowest point. But in general, there are more signal in this lung cancer data set compared to the ovarian cancer data set.

5 Discussion

-

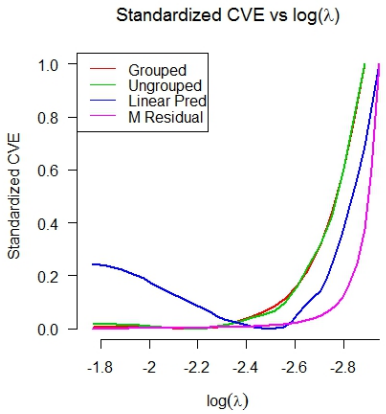

Figure 10: CVE for Ovarian Cancer Data

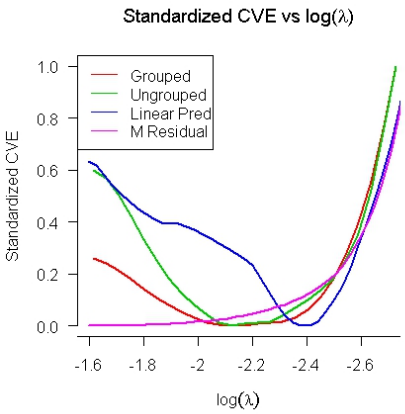

Figure 11: CVE for Lung Cancer Data

•

References

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

P. J. Verweij and H. C. Van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12 (24):2305–2314, 1993.