

Comparing Cross Validation Methods for Penalized Cox Regression

March 24, 2017

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 proportional hazard regression is one of the most commonly used statistical model for analyzing data with survival outcomes. Despite its popularity, the cox model has its own limitations due to the semiparametric nature of the model. Since the model only estimates relative risks, it is challenging to evaluate the model's predictive accuracy.

In the Cox model (Cox, 1972), the hazard function for individual observation is given by

$$h_j(t) = h_0(t) \exp(X_j^T \beta), \quad (1)$$

where h_0 is the baseline hazard and $\exp(X_j^T \beta)$ is the relative risk. The estimation of the coefficients β s are obtained by maximizing the partial likelihood

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

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. Instead of dealing with the partial likelihood, estimates of the β coefficients are obtained by minimizing the objective function

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

The first part of the objective function is based on the partial likelihood. The second part is a penalty term that depends on tuning parameter λ . With LASSO penalty $P_\lambda(\beta) = \lambda(\sum_j |\beta_j|)$, the penalized cox regression extends cox regression to deal with variable selections in high dimensional settings, where the number of variable $p \gg n$, the number of observation. An example of its application would be to study how the overall survival time of a particular cancer is associated with gene expressions.

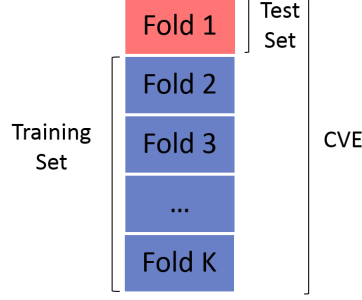


Figure 1: K-fold Cross Validation

To get LASSO estimations for the β coefficients, the choice of the tuning parameter λ is crucial. As λ increases, more β coefficients would be set to 0 and fewer variables will be selected by the model. When λ is too small, the model tends to be redundant. And λ also should not be set too large. In this case, 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). To select the most suitable lambda, the CVE would be calculated for each candidate lambda. And the one that has the smallest cross validated error would be selected.

However, for penalized cox regression, there is no natural way to carry out cross validation. Since the partial likelihood of the cox model does not give estimates for the baseline, the outcomes can only be compared within the same framework, in terms of their relative risks. This is against the idea of cross validation, which relies on splitting the data into different folds.

The most commonly used approach of conducting cross validation in Cox regression is proposed by Verweij et. (1993). They propose a cross-validated log likelihood as a measure for the model's predictive accuracy. It is currently implemented in the glmnet package as the "grouped" option. Similarly, the glmnet package also includes the "ungrouped" option where the cross-validated likelihood is calculated only based on the observations from the test set. 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 Martingale residual 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 genomic 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 $\hat{\beta}_{-i}$. Verweij et. defined cross-validated log likelihood of leaving the ith fold out to be

$$cvl = \sum_{i=1}^K \{l(\hat{\beta}_{-i}) - l_{-i}(\hat{\beta}_{-i})\}. \quad (4)$$

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

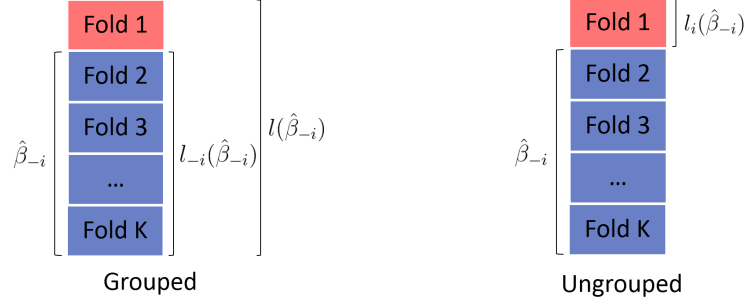


Figure 2: Grouped vs Ungrouped Cross Validated Likelihood

folds.

An alternative way of conducting cross-validation is implemented in the glmnet package. The cross-validated log likelihood is defined as

$$cvl = \sum_{i=1}^K l_i(\hat{\beta}_{-i}). \quad (5)$$

Instead of using subtraction, $l_i(\hat{\beta}_{-i})$ is the partial likelihood evaluated at $\hat{\beta}_{-i}$ with observations from the fold that is left out.

We refer to the first definition as grouped cvl and the second definition as ungrouped cvl. Figure 2 illustrates the ideas of the two methods. The grouped cvl uses observations more efficiently than the ungrouped one. Since only the left-out fold would be used, the ungrouped cvl would not very stable if number of observations is too small.

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:

$$\hat{\eta}_{-i} = X_i \hat{\beta}_{-i}. \quad (6)$$

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

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

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 ncvg. 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 Martingale 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 $\hat{\Lambda}_0$. Based on the observations in the training set, a baseline hazard $\hat{\Lambda}_0$ can be estimated via Kalbfleisch and Prentice's method. Then for each observation j in the test set, the Martingale Residual can be calculated:

$$\hat{M}_j = d_j - \hat{\Lambda}_0(t_j) e^{\hat{\eta}_j}, \quad (8)$$

$$\hat{\beta}_{-i} \left[\begin{array}{c} \text{Fold 1} \\ \text{Fold 2} \\ \text{Fold 3} \\ \dots \\ \text{Fold K} \end{array} \right] \hat{\eta}_{-i} = \mathbf{x}_i' \hat{\beta}_{-i} \quad l(\hat{\eta}_{-})$$

Figure 3: Cross Validated Linear Predictors

$$\hat{\Lambda}_{-i} \left[\begin{array}{c} \text{Fold 1} \\ \text{Fold 2} \\ \text{Fold 3} \\ \dots \\ \text{Fold K} \end{array} \right] \hat{M}_j = d_j - \hat{\Lambda}_0(t_j) e^{\hat{\eta}_{ij}} \quad \sum_j \hat{M}_j^2$$

Figure 4: Cross Validated Martingale Residuals

where d_j is the status of the j th observation and t_j is the time component of the j th observation. After repeating this procedure for all K folds, we would obtain a set of martingale residuals for all observations. They are pulled together by sum of squares as the cross validated error:

$$CVE = \sum_j \hat{M}_j^2 \quad (9)$$

3 Simulation Studies

3.1 Simulation Settings

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 dimensional 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: $\beta_{p \times 1} = (\beta_1, \beta_2, \dots, \beta_{10}, 0, \dots, 0)^T$. 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 λ s 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.

3.2 Simulation Results

3.2.1 Simulations in Low Dimensional Settings

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 β s and the magnitude of the β s 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 β s 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.

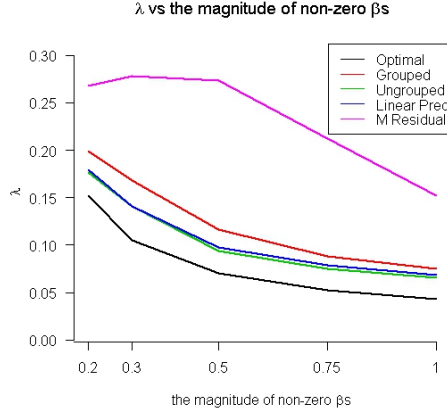


Figure 5

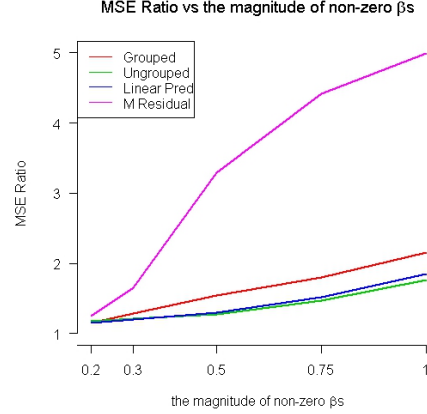


Figure 6

λ s 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 λ s 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 martingale 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.

3.2.2 Simulations in High Dimensional Setting

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 .

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

3.2.3 Comparisons with Information Criteria

We also compared cross-validation with information criteria AIC and BIC. ...

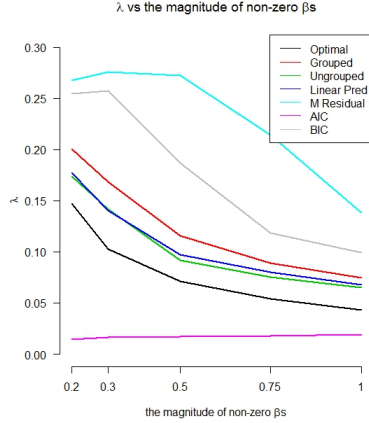


Figure 7: Comparison to Information Criteria in Lower Dimension

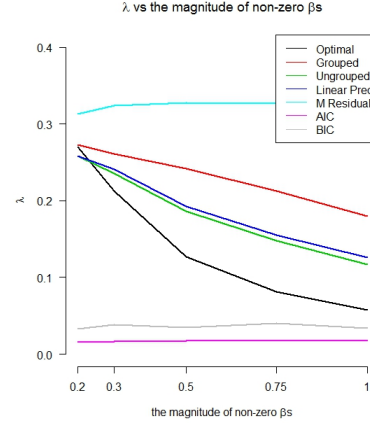


Figure 8: Comparison to Information Criteria in Higher Dimension

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

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 Martingale 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

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

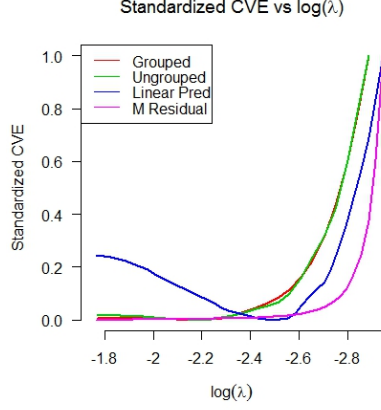


Figure 9: CVE for Ovarian Cancer Data

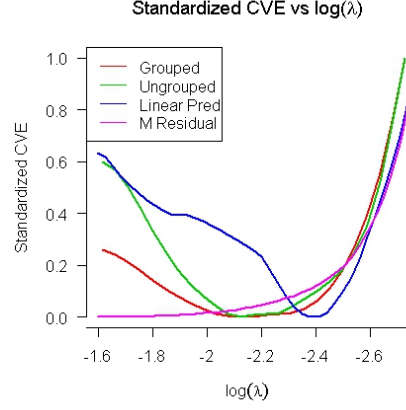


Figure 10: CVE for Lung Cancer Data

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

References

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34** 187–220.