Comparing Cross Validation Methods for Penalized Cox Regression

Abstract

While cross validation is commonly used for selecting tuning parameters in penalized regression, little research has been conducted to study cross validation methods for penalized Cox regression. Due to its semi-parametric nature, conducting cross validation for Cox models has always been a challenge. 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 overall 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 LASSO cox regression can be used to build prediction models for survival outcome based on high dimensional data, where the number of covariates p ≫ n. An example of its application would be to make predictions of cancer patients' the overall survival time with their certain gene expressions.

Selecting the tuning parameter λ is crutial to getting LASSO estimations for β. The LASSO penality enforces sparcity to the regression model, so that it carries out an automatic variable selection along the solution path. 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 model tends to be redundant. When λ is too large, the model would be too conservative to select any variables. In either case, model's predictive accuray would be affected.

One common approach for selecting the tuning parameter λ for linear regression and logistic regression is via the K-fold cross validation. 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.

Currently, there are two common cross-validation approaches for Cox regression, both having its disadvantages. 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 2, this approach is unstable when number of events gets small. The second approach was propsoed by Verweij et al. [Verweij and Van Houwelingen, 1993]. This is a stablized version of cross-validated log likehood, which has been widely used and implemented in R packages such as `glmnet` [Friedman et al., 2010]. Although this approach fixes the stability issue, it tends to yield more conservative models in practice. We hope to improve cross-validation procedure for Cox regression to achieve a good balance between stability and performance.

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 existing cross-validated partial likelihood approaches. 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 from real studies 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.

An alternative way to calculate cross-validated log likelihood was proposed by Verweij et al. 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. The upper panels in 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. In the first step, the ith fold would be left out and the other K-1 folds would be 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:

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

.

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` [Breheny and Huang, 2011]. 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

The fundamental challenge of conducting cross validation for cox regression is that the baseline hazard is not estimated from the model. Hence we proposed our 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 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:

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

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:

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

The sum of squares of the deviance residuals, $\sum_i \hat{d}_i^2$, would be used as the cross validated error. An illustration of this method is drawn in the bottom right panel of Figure 1.

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 generated 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.

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: $\beta_{p \times 1} = (\beta_1, \beta_2, ..., \beta_j, 0, ..., 0)^T$. The covariate matrix $X_{n \times p}$ was centered before it was used to generate outcomes. Survival times were generated from exponential distribution $h(t) = h_0 exp(X\beta)$, conditioned on covariates. Censoring status were generated based on binomial distribution. All simulations were implemented in R [R Core Team, 2017].

Suppose $\hat{\beta}$ is the coefficients estimated by a fitted cox model. We quantify the predictive accuracy by measuring 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.
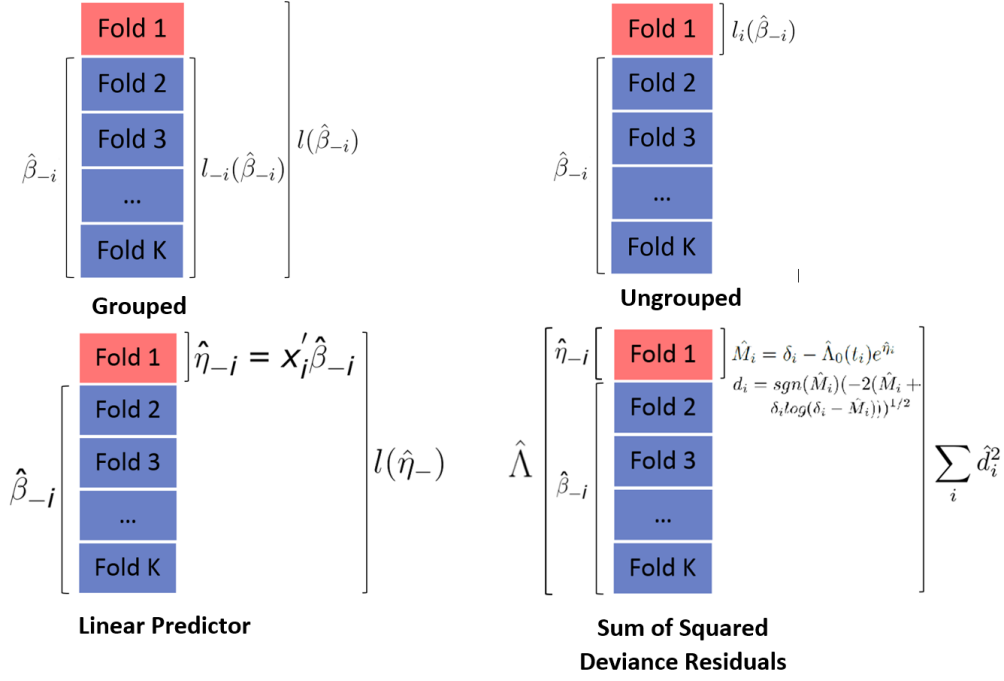
Figure 1: Illustration of K-fold Cross Validation Methods for Cox Regression

3.1 Simulations Comparing Cross-validation Approaches

The simulation experiments were first conducted when the signal in the data was varied. In each simulated data set, there are n = 100 observations and p = 1000 covariates. 10% of the observations were censored. 10-fold cross-validation was implemented for all four methods. Number of non-zero βs was 20. The signal in the data set was varied by varing the magnitude of the βs from 0.50 to 1.50. For each scenario, 200 replications were used.
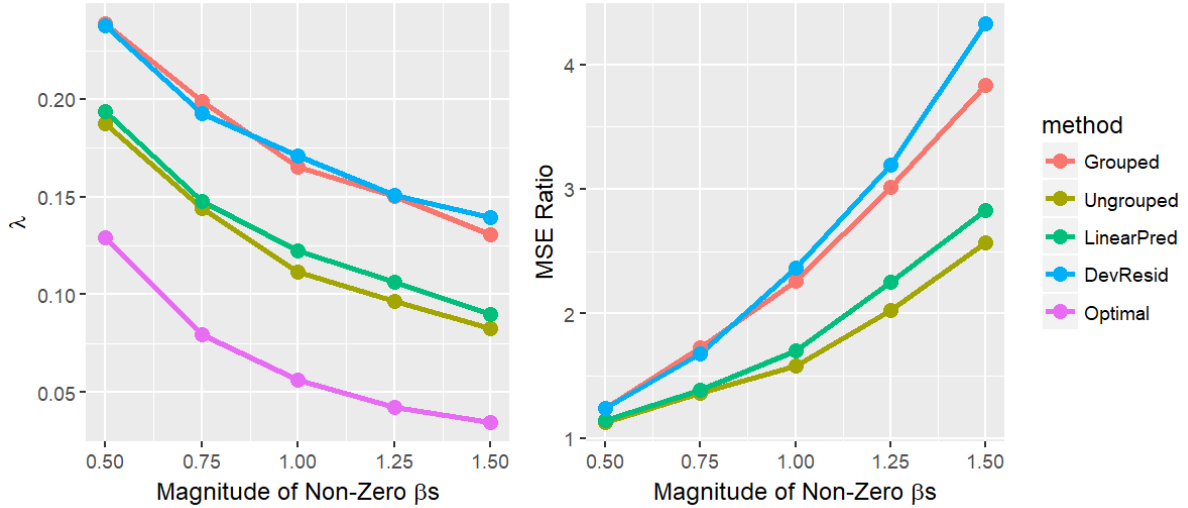

Figure 2: Simulation Studies Comparing 4 CV methods in terms of λ and MSE ratio. n = 200, p = 1000, 10% censoring.

Results of this simulation were illustrated in Figure 2. The x axis of the plots are the magnitude of the non-zero β coefficients. The left panel compares the λ selected by different methods. The grouped cross-validated partial likelihood approach, which is currently most widely used approach, consistently selects λs that are greater than the optimal one. Greater λ leads to greater penalties and fewer non-zero variables, hence the grouped approach is conservative in terms of selecting variables. While the proposed cross-validated Deviance residual has very close performance as the grouped approach, the ungrouped cross-validated log likelihood and the cross-validated linear predictor were more liberal than the other two

4

methods. In general, all four methods tends to act conservatively and selects smaller models.

The right panel compares the MSE ratios of the model selected. 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 has the smallest MSE ratio. The linear predictor approach's performance is also very close the ungrouped appraoch. Both the grouped approach and the proposed deviance residual approach yield models that have greater MSE, hence are not recommended for use in practice.

Table 1: Simulations Comparing all four Cross-Validation methods, Varying n and p

	n = 100, p = 20 #of non-zero = 10		n = 100, p = 100 #of non-zero = 10		n = 500, p = 10000 #of non-zero = 20		n = 250, p = 10000 #of non-zero = 20	
	λ (SD)	MSE Ratio	λ (SD)	MSE Ratio	λ (SD)	MSE Ratio	λ (SD)	MSE Ratio
Optimal	0.016 (0.009)	/	0.042 (0.003)	/	0.026 (0.001)	/	0.026 (0.001)	/
Grouped	0.035 (0.007)	2.597	0.073 (0.009)	2.184	0.053 (0.003)	5.627	0.123 (0.036)	3.931
Ungrouped	0.030 (0.009)	2.246	0.064 (0.01)	1.815	0.040 (0.003)	3.131	0.070 (0.016)	2.560
LinearPred	0.034 (0.007)	2.559	0.066 (0.008)	1.880	0.041 (0.003)	3.350	0.074 (0.016)	2.713
DevResid	0.100 (0.009)	14.005	0.108 (0.008)	4.087	0.076 (0.003)	9.133	0.111 (0.028)	3.767

Simulation studies were also conducted in scenarios with various sample sizes and dimensions. The results are shown in Table 1. While the ungrouped approach yields the smallest MSE and best predictive accuracy, the Linear Predictor and the Ungrouped approaches perform closely to each other across all settings. Grouped approach performs well in low-dimensions or with sufficient samples, while deviance residual approach performs better in high-dimensional settings.

3.2 Stability

In this section, simulations studies were conducted to address the stability issue of the ungrouped cross-validated likelihood. The ungrouped approach is an intuitive way to carry out cross-validation. It also has the best predictive accuarcy based on the simulation results in previous sections. However, one big disadvantage of this approach is that it is unstable when number of observed events in some fold is really small. As is illustrated in Figure 3, when censoring increases or when number of folds increases, the model would fail to converge due to insufficient number of events in the test set. Hence, the ungrouped cross-validated likelihood should not be used when there is no sufficient number of events.
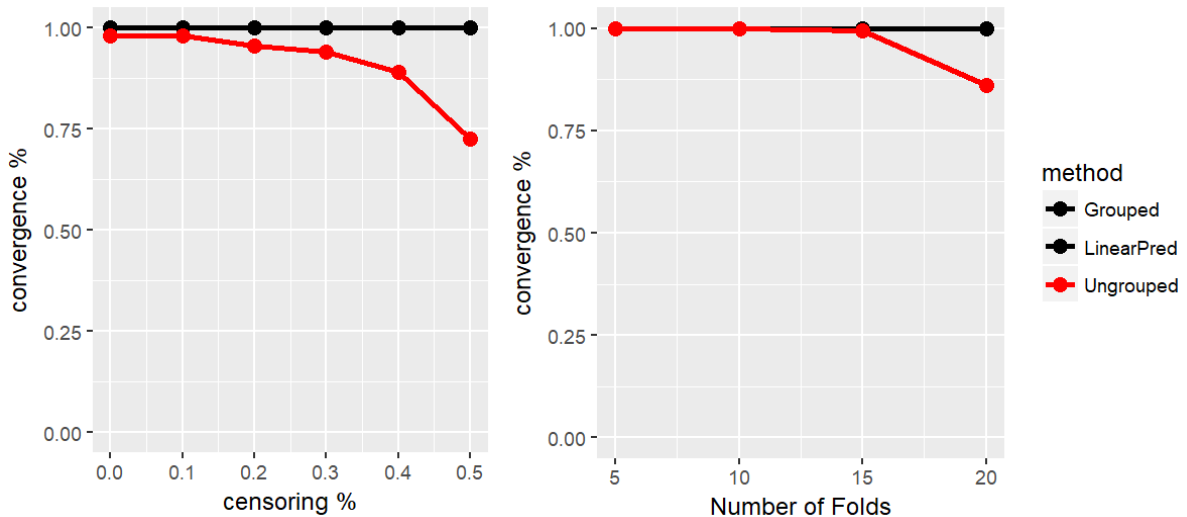

Figure 3: Comparing Stability

3.3 Results of Leave-One-Out Cross Validation

In this section, Leave-One-Out Cross Validation (LOOCV) were studied, in comparison with 10-fold CV. When first introduced, the grouped cross-validated partial likelihood was proposed so that LOOCV for Cox regression became feasible. It is

impossible to carry out LOO cross-validation using the ungrouped approach, because the partial likelihood cannot be built on the one left-out observation. While the grouped cross-validated partial likelihood solves the stability issue by aggregating more observations together, it is not an efficient way to evaluate the model fit. Simulation studies in Section 3.1 have illustrated that it tends to be conservative in variable selections and selects smaller models in various scenarios. This fact can be even more illustrated via the following simulation studies, where both 10-fold and LOOCV were conducted via grouped, linear predictor and deviance residual approaches and 10 fold CV was conducted for the ungrouped approach.

As is illustrated in Figure 4, LOOCV selected models that have smaller MSE than 10-fold CV for all three CV approaches. As is shown in left and middle panels, LOOCV based on deviance residuals and grouped cross-validated partial likelihood still perform more conservatively than the ungrouped approach. Even if the grouped approach was proposed to allow LOOCV, the LOOCV based on the grouped approach does not surpass the 10-fold ungrouped approach. On the other hand, the Linear Predictors also allows LOOCV, and the LOOCV based on Linear Predictors performs better than the ungrouped approach.
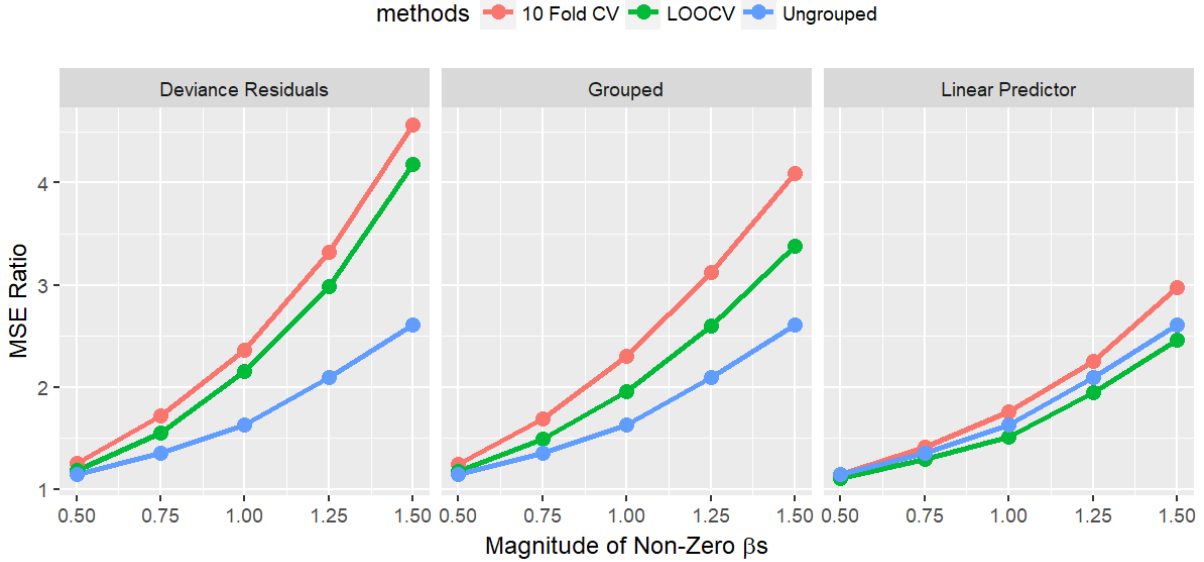

Figure 4: LOOCV

3.4 Simulations on Deviance Residuals and Baseline Hazard Estimations

While the Deviance Residual approach has a reasonable performance in scenarios where dimension is high, it tends to act more conservatively than all the other three approaches in general. As is noted by Therneau et al, the sums of squared martingale residuals and deviance residuals do not necessary reflect how good the model fit is [Therneau and Grambsch, 2000]. The difficulty of estimating the baseline hazard seems to play an important role. A series of simulations were thus designed to study how baseline hazard estimation affects the deviance residuals' performance as a measurement for model fit.

In the following simulation studies, the defined optimal λ and the λ selected by the Linear Predictor approach were used as benchmarks. Various sample sizes and dimensions were varied. The survival outcomes were generated both from exponential distribution and Weibull distribution. Six different approaches were compared in terms of baseline hazard estimation and their impact on sums of squared deviance residuals. Two non-parametric approaches, Kalbfleisch and Prentices' estimator and Breslow's estimator, were used first for estimating the cumulative hazard baseline [Kalbfleisch and Prentice, 2011] [Breslow, 1972]. As is presented in column 3 and column 5 in Table 2, the sum of squared deviance residuals computed based on both approaches are all very conservative compared to the optimal λ. Their performance is extremely bad when n is large and p is small.

Since both of the non-parameteric estimations seemed to perform well at early time points but poorly later, a weighted approach was then applied to emphasize earlier time points. Weights are assigned proportional to number of subjects at risk, making earlier time points weighted more than later time points. Results of these two approaches are presented in column 4 and 6 in Table 2. The attempt to improve the non-parameteric estimations via weighting does not seem to improve the selection too much.

In the next step, the non-parametric baseline hazard estimation was replaced by the exact baseline hazard that was used to generate the outcome. Results were listed in the seventh column in Table 2. When the baseline was known, the performance

of sum of squared deviance residuals actually performed really close to the optimal λ. This simulation suggests that when the baseline hazard was accurately estimated, then the sum of squared deviance residuals gave really good measures for the model fit.

In the last step, we replaced the non-parametric baseline hazard estimation with a parametric approach, where we assume the baseline hazard to follow an exponential distribution. We derived the estimation for baseline hazard by maximum likelihood approach conditioned on covariates: $\hat{\lambda}_0 = \frac{\sum d_i}{\sum t_i e^{\eta_i}}$. The results are listed in the eighth column in Table 2. When this estimation was used for data that were generated from exponential distribution, its performance was as well as when the exact baseline was specified. When it was fitted to data that were generated from the Weibull distribution, that is when the model is misspecified, then its performance was even worse than the non-parametric approaches.

Hence, getting accurate estimation of the baseline hazard is crucial to whether or not the sum of squared deviance residuals can be a good measurement for model fit.

Table 2: Average of Selected λ with Different Baseline Estimation for Deviance Residuals

	Optimal	Linear Pred	KP	KP weighted	Breslow	Breslow weighted	Exact	Exp
True Baseline: $Exponential(1)$								
n = 100, p = 10, non-zero β: 2	0.0461	0.0665	0.1986	0.1970	0.1908	0.1975	0.0540	0.0539
n = 100, p = 100, non-zero β: 20	0.0555	0.1841	0.2359	0.2201	0.2146	0.2244	0.0805	0.0985
n = 100, p = 1000, non-zero β: 20	0.0268	0.0505	0.0948	0.0798	0.0867	0.0809	0.0288	0.0270
True Baseline: $Weibull$								
n = 100, p = 10, non-zero β: 2	0.0476	0.0710	0.2004	0.1969	0.1907	0.1907	N/A	0.2330

4 Application to Real Data

In this section, we compared the performances of those cross validation methods when they are applied to data from real studies. The first data set is collected from a study on ovariance cancer from The Cancer Genome Atlas. The second dataset is a study on lung cancer [Shedden et al., 2008]. The outcomes of both data sets are patients' overall survival time.

In the ovarian cancer data set, there are 460 patients and 236 events . About 40% of the observations are censored. The covariate matrix records whether or not there is mutation at certain gene location and has dimension of 12376. Three clinical variables, age, gender and residuals, which were known to have large impact on survival were adjusted in the analysis. That is to say, the clinical covariates were first fitted into regular cox regression and only linear predictors were kept so that the selection for the penalty term would not affect the estimation of those clinical variables. Cross-validation was used for the selection of the penalty term, which affects the selection of gene mutations. The results are illustrated in Table 3.

In the lung cancer data set , researchers collected gene expression data for 22283 genes. 236 death events were observed for 442 patients. About 50% of the patients are censored. Again clinical variables, age, gender, status of getting adjuvant chemotherapy were adjusted before the LASSO cox regression was fitted. Cross-validation was again used for the selection of the penalty term, which selects the gene expressions that are associated with patients' overall survival. The results of the analysis are also listed in Table 3.

Table 3: Application to Real Data Sets

	Ovarian		Shedden	
	λ	$\log(\lambda)$	λ	$\log(\lambda)$
Grouped	0.1226	-2.0985	0.1215	-2.1079
Ungrouped	0.0803	-2.5222	0.0983	-2.3197
Linear Pred	0.0828	-2.4919	0.1013	-2.2894
Dev Resid	0.0992	-2.3104	0.1110	-2.1987

How the CV methods perform on real data reflects what is shown in simulation studies. The ungrouped CVL is the most liberal approach. Deviance Residual and Grouped CVL approaches were both really conversative, while the ungrouped approach was the second liberal to ungrouped CVL. In Figure 5, 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, 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. For the grouped cross-validated partial likelihood, there's almost no signal there.
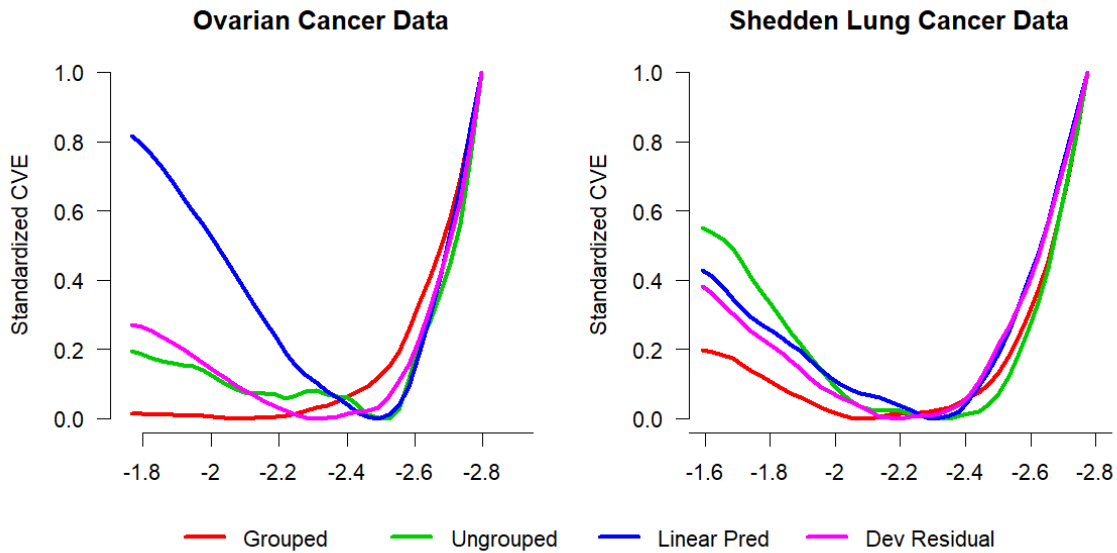
Figure 5: Comparing Cross-Validated Error for Different Methods on Real Data Sets

5   Discussion

Four methods of conducting cross-validation for Cox Regression were compared through simulation studies in terms of their predictive accuracy. All cross-validation methods tend to select the more conservative model: with greater value of λ, and fewer variables selected into the model. Despite of the stability issue, the ungrouped approach is the most liberal one across all scenarios and yield the least MSE. The proposed linear predictor approach performs closely to the ungrouped approach. Under LOOCV, the linear predictor approach out-performs the ungrouped approach. Both the grouped approach and the proposed Deviance Residual approach perform conservatively, hence not recommended for use in practice. Results from real data application reflects results from the simulation studies.

References

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253, 2011.

N. Breslow. Discussion following "regression models and life-tables". *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. URL http://www.jstor.org/stable/2985181.

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

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

K. Shedden, J. M. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, et al. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822, 2008.

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 and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer, 2000.

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.