Journal of the Korean Data & Information Science Society
2011, **22**(4), 795–801

한국데이터정보과학회지

Variable selection in the kernel Cox regression[†]

Jooyong Shim[1]

[1]Department of Data Science, Inje University

Received 29 May 2011, revised 21 June 2011, accepted 26 June 2011

Abstract

In machine learning and statistics it is often the case that some variables are not important, while some variables are more important than others. We propose a novel algorithm for selecting such relevant variables in the kernel Cox regression. We employ the weighted version of ANOVA decomposition kernels to choose optimal subset of relevant variables in the kernel Cox regression. Experimental results are then presented which indicate the performance of the proposed method.

Keywords: ANOVA decomposition kernel, generalized cross validation function, kernel Cox regression model, variable selection.

1. Introduction

Let t_i be the response variables corresponding to input vector, $\boldsymbol{x}_i$ or transformation on it, where $i = 1, 2, \cdots, n$. In fact we can not observe t_i's but the observed variable, $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$, where $I(\cdot)$ denotes the indicator function and c_i is the censoring variable corresponding to $\boldsymbol{x}_i$ for $i = 1, 2, \cdots, n$. c_i's are assumed to be independently distributed with unknown survival distribution functions.
The Cox regression model (proportional hazard model; Cox, 1972, 1975) includes the hazard function of the ith subject with input vector $\boldsymbol{x}_i$ of the form such that

$$h(t_i|\boldsymbol{x}_i) = h_0(t_i) \exp(\boldsymbol{\beta}'\boldsymbol{x}_i), \tag{1.1}$$

where $h_0(t_i)$ is a unspecified baseline hazard function and $\boldsymbol{\beta}$ is a $d \times 1$ regression parameter vector. We assume the following general Cox regression model, where the hazard function for the ith subject is modeled as

$$h(t|\boldsymbol{x}_i) = h_0(t) \exp(f(\boldsymbol{x}_i)) \tag{1.2}$$

where $f(\boldsymbol{x}_i)$ is an arbitrary nonlinear function of the input vector $\boldsymbol{x}_i$. Li and Luan (2003), Evers and Messow (2008) applied the kernel methods to estimate the function $f(\boldsymbol{x}_i)$.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0028135).
[1] Adjunct professor, Department of Data Science, Inje University, Obang-Dong, Kimhae 621-749, Korea. E-mail: ds1631@hanmail.net

Variable selection is the method of selecting a subset of relevant variables (features) on the response. It can be shown that optimal variable selection requires an exhaustive search of all possible subsets of variables. If large numbers of variables are available, this is impractical. For the Cox regression, all the input variables may not affect the survival patterns so that some corresponding regression parameters may be zeros in true linear hazard function. Many variable selection techniques for linear regression models have been extended to the context of survival models, including the best-subset selection, stepwise selection, and Bootstrap procedures (Sauerbrei and Schumacher, 1992). LASSO (least absolute shrinkage and selection operator, Tibshirani; 1996) which belongs to the variable selection methods based on the penalized likelihood approach (Fan and Li, 2001) has been proposed for the Cox regression (Tibshirani, 1997). By shrinking some regression parameters to zero, this method provides the selection of important variables and the estimation of regression parameters simultaneously. Recently Zhang and Lu (2007) applied the adaptive LASSO for the Cox regression.

In this paper we propose a variable selection method in the kernel Cox regression, which uses ANOVA decomposition kernel (Schoelkpf *et al.*, 1998), which can be used even for nonlinear hazard function. From the quadratic programming problem we obtain weights whose magnitudes imply the importance of variables on the kernel Cox regression. The rest of paper is organized as follows. In Section 2 we present the kernel Cox model and model selection methods. In Section 3 we propose the variable selection method using ANOVA decomposition kernel. In Section 4 we perform the numerical studies with simulated datasets. In Section 5 we give the concluding remarks.

2. The kernel Cox regression model

Under the assumption of no ties, we can see that for each uncensored time y_i,

$$P(\text{a failure in } [y_i, y_i + \triangle y) | R_i) \approx \sum_{j \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{x}_j) h_0(y_i) \triangle y, \tag{2.1}$$

$$P(\text{a failure of } i \text{ at } y_i \mid \text{a failure in } R_i \text{ at } y_i) = \frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{x}_j)}, \tag{2.2}$$

where R_i is the risk set at time y_i. Cox (1972, 1975) proposed the proportional hazard regression model by treating the conditional likelihood (2.2) as an ordinary likelihood. The log partial likelihood of the Cox regression model is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i [\boldsymbol{\beta}' \boldsymbol{x}_i - \log\{ \sum_{j=i}^{n} \exp(\boldsymbol{\beta}' \boldsymbol{x}_j) \}]. \tag{2.3}$$

When ties are present, the technique in Breslow (1974) can be used. The maximum likelihood estimate of $\boldsymbol{\beta}$ is obtained by solving $\partial l(\beta)/\partial \beta = \mathbf{0}$, which usually needs the iterative methods. Under the Cox regression model, the survival function is obtained as follows,

$$S(t : \boldsymbol{x}) = S_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{x}), \text{ where } S_0(t) = \exp(-H_0(t)), \tag{2.4}$$

where $H_0(t)$ is the baseline cumulative hazard function. Breslow (1974) proposed the estimates of the survival function by assuming the piecewise constant baseline hazard functions.

Tsiatis (1978) obtained the estimate of the survival function by assuming that the cumulative baseline hazard function is a step function.

With a nonlinear feature mapping function $\phi(\boldsymbol{x}_i)$the hazard function (1.2) can be written as

$$h(t_i|\boldsymbol{x}_i) = h_0(t_i)\exp(\boldsymbol{w}'\phi(\boldsymbol{x}_i)), \tag{2.5}$$

where $\boldsymbol{w}$ is a corresponding weight vector. Known that $\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which are obtained from the application of Mercer's (1909) conditions. Under the assumption of no ties, the log partial likelihood of the kernel Cox model (Evers and Messow, 2008) is given by

$$l(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \delta_i [K_i\boldsymbol{\alpha} - \log\{ \sum_{j=i}^{n} \exp(K_j\boldsymbol{\alpha})\}] \tag{2.6}$$

where K_i is the ith row of $K = K(\boldsymbol{x}, \boldsymbol{x})$.We consider the minimization of the penalized log partial likelihood function,

$$L(\boldsymbol{\alpha}) = -l(\boldsymbol{\alpha}) + \frac{\lambda}{2}\boldsymbol{\alpha}'K\boldsymbol{\alpha}, \tag{2.7}$$

where $\lambda > 0$ is the regularization parameter.

The optimal values of $\boldsymbol{\alpha}$ is usually obtained from the penalized log partial likelihood function (2.7) by Newton-Raphson method, in which $\boldsymbol{\alpha}$ at $(t+1)$th iteration can be obtained from

$$\boldsymbol{\alpha^{(t+1)}} = \boldsymbol{\alpha^{(t)}} - (H + \lambda K)^{-1}\left(G + \lambda K\boldsymbol{\alpha^{(t)}}\right), \tag{2.8}$$

where G is the gradient vector of $-l(\boldsymbol{\alpha^{(t)}})$ and H is the Hessian matrix of $-l(\boldsymbol{\alpha^{(t)}})$. Using first order Taylor expansion of $L(\boldsymbol{\alpha})$, we can express the Newton-Raphson method (2.8) as the iterative reweighted least squares (IRWLS) procedure as follows:

$$\boldsymbol{\alpha^{(t+1)}} = \left(K + \frac{1}{\lambda}H\right)^{-1}\boldsymbol{z}, \tag{2.9}$$

where $\boldsymbol{z} = (H\boldsymbol{\alpha^{(t)}} - G)/\lambda$. Note that $\boldsymbol{\alpha}$ in (2.9) is the minimizer of

$$\frac{1}{2}(\boldsymbol{z} - K\boldsymbol{\alpha})'H^{-1}(\boldsymbol{z} - K\boldsymbol{\alpha}) + \frac{1}{2\lambda}\boldsymbol{\alpha}'K\boldsymbol{\alpha} \tag{2.10}$$

which will be used in the model selection and the variable selection.

The functional structures of the penalized log partial likelihood for the kernel Cox regression model is characterized by the regularization parameter λ and the kernel parameter. Li and Luan (2003) proposed the leave-one-out cross validation (CV) function based on the partial likelihood of the Cox regression and choose the optimal parameters which maximizes the leave-one-out cross validation (CV) function,

$$CV(\theta) = \prod_{i=1,\ \delta_i=1}^{n} \frac{\exp(\widehat{f}^{(-i)}(\boldsymbol{x}_i|\theta))}{\sum_{j\geq i}\exp(\widehat{f}^{(-i)}(\boldsymbol{x}_i|\theta))}, \tag{2.11}$$

where θ is the set of the regularization parameter λ and the kernel parameter, and $\widehat{f}^{(-i)}(\boldsymbol{x}_i|\theta)$ is the function estimated without ith observation uncensored.

Since for each candidate of sets of parameters, $\widehat{f}^{(-i)}(\boldsymbol{x}_i|\theta)$ for $i = 1, \cdots, n_u$ (number of the uncensored), should be evaluated, selecting parameters using CV function is computationally formidable.

With the final estimate of $\boldsymbol{z}$ given from (2.9), the optimal values of hyper parameters can be chosen by minimizing the generalized cross validation (GCV) function (Craven and Wahba, 1979) as follows:

$$GCV(\theta) = \frac{1}{n}\frac{(\boldsymbol{z} - K\boldsymbol{\alpha})' H^{-1}(\boldsymbol{z} - K\boldsymbol{\alpha})}{(1 - tr\{K(K + H/\lambda)^{-1}\}/n)^2}, \tag{2.12}$$

which has a similar formular as GCV function used in kernel regression of Cho *et al.* (2010), Hwang and Shim (2010), Shim (2005), Shim and Lee (2009).

3. Variable selection using ANOVA decomposition kernels

The ANOVA decomposition kernels are inspired by ANOVA in Statistics, which can be seen as the sum of kernels constructed by different subsets of variables (Schoelkopf *et al.*, 1998). The ANOVA decomposition kernel is known to has two main advantages (Saunders *et al.*, 1998) - the improvement of prediction performance by considering the different subsets as group together like variables and the avoidance of overfitting by considering only some input variables.

Let $\boldsymbol{x}$ be the $n \times d$ input matrix and $\boldsymbol{x}_{:k}$ is the $n \times p$ ($\leq d$) submatrix of $\boldsymbol{x}$ consisting of the kth subset of $I_p = \{(k_1, \cdots, k_p) | 1 \leq k_1 < \cdots < k_p \leq d\}$ and $d_p = \binom{d}{p}$ is the size of I_p, the ANOVA decomposition kernel is given as

$$\boldsymbol{K}_A = \sum_{k=1}^{d_p} K(\boldsymbol{x}_{:k}, \boldsymbol{x}_{:k}) \tag{3.1}$$

In this paper we modify the ANOVA decomposition kernel as the weighted version such as

$$\boldsymbol{K}_A = \sum_{k=1}^{d_p} v_k K(\boldsymbol{x}_{:k}, \boldsymbol{x}_{:k}) \tag{3.2}$$

where $\boldsymbol{x}_{:k}$ is the $n \times p$ submatrix of $\boldsymbol{x}$ consisting of the kth subset of I_p, $v_k \geq 0$, $\sum_{k=1}^{d_p} v_k = 1$ and v_k is the weight representing the magnitude of influence of kth set of input variables on the hazard function. The important set of input variables can be selected according to magnitude of weight $v_k's$.

The weights v_k's and $\boldsymbol{\alpha}$ cannot be obtained in a step but by the iterative procedure since $\boldsymbol{\alpha}$ contains v_k. The iterative procedure of variable selection in the kernel Cox regression can be carried out as follows:

(i) Initialize weights $\boldsymbol{v} = (1/d_p, \cdots, 1/d_p)'$.

(ii) Find $\boldsymbol{\alpha}$ from the IRWLS procedure (2.9) by replacing $\boldsymbol{K}$ with $\boldsymbol{K}_A$ in (3.2).

(iii) With $\boldsymbol{\alpha}$, find weights $\boldsymbol{v}$ from the quadratic programming problem,

$$\min \frac{1}{2}\boldsymbol{v}'A'H^{-1}A\boldsymbol{v} - \boldsymbol{z}'H^{-1}A\boldsymbol{v} \tag{3.3}$$

subject to $\mathbf{0} \le \boldsymbol{v} \le \mathbf{1}$ and $\mathbf{1}'\boldsymbol{v} = 1$,
where $A_k = K(\boldsymbol{x}_{:k}, \boldsymbol{x}_{:k})\boldsymbol{\alpha}$ is the $n \times 1$ vector, $A = (A_1, \cdots, A_d)$. Note that (3.3) is derived from $\frac{1}{2}(\boldsymbol{z} - K_A\boldsymbol{\alpha})'H^{-1}(\boldsymbol{z} - K_A\boldsymbol{\alpha})$.
(iv) Iterate (ii) and (iii) until $||\boldsymbol{v}^{(t)} - \boldsymbol{v}^{(t+1)}|| <$ tolerance.

4. Numerical studies

We illustrate the performance of the proposed method of variable selection in the kernel Cox regression by comparing its performance with adaptive LASSO (Zhang and Lu, 2007) and the exhaustive search using kernel Cox regression via 50 simulated datasets, respectively.

Example 4.1 In each dataset of size $n = 44$, the hazard function of ith subject is set to $h(t_i|\boldsymbol{x}_i) = 0.1\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ with $\boldsymbol{\beta}' = (1, 0, 1, 0, 0, 0)'$. The input vector $\boldsymbol{x}_i$ is generated from $N(\mathbf{0}_{6\times1}, \mathrm{I}_6 \times 6)$, the survival time t_i is set to $t_i = -\log(u_i)/h(t_i|\boldsymbol{x}_i)$ with u_i generated from $U(0, 1)$ and the censored time c_i is generated from $U(0, 6)$ and the observed time y_i is $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \le c_i)$, $i = 1, 2, \cdots, 44$. In each dataset of 50 datasets, we employ the linear kernel and consider the selection of two important variables, which leads 15 weights v_k, $k = 1, \cdots, \binom{6}{2} = 15$. The box plots of weights of sets of two variables and the estimated regression parameters by the Cox method with adaptive LASSO (Zhang and Lu, 2007) are shown as in Figure 4.1. From the figure we can see that the proposed method agree with adaptive LASSO (Zhang and Lu, 2007) in the variable selection of x_1 and x_2.

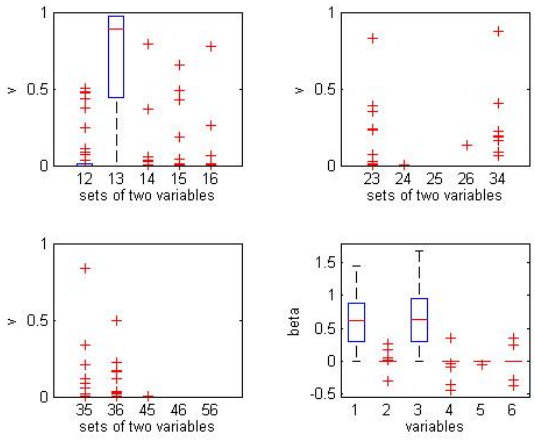

Figure 4.1 Weights of sets of two variables and the estimated regression parameters by adaptive LASSO

Example 4.2 In each dataset of size $n = 44$, the hazard function of ith subject is set to $h(t_i|\boldsymbol{x}_i) = 0.1\exp(|\sin(6\pi x_2 + 6\pi x_3)|)$. The input vector $\boldsymbol{x}_i$ is generated from $N(\mathbf{0}_{6\times 1}, \mathrm{I}_6 \times 6)$, the survival time t_i is set to $t_i = -\log(u_i)/h(t_i|\boldsymbol{x}_i)$ with u_i generated from $U(0,1)$ and the censored time c_i is generated from $U(0,6)$ and the observed time y_i is $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \le c_i)$, $i = 1, 2, \cdots, 44$. In each dataset of 50 datasets, we employ RBF kernel,

$$K(\boldsymbol{x}_1, \boldsymbol{x}_j) = \exp(-\frac{1}{\sigma^2}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2) \tag{4.1}$$

and use 15 sets of two variables $v_k, k = 1, \cdots, \binom{6}{2} = 15$, which are shown in Table 4.1. For the comparison we apply the kernel Cox model (Evers and Messow, 2008). We divide each dataset into 15 sub-datasets according to 15 sets of two variables such as $\{y_i, \delta_i, x_{i1}, x_{i2}\}_{i=1}^{44}$, $\{y_i, \delta_i, x_{i1}, x_{i3}\}_{i=1}^{44}, \cdots, \{y_i, \delta_i, x_{i5}, x_{i6}\}_{i=1}^{44}$ and obtain 15 likelihoods $l(\boldsymbol{\alpha})$'s in (2.6) for 15 sub-datasets. We compute the averages of 50 $l(\boldsymbol{\alpha})$'s for each set of two variables to choose two most important variable, which are shown in Table 4.1. From the table we can see that the proposed method agree with the exhaustive search using the kernel Cox model in the variable selection of (x_2, x_3) as the first important set of two variables and (x_3, x_5) as the second important set of two variables.

Table 4.1 Results of Example 4.2 (standard deviation in parenthesis)

variables	average of v_k	average of $l(\alpha)$
1 2	0.0761 (0.1481)	46.1136 (9.1283)
1 3	0.0748 (0.1554)	46.0041 (9.5062)
1 4	0.0421 (0.1049)	49.8397 (9.8303)
1 5	0.0682 (0.1564)	50.0529 (10.0471)
1 6	0.0343 (0.0676)	50.2795 (9.7087)
2 3	0.1403 (0.2221)	42.4665 (9.0329)
2 4	0.0942 (0.2085)	46.0914 (9.5227)
2 5	0.0712 (0.1487)	46.1671 (9.8717)
2 6	0.0561 (0.1312)	46.4810 (9.6283)
3 4	0.0684 (0.1556)	46.0569 (9.0557)
3 5	0.1358 (0.1973)	45.9369 (9.1771)
3 6	0.0701 (0.1342)	46.4461 (8.9443)
4 5	0.0231 (0.0678)	50.4073 (10.2590)
4 6	0.0376 (0.1024)	50.2113 (9.5618)
5 6	0.0078 (0.0239)	50.5981 (10.0352)

5. Concluding remarks

In this paper we dealt with variable selection in the kernel Cox regression model. We modify the penalized log partial likelihood function of the kernel Cox regression model into the objective function of penalized least squares regression consisted of working variable. This provides not only easy derivation of the generalized cross validation function which enables the model selection faster than the leave-one-out cross validation function but also easy variable selection method for high-dimensional data. From the simulated data we found that the proposed method provides the satisfying results.

References

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics,* **30**, 89-99.

Cho, D. H., Shim, J. and Seok, K. H. (2010). Doubly penalized kernel method for heteroscedastic autoregressive data. *Journal of the Korean Data & Information Science Society*, **21**, 155-162.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, **31**, 377-403.

Evers, L. and Messow, C. M. (2008). Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, **24**, 1632-1638.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association,* **96**, 348-1360.

Hwang, C. and Shim, J. (2010). Semiparametric support vector machine for accelerated failure time model. *Journal of the Korean Data & Information Science Society*, **21**, 467-477.

Li, H. and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, **8**, 865-876.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society A*, 415-446.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistical Medicine,* **11**, 2093-2099.

Saunders, C., Gammerman, A. and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. *Proceedings of the 15th International Conference on Machine Learning*, 515-521.

Schoelkopf, B., Burge, C. and Smola, A. (1998). *Advances in kernel methods: Support vector learning*, MIT Press, MA.

Shim, J. (2005). Censored kernel ridge regression. *Journal of the Korean Data & Information Science Society*, **16**, 1045-1052.

Shim, J. and Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of the Korean Data & Information Science Society*, **20**, 467-472.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Tibshirani, R.(1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Tsiatis, R. (1978). *A heuristic estimate of the asymptotic variance of survival probability in Cox' regression model,* Technical report of University of Wisconsin, number 524.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691-703.