Ultra Scientist Vol. 25(1)A, 103-116 (2013).

Variable Selection for Multivariate Survival data

[1]A. LOKESHMARAN* and [2]R. ELANGOVAN

(Acceptance Date 14th March, 2013)

Abstract

It is assumed for the Cox's proportional hazards model that the survival times of subjects are independent. This assumption might be violated in some situations, in which the collected data are correlated. The well-known Cox model is not valid in this situation because independence assumption among individuals is violated. For this purpose Cox's proportional hazard model is extent to the analysis of multivariate failure time data, which includes frailty models and marginal model. In this paper frailty and marginal hazard models are discussed using nonconcave penalized likelihood approach. Detailed illustrations are also provided.

Key words: Cox's Proportional Hazards Model, Multivariate Failure Time Data, Frailty Model, Marginal Model, Nonconcave Penalized Likelihood Approach.

1. Introduction

Variable selection is vital to survival analysis. In practice, many covariates are often available as potential risk factors. At the initial stage of modeling, data analysis usually introduces a large number of predictors. To enhance model predictability and interpretation, a parsimonious model is always desirable. Thus, selecting significant variables plays crucial roles in model building and is very challenging in the presence of a large number of predicators. Most variable selection criteria are closely related to penalized least squares and penalized likelihood. Variable selection for multivariate failure data received much

* Presented in the International Conference on Frontiers of Statistics and its applications (ICONFROST-2012) and 32nd Annual Convention of Indian Society for Probability and Statistics (ISPS), Department of Statistics, Pondicherry University (A Central University), Dec 21-23, 2012, Puducherry, India.

attention by many researchers, to name a few Akaike[1], Schwarz[26], Volinsky and Raftery[29], Fan and Li[12-15]. The basic assumption in Cox's proportion hazard model is that the survival time of subjects are independent. This assumption may be violated some time and the collected data may exhibit the existence of correlation among the survival times of the chosen subjects. One popular approach to model correlated survival times is to use a frailty model. Unlike the Cox regression model, there are some challenges in parameter estimation in the Cox frailty model even without the task of model selection[4-7].

The interpretations of the regression coefficients in the frailty model are different from those in the Cox model. Consequently, when the correlation among the observations is not of interest, the marginal proportional hazard models have received much attention in the recent literature because they are semi-parametric models and retain the virtue of the Cox model[8,9]. In this paper, the extension of the Cox regression model to the analysis of multivariate failure time data include Frailty and Marginal hazard models are discussed. Detailed illustrations are also provided.

Recent development in Cox's proportional hazard model:

A real survival analysis application via variable selection for Cox proportional model has been discussed by Androulakis *et. al.*[3]. In this paper different statistical methods are applied to analyze trauma annual data, collected by 30 general hospitals in Greece. L_1 penalized estimation in the Cox proportional hazard model is discussed by Geoman[17]. In this paper the algorithm is demonstrated in Cox proportional hazard model, predictive survival of breast cancer patients use Gene expression data. Variable selection for multivariate failure time data has been discussed by Fan *et. al*. (2005). The proposed variable selection procedure has been analyzed for the data set collected in the Framingham Heart study[11].

The paper is organized as follows, In section 2, we briefly introduce, HARD, SCAD, LASSO, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Penalized least square, Penalized likelihood and Cox's proportional hazard model. It is assumed for the Cox's proportional hazard models that the survival times of subjects are independent. This assumption might be violated some situation in which the collected data correlated. The well known Cox's model is not valid in this situation, because independent assumption among individual is violated. The extension of the Cox's regression model for the analysis of multivariate failure time data include Frailty and Marginal Models are discussed in section 3. Detailed illustrations are provided in section 4.

2.1. HARD Threshold Penality :

In the discussion of Antoniadis[4], Fan observed that the penalized least-squares estimator with the penalty function $p(|\theta|) = |\theta|I(|\theta| \leq \lambda) + \lambda/2I(|\theta| > \lambda)$ leads to the hard-thresholding rule

$$\hat{\theta} = zI(|z| > \lambda)$$

This penalty function does not over penalize the large value of $|\theta|$. Fan proposed the following hard thresholding penalty function:

$p\lambda(|\theta|)=\lambda^2-(|\theta|-\lambda)^2\ I(|\theta|<\lambda)$
With the clipped L1-penalty function
$p\lambda(|\theta|)=\lambda \min(|\theta|,\lambda)$
the solution is a mixture of soft and hard thresholding rule

$\hat{\theta}=\text{sgn}(z)(|z|-\lambda)_+ I(|z|\leq 1.5\lambda)+zI(|z|>1.5\lambda)$

2.2. Smoothly Clipped Absolute Deviation Penalty (SCAD) :

All of penalty functions introduced so far do not satisfy mathematical conditions imposed for a continuous and thresholding rule. The continuous differentiable penalty function defined by

$$p'(\theta) = I(\theta \leq \lambda) + \frac{(\alpha\lambda - \theta)_+}{(\alpha - 1)} I(\theta > \lambda) \text{ for}$$

some $\alpha > 2$ and $\theta > 0$,
improves the properties of the L_1-penalty and the hard-thresholding penalty function given by (2.1). We will call this penalty function as smoothly clipped absolute deviation (SCAD) penalty. This corresponds to a quadratic spline function with knots at λ and $\alpha\lambda$. This penalty function leaves large value of θ not excessively penalized and makes the solution continuous. The resulting solution is given by

$$\hat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+ & \text{when } |z| \leq 2\lambda; \\ \{(\alpha - 1)z - \text{sgn}(z)\alpha\lambda\} & \text{when } 2\lambda < |z| \leq \alpha\lambda; \\ \alpha - 2 & \\ z & \text{when } |z| > \alpha\lambda. \end{cases}$$

This solution is due to Fan[13]. The procedures using the SCAD penalty simply referred as SCAD.

2.3. Least Absolute Shrinkage and Selection Operator (LASSO) :

Suppose that we have data (x^i, y_i), i=1, 2, . . ., N, where $x^i=(x_{i1},..., x_{ip})^T$ are the predictor variables and y_i are the responses[16]. As in the usual regression set-up, we assume either that the observations are independent or that the y_is are conditionally independent given the x_{ij}s. We assume that the x_{ij} are standardized so that $\Sigma_i x_{ij}/N = 0$, $\Sigma_i x_{ij}^2/N = 1$.

Letting $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$, the lasso estimate $(\hat{\alpha}, \hat{\beta})$ is defined by

$$(\hat{\alpha}, \hat{\beta}) = \arg \min \left\{ \Sigma_{i=1}^N \left(y_i - \alpha - \Sigma_j \beta_j x_{ij} \right)^2 \right\}$$

subject to $\Sigma_j |\beta_j| \leq t$

Here $t \geq 0$ is a tuning parameter. For all t, the solution for is $\hat{\alpha} = \bar{y}$. For a detailed study refer to Tibshirani[27]

2.4. Akaike Information Criterion (AIC):

The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. AIC values provide a means for model selection. In the general case[18-21], the AIC is
AIC = 2k - 2 ln(L)
where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model. For a detailed study, refer to Akaike[1].

2.5. Bayesian Information Criterion (BIC):

The Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood

function, and it is closely related to Akaike information criterion (AIC). When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. Akaike was so impressed with Schwarz's Bayesian formalism that he developed his own Bayesian formalism, now often referred to as the ABIC for "a Bayesian Information Criterion" or more casually "Akaike's Bayesian Information Criterion".

Let $\{(y_i, x_i) : i = 1, \ldots, n\}$ be independent observations. Suppose that the conditional density function of y_i given x_i is $f(y_i \mid x_i, \theta)$, where $\theta \in \Theta R^P$, P being a positive integer. The likelihood function of θ is given by

$$L_n(\theta) = f(x;\theta) = \prod_{i=1}^{n} f(y_i|x_i,\theta)$$

where $Y = (y_1, \ldots, y_n)$. Let s be a subset of $\{1, \ldots, P\}$. Denote by $\theta(s)$ the parameter θ with those components outside s being set to 0 or some prespecified values. The BIC proposed by Schwarz[26] selects the model that minimizes

$$BIC(s) = -2 \log L_n\{\hat{\theta}(s)\} + \nu(s) \log n$$

where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$ and $\nu(s)$ is the number of components in s. For a detailed study, refer to Akaike[2], Schwarz[26]

2.6. Penalized Least Square and Penalized Likelihood :

Most variable selection procedures are related to penalized least squares. Suppose that we have the (d + 1) – dimensional random sample (x_i, y_i), i=1,…,n, from a population (x, y), where x is a d – dimensional random vector, and y is a continuous random variable. Consider a linear regression model

$$y_i = x_i^T\beta + \varepsilon_i$$

where β is unknown regression coefficients, and ε_i is a random error with mean zero and variance σ^2. Define a penalized least squares as

$$Q(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + n\sum_{j=1}^{d} p\lambda_{jn}\left(|\beta_j|\right) \quad (2.6.1)$$

where $p\lambda_{jn}(\cdot)$ is a given non-negative penalty function, and λ_{jn}'s are regularization parameters, which may depend on n and can be chosen by a data-driven criterion, such as cross-validation (CV) and generalized cross-validation (GCV), refer to Craven and Wahba[10]. Minimizing equ. (2.6.1) yields a penalized least square estimator. Conditioning on x_i, suppose that y_i has a density $f_i\{g(x_i^T\beta), y_i\}$, where g is a known link function. Let $\ell_i = \log f_i$ denote the conditional log-likelihood of y_i. Define a penalized likelihood as

$$\sum_{i=1}^{n} \ell_i(g(x_i^T\beta), y_i) - n\sum_{j=1}^{d} p\lambda_{jn}\left(|\beta_j|\right)$$

The penalized likelihood approach can be directly applied for parametric models in survival analysis. Let T, C and x be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T,C\}$ be the observed time and $\delta = I(T \le C)$ be the censoring indicator. It is

assumed that T and C are conditionally independent given x and that the censoring mechanism is non-informative. When the observed data $\{(x_i, Z_i, \delta_i): i = 1,\ldots,n\}$ is an independently and identically distributed random sample from a certain population (x, Z, δ), a complete likelihood of the data is given be

$$L = \prod_u f(Z_i|x_i) \prod_c \bar{F}(Z_i|x_i) = \prod_u h(Z_i|x_i) \prod_{i=1}^{n} \bar{F}(Z_i|x_i) \quad (2.6.2)$$

where the subscripts c and u denote the product of the censored and uncensored data respectively, and $f(t|x)$, $\bar{F}(t|x)$ and h(t|x) are the conditional density function, the conditional survival function and the conditional hazard function of T given x.

2.7. Generalized Cross Validation (GCV) estimate :

The generalized Cross Validation estimation of λ is the minimizer of V(λ)

$$V(\lambda) = \frac{(1/n)\left\|(I - A(\lambda))y\right\|^2}{\left[(1/n)\mathrm{tr}(I - A(\lambda))\right]^2}$$

where A(λ) is the n × n influence matrix, which statisfies

$$\begin{pmatrix} f_{n,\lambda}(t_1) \\ \vdots \\ f_{n,\lambda}(t_n) \end{pmatrix} = A(\lambda)y, \qquad y = (y_1, \ldots, y_n)$$

The GCV estimates the λ which minimizes the predictive mean square error R(λ) defined by

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(f(t_i) - f_{n,\lambda}(t_i)\right)^2$$

$f_{n,\lambda}(t)$, $t \in [0,1]$ is also a Bayes estimate of f(t), if f is endowed with a certain zero mean Gaussan prior, which is partially improper.

2.8. Cox's Proportional Hazard Model:

Let $t_1^0 < \cdots < t_N^0$ denote the ordered observed failure times. Let (j) provide the label for the item falling at t_j^0 so that the covariates associated with the N failures are $\mathbf{x}_{(1)},\ldots,\mathbf{x}_{(N)}$. Let R_j denote the risk set right before the time t_j^0:

$$R_j = \{i: Z_i \geq t_j^0\}$$

The Cox's proportional hazards models is given by,

$$h(t|\mathbf{x}) = h_0(t)\exp(x^T\beta) \quad (2.8.1)$$

with the baseline hazard functions $h_0(t)$ and parameter β. The likelihood discussed in equ. (2.6.2) of section 2.6, the likelihood becomes[25,28,30]

$$L = \prod_{i=1}^{N} h_0(Z_{(i)})\exp(x_{(i)}^T\beta)\prod_{i=1}^{n}\exp\{-H_0(Z_i)\exp(x_i^T\beta)\}$$

where $H_0(\cdot)$ is the cumulative baseline hazard function. If the baseline hazard function has a parametric form, $h_0(\theta,\cdot)$ say, then the corresponding penalized log-likelihood function is

$$\sum_{i=1}^{N}\left[\log\{h_0(\theta, Z_{(i)})\} + x_{(i)}^T\beta\right] - \sum_{i=1}^{N}\{H_0(\theta, Z_i)\exp(x_i^T\beta)\} - n\sum_{j=1}^{d} p\lambda\left(|\beta_j|\right) \quad (2.8.2)$$

Maximizing (2.8.2) with respect to (θ, β) yields the maximum penalized likelihood estimator.

2.9. Pseudo Likelihood :

A Pseudo likelihood is an approximation to the joint probability distribution of a collection of random variables. The practical use of this is that it can provide an approximation to the likelihood function of a set of observed data which may either provide a computationally simpler problem for estimation, or may provide a way of obtaining explicit estimates of model parameters.

Given a set of random variables X = X_1, X_2, …, X_n and a set E of dependencies between these random variables, where {X_i, X_j} ∈ E implies X_i is conditionally independent of X_j given X_i's neighbors, the pseudo likelihood of X = x =(x_1, x_2,…,x_n) is

$$\Pr(X - x) - \prod_i \Pr\left(X_i - x_i | X_j - x_j \; for \; all \; j \; for \; which \; \{X_i, X_j\} \epsilon E\right)$$

Here X is a vector of variables, x is a vector of values. The expression X = x above means that each variable X_i in the vector X has a corresponding value x_i in the vector x. The expression Pr(X = x) is the probability that the vector of variables X has values equal to the vector x. Because situations can often be described using state variables ranging over a set of possible values, the expression Pr(X = x) can therefore represent the probability of a certain state among all possible states allowed by the state variables. The Pseudo-log-likelihood is a similar measure derived from the above expression. Thus

$$\log \Pr(X = x) = \sum_i \log \Pr\left(X_i - x_i | X_j - x_j \; for \; all \; \{X_i, X_j\} \epsilon E\right)$$

One use of the pseudo-likelihood measure is as an approximation for inference about a Markov or Bayesian network, as the pseudo-likelihood of an assignment to X_i may often be computed more efficiently than the likelihood, particularly when the latter may require marginalization over a large number of variables.

2. Frailty and Marginal Hazard Model :
2.1. Frailty Model:

The popular approach to modeling correlated survival times is to use a frailty model. Consider the Cox proportional hazard frailty model, in which it is assumed that the hazard rate for the j^{th} subject in the i^{th} subgroup is

$$h_{ij}\left(t | x_{ij}, u_i\right) = h_0(t) u_i \, exp\left(x_{ij}^T \beta\right), \quad i = 1,2,\dots,n; j = 1,2,\dots,J_i \qquad (3.1.1)$$

where the u_i's are associated with frailties, and they are a random sample from some population. It is frequently assumed that given the frailty u_i, the data in the i^{th} group are independent. The most frequently used distribution for frailty is the gamma distribution due to its simplicity. Assume without loss of generality that the mean of frailty is 1 so that all parameters involved are estimable. For the gamma frailty model, the density of u is

$$g(u) = \frac{\alpha^\alpha u^{\alpha-1} exp\,(-\alpha u)}{\Gamma(\alpha)}$$

From equ. (2.6.2), the full likelihood of "pseudo-data" {(u_i,x_{ij},Z_{ij},δ_{ij}):i=1,2,…,n; j=1, 2,…, J_i} is

$$\prod_{i=1}^{n}\prod_{j=1}^{J_i}\left[\{h(z_{ij}|x_{ij},u_i)\}^{\delta_{ij}}\bar{F}(z_{ij}|x_{ij},u_i)\right]\prod_{i=1}^{n}g(u_i)$$

Integrating the full likelihood function with respect to $u_1,\ldots,u_n$, the likelihood of the observed data is given by

$$L(\beta,\theta) = exp\left\{\beta^T\left(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}x_{ij}\right)\right\}\prod_{i=1}^{n}\frac{\alpha^\alpha\prod_{j=1}^{J_i}\{h_0(z_{ij})\}^{\delta_{ij}}}{\Gamma(\alpha)\left\{\sum_{j=1}^{J_i}H_0(z_{ij})\,exp(x_{ij}^T\beta)+\alpha\right\}^{A_i+\alpha}} \quad (3.1.2)$$

where $\theta=(\alpha, H)$ and $A_i=\sum_{j=1}^{J_i}\delta_{ij}$. The log-likelihood of the observed data is

$$\begin{aligned}\ell_f(\beta,\theta) = \sum_{i=1}^{n}\Bigg\{&\sum_{j=1}^{J_i}\delta_{ij}\,log\,h(z_{ij})\\ &-\left[(A_i+\alpha)\,log\left\{\sum_{j=1}^{J_i}H_0(z_{ij})\,exp(x_{ij}^T\beta)+\alpha\right\}\right]\Bigg\}\\ &+\sum_{i=1}^{n}\left\{\beta^T\left(\sum_{j=1}^{J_i}\delta_{ij}x_{ij}\right)+\alpha\,log\,\alpha-log\,\Gamma(\alpha)\right\}\end{aligned} \quad (3.1.3)$$

To eliminate the nuisance parameter h(·), we again employ the profile likelihood method. Consider the "least informative" non parametric modeling for $H_0(\cdot)$:

$$H_0(z)=\sum_{l=1}^{N}\lambda_l I(z_l\le z) \quad (3.1.4)$$

where$\{z_1,\ldots,z_N\}$ are pooled observed failure times. Substituting equ.(3.1.4) in equ. (3.1.3), then differentiating it with respect to λ_l, l = 1,…,N, the root of the corresponding score functions should satisfy the following equations:

$$\lambda_l^{-1}=\sum_{i=1}^{n}\frac{(A_i+\alpha)\sum_{j=1}^{J_i}I(z_l\le z_{ij})\,exp(x_{ij}^T\beta)}{\sum_{k=1}^{N}\sum_{j=1}^{J_i}H_0(z_{ij})\,exp(x_{ij}^T\beta)+\alpha},\ \text{for l = 1, … , n} \quad (3.1.5)$$

the above solution does not admit a close form, neither does the profile likelihood function. However, the maximum profile likelihood can be implemented as follows. With initial values α, β and λ_I, update $\{\lambda_i\}$ from equ. (3.1.5) and obtain the penalized profile likelihood of equ. (3.1.3). with known $H_0(.)$ defined by equ. (3.1.4), maximize the penalized likelihood equ. (3.1.3) with respect to (α,β), and iterate between these two steps. When the Newton-

Raphson algorithm is applied to the penalized likelihood equ. (3.1.3), it involves the first two order derivatives of the gamma function, which may not exist for certain value of α. One approach to avoid this difficulty is the use of a grid of possible values for the frailty parameter α and finding the maxima over this discrete grid, as suggested by Nielsen *et. al.*[24].

3.2. Prediction and model error :

When the covariate x is random, if $\hat{\mu}(x)$ is a prediction procedure constructed using the present data, the prediction error is defined as

$$PE(\hat{\mu}) = E\{Y - \hat{\mu}(x)\}^2$$

where the expectation is only taken with respect to the new observation (x,Y). The prediction error can be decomposed as

$$PE(\hat{\mu}) = EVar(Y|x) + E\{(Y|x) - \hat{\mu}(x)\}^2$$

The first component is inherently due to stochastic errors. The second component is due to lack of fir to an underlying model. This component is called a model error and is denoted by $ME(\hat{\mu})$. For the Cox proportional hazards model (2.8.1),

$$\mu(x) = E(T|x) = \int_0^\infty h_0(t)\exp(x^T\beta)$$

$$\exp\left\{-\int_0^t h_0(t)\exp(x^T\beta)du\right\}dt$$

In the following simulation examples, it will be taken that $h_0(t) \equiv 1$. Thus by some algebra calculation,

$$\mu(x) = \exp(x^T\beta)$$

For the Cox frailty model with $h_0(t) \equiv 1$,

$$\mu(x) = \exp(x^T\beta)\, E(u^{-1})$$

The factor $E(u^{-1})$, due to the frailty, is dropped off when the performance of two different approaches is compared in terms of their Relative Model Errors (RME), defined as the ratio of the model errors of the two approaches. Therefore, the model error will be defined as

$$E\left\{\exp\left(-x^T\hat{\beta}\right) - \exp\left(-x^T\beta_0\right)\right\}^2$$

for both the Cox model and the frailty model.

3.3. Marginal Hazard Model :

As discussed in section 1, when the correlation among the observations is not of interest, the marginal proportional hazard models have received much attention in the recent literature because they are semi-parametric models and retain the virtue of the Cox model. Let T_{ik} be the k^{th} type of failure occurs on the i^{th} unit, and let C_{ik} be the corresponding censoring time. Define $X_{ik} = \min(T_{ik}, C_{ik})$ and $\Delta_{ik} = I(T_{ik} \leq C_{ik})$. Also, let $Z_{ik} = (Z_{1ik},\ldots, Z_{pik})'$ denote the covariate vector for the i^{th} unit with respect to the k^{th} type of failure. The failure time vector $T_i = (T_{i1},\ldots, T_{ik})$ and the censoring time vector $C_i = (C_{i1},\ldots,C_{ik})$ are assumed to be independent conditional on the covariates vector $Z_i = (Z_{i1}', \ldots, Z_{ik}')$ (i = 1,…, n). further assume that (X_i, C_i, A_i) (i = 1,…, n) are independent and identically distributed random elements.

If T_{ik} or Z_{ik} is missing, we set $C_{ik} = 0$, which ensures that $X_{ik} = 0$ and $\Delta_{ik} = 0$. It is natural to formulate the marginal distribution for each type of failure with a proportional hazard model. Depending on whether the baseline hazard functions are identical or are different among the M types of failures, the hazard function of the i^{th} unit for the k^{th} type of failure is

$$\lambda_k(t, Z_{ik}) = \lambda_0(t)e^{\beta' Z_{ik}(t)} \qquad (3.3.1)$$

where $\lambda_0(t)$ is unspecified baseline hazard functions, and $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown regression parameters. Then the 'partial likelihood functions' for β are

$$\tilde{L}(\beta) = \prod_{i=1}^{n}\prod_{k=1}^{M}\left\{\frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n}\sum_{l=1}^{M} Y_{jl}(X_{ik})e^{\beta' Z_{ik}(X_{ik})}}\right\}^{\Delta_{ik}}$$

The corresponding 'score functions' is

$$\tilde{U}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{M}\Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta, X_{ik})}{\bar{S}^{(0)}(\beta, X_{ik})}\right\}^{\Delta_{ik}}$$

where $\bar{S}^{(r)}(\beta,t) = \sum_{k=1}^{M} S_k^{(r)}(\beta,t)$, (r =0,1) and $S_k^{(0)}(\beta,t) = \sum_{j=1}^{n} Y_{jk}(t)e^{\beta' z_{jk}(t)}$, $S_k^{(1)}(\beta,t) = \sum_{j=1}^{n} Y_{jk}(t)e^{\beta' z_{jk}(t)} Z_{jk}(t)$, (k=1,…,M)it is observed that the unique estimator $\tilde{\beta}$ by solving $\{\tilde{U}(\beta) = 0\}$. Although observations are generally correlated within the same unit, the estimator $\tilde{\beta}$ can be proven to be consistent for β as long as the marginal models are correctly specified. The derivative matrix $-\frac{\partial^2 \log \tilde{L}(\beta)}{\partial \beta^2}\Big|_{\beta=\tilde{\beta}}$ however, does not provide a valid variance-covariance estimator for $\tilde{U}(\beta)$.

For large n and relatively small M, the statistic $\tilde{U}(\beta)$ is approximately p-variate normal with mean 0 and with (estimated) covariance matrix $\tilde{B}(\tilde{\beta}) = \sum_{i=1}^{n}\sum_{k=1}^{M}\sum_{l=1}^{M}\tilde{W}_{ik}(\tilde{\beta})\tilde{W}_{il}(\tilde{\beta})'$, where under equ. (3.3.1)

$$\tilde{W}_{ik}(\beta) = \Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta, X_{ik})}{\bar{S}^{(0)}(\beta, X_{ik})}\right\}$$

$$-\sum_{j=1}^{n}\sum_{l=1}^{M}\frac{\Delta_{jl}Y_{ik}(X_{jl})e^{\beta' Z_{ik}(X_{jl})}}{\bar{S}^{(0)}(\beta, X_{jl})}$$

$$\left\{Z_{ik}(X_{jl}) - \frac{\bar{S}^{(1)}(\beta, X_{jl})}{\bar{S}^{(0)}(\beta, X_{jl})}\right\}$$

Furthermore, the estimator $\tilde{\beta}$ is approxi-mately p-variate normal with mean β with (estimated) covariance matrix $\tilde{D}(\tilde{\beta}) = \tilde{A}^{-1}(\tilde{\beta})\,\tilde{B}(\tilde{\beta})\tilde{A}^{-1}(\tilde{\beta})$, where

$$\tilde{A}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{M}\Delta_{ik}\left\{\frac{\bar{S}^{(2)}(\beta, X_{ik})}{\bar{S}^{(0)}(\beta, X_{ik})} - \frac{\bar{S}^{(1)}(\beta, X_{ik})\bar{S}^{(1)}(\beta, X_{ik})'}{\bar{S}^{(0)}(\beta, X_{ik})^2}\right\}$$

under eq. (3.3.1)

$\tilde{A}(\beta) = -\frac{\partial^2 \log \tilde{L}(\beta)}{\partial \beta^2}$. In the case of M=1, the matrix $\tilde{D}(\tilde{\beta})$ reduces to the Lin-Wei robust variance-covariance estimator. If the marginal

models are correctly specified and if the observations' failure times within the same unit are independent, then $\tilde{B}(\tilde{\beta})$ is asymptotically equivalent to $\tilde{A}(\tilde{\beta}).\tilde{A}^{-1}(\tilde{\beta})$ and $\tilde{D}(\tilde{\beta})$ as, respectively, the naive and robust variance covariance estimators for $\tilde{\beta}$, and call $\tilde{U}'(0)\tilde{A}^{-1}(0)\tilde{U}(0)$ and $\tilde{U}'(0)\tilde{B}^{-1}(0)\tilde{U}(0)$ the naive and robust log-rank statistics, respectively. To test hypotheses involving several components of , the multivariate general linear hypothesis can be expressed as H_0: $L\beta = d$, where L is a $r \times p$ matrix of constants and d is a $r \times 1$ vector of constants. The robust Wald statistic for testing H_0 is $(L\tilde{\beta} - d)'\{L\tilde{D}(\tilde{\beta})L'\}^{-1}(L\tilde{\beta} - d)'$, which has an approximate χ^2 distribution with r degrees of freedom.

4. Numerical illustrations :

4.1. Numerical illustration for Frailty Model:

Following the approach of Morris *et. al.*[23], the proposed frailty model is applied to the Mahatma Gandhi Medical College and Research Institute hospital "Nursing Home", located at Pillaiyarkuppam, 14 km from Puducherry, near by district Head quarter Cuddalore of Tamil Nadu. A full description of this data set is as follows:

x_1 – treatment indicator

$$x_1 = \begin{cases} 1 \text{ if treated at a nursing home} \\ 0 \text{ otherwise} \end{cases}$$

x_2 – variable age

$$x_2 = \{k \; such \; that \; k \in (60,70)\}$$

x_3 – gender

$$x_3 = \begin{cases} 1 \text{ if Male} \\ 0 \text{ if Female} \end{cases}$$

x_4 – marital status

$$x_3 = \begin{cases} 1 \text{ if Married} \\ 0 \text{ otherwise} \end{cases}$$

x_5, x_6 and x_7 are three binary health status indicators, corresponding from the best health to the worst health. The model suggested by Morris *et. al.*[23] is

$$h(t|x) = h_0(t)\exp\left(\sum_{i=0}^{7} x_i\beta_i\right)$$

where $h_0(t)$ is the base line hazard function, using gamma frailty as discussed in section 3.1. using the algorithm suggested by Lin[22]. The Cox model is fitted with three parametric and the nonparametric baseline hazard models to this data set. Only x_2 is standardized as other variables are binary. Penalized partial likelihood approach with the SCAD, L_1 and hard penalty are applied to this data set. The thresholding parameter λ, selected by the GCV, is 0.02, 0.01 and 0.09 for the SCAD, LASSO and HARD, respectively. The best subset variable selection with AIC and BIC is also computed. For estimating the parameters, the algorithm and programme suggested by Lin (MULCOX, 1990)*, Lin (MULCOX2, 1993)** has been used. The Estimated coefficients and their standard errors are shown in Table 1.

From the above table is is observed that the age variable is not significant. However it is very significant when compare with interactions. It is evident from the above table

Table 1. Estimated Coefficients and Standard Errors

	MLE	Best (BIC)	Best (AIC)	SCAD	LASSO	HARD
TRT	-0.03(0.06)	0(–)	0(–)	0(–)	0(–)	0(–)
Age	-0.10(0.04)	0(–)	-0.07(0.02)	-0.07(0.03)	-0.03(0.01)	0(–)
Gender	0.38(0.09)	0.35(0.05)	0.39(0.07)	0.39(0.07)	0.26(0.04)	0.39(0.07)
Married	0.18(0.13)	0(–)	0.12(0.07)	0.14(0.07)	0.04(0.02)	0.14(0.07)
Health1	0.02(0.07)	0(–)	0(–)	0(–)	0(–)	0(–)
Health2	0.21(0.06)	0.21(0.05)	0.21(0.05)	0.21(0.05)	0.11(0.03)	0.19(0.05)
Health3	0.60(0.09)	0.56(0.08)	0.57(0.08)	0.57(0.08)	0.38(0.05)	0.58(0.08)
TRT*Age	0.10(0.05)	0(–)	0(–)	0(–)	0(–)	0(–)
TRT*Gender	-0.08(0.12)	0(–)	-0.13(0.10)	-0.14(0.10)	0(–)	-0.13(0.10)
TRT*Married	-0.01(0.15)	0(–)	0(–)	0(–)	0(–)	0(–)
Age*Gender	0.12(0.05)	0(–)	0.13(0.05)	0.12(0.05)	0.03(0.02)	0.05(0.04)
Age*Married	0.07(0.07)	0(–)	0(–)	0.07(0.07)	0(–)	0(–)
Gender*Married	-0.06(0.15)	0(–)	0(–)	0(–)	0(–)	0(–)

* a computer program for the Cox regression analysis of multiple failure time variables

** a general computer program for the Cox regression analysis of multivariate failure time data.

that elderly patients are likely stay at nursing home. The interaction between the variables treatment and gender selected by SCAD and HARD seems to be significant, although the treatment is not significant. It is clearly evident from the real life phenomena that, men prefer to stay at a nursing home with treatment, while elderly men like to leave a nursing home earlier. The result exhibit the same scenario as suggested by Morris *et. al.*[23].

4.2. Numerical illustration for Marginal Model :

The Diabetic Retinopathy study was conducted by the National Eye Institute to assess the effectiveness of laser photocoagulation in delaying the onset of blindness in patients with diabetic retinopathy (1981). Prevalence of Cataract Blindness in a rural Puducherry was conducted by the Mahatma Gandhi Medical College & Research Institute, Puducherry, to assess the cataract blindness among male and female patients (2011). Among the patients, the Diabetic Retinopathy has been identified. Between January 2010 to December 2011, 76 patients were entered the study. Following the approach of Huster et al. and Liang et al., the data were collected from Mahatma Gandhi Medical College and Research Institute, one eye of each patient was randomly selected for photocoagulation and the other eye was

Table 2. Estimates of Regression Parameters for the Diabetic Retinopathy Study*

Covariate	Methods			
	Naive	Robust	Liang	Huster
Treatment (Z_1)	-0.302 (0.158)	-0.302 (0.135)	-0.301 (0.135)	-0.31 (0.16)
Diabetic type (Z_2)	0.229 (0.125)	0.229 (0.122)	0.228 (0.122)	0.26 (0.13)
Interaction ($Z_1 \times Z_2$)	-0.715 (0.261)	-0.715 (0.214)	-0.713 (0.213)	-0.71 (0.26)

* The standard errors estimates are given in parentheses

observed without treatment. The patients were observed for the occurrence of blindness in the left and right eyes. One anticipates some dependence between a patient's two eyes.

Consider the model given in equ. (3.3.1) with $Z_{ik} = (Z_{1ik}, Z_{2ik}, Z_{3ik})'$ (i = 1,…,126; k = 1,2), where

$$Z_{1ik} = \begin{cases} 1 \; if \; the \; k^{th} \; eye \; of \; the \; i^{th} \; patient \; was \; on \; treatment, \\ 0 \; otherwise; \end{cases}$$

$$Z_{2ik} = \begin{cases} 1 \; if \; the \; i^{th} \; patient \; adult \; onset \; diabetes, \\ 0 \; if \; the \; i^{th} \; patient \; juvenile \; onset \; diabetes; \end{cases}$$

and $Z_{3ik} = Z_{1ik} \times Z_{2ik}$.

The estimates of regression parameters for the Diabetic Retinopathy study based on the data set is given in Table 2.
The robust standard error estimates are appreciably smaller than the naive estimates. The treatment appears to be effective, and this effect is much stronger for adult onset diabetes than for juvenile onset diabetes. The Liang et al. method produces very similar parameter estimates and the standard error estimates are almost identical to our robust ones.

Conclusion

The Liang *et al.* method produces similar results comparing to other methods and almost identical to robust ones. In the case of Huster et al., the estimates are fairly close to the naive estimates. The marginal approach is expected to be more efficient than the Frailty model provided that the Frailty distribution is correctly specified. However the types of dependence by the Frailty model are quite limited and fitting is rather difficult, cumbersome.

References

1. Akaike, H., "A new look at the statistical model identification". *IEEE Transactions on Automatic Control, 19(6)*, 716–723 (1974).
2. Akaike, H., "On entropy maximization principle". *Applications of Statistics*, North-Holland, Amsterdam, 27–41 (1977).
3. Androulakis, E., Koukouvinos, C., Mylona, K. and Vonta, F., A real survival analysis application via variable selection methods

for Cox´s proportional hazards model, *Journal of Applied Statistics*, *47,* No 8, 1399-1406 (2010).

4. Antoniadis, A., Wavelets in statistics: A review (with discussion). *J. Italian Statist. Soc. 6,* 97-144 (1997).
5. Besag, J., Statistical Analysis of Non-Lattice Data. *The Statistician, 24(3),* 179–195 (1975).
6. Breiman, L., Heuristics of instability and stabilization in model selection. *Ann. Statist. 24,* 2350–2383 (1996).
7. Cai, J., Fan, J., Li, R. and Zhou, H., Variable selection for multivariate failure time data. *Biometrika, 92,* 303-316 (2005).
8. Cox D.R., Regression model and life tables (with discussion). *Journal of the Royal Statistical Society*, B *34,* 187-220 (1972).
9. Cox D.R., Partial likelihood, *Biometrika, 62,* 269-296 (1975).
10. Craven, P. and Wahba, G., Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math., 31,* 377-403 (1979).
11. Diabetic Retinopathy Study Research Group, Diabetic retinopathy study. *Investigative Ophthalmology and Visual Science, 21,* Part 2, 149-226 (1981).
12. Fan, J. and Li, R., Variable selection for Cox's proportional hazards model and frailty model. *Ann Statist, 30,* 74–99 (2002).
13. Fan, J., Comment on "Wavelets in statistics: a review" by A. Antoniadis. *J. Italian Statist. Assoc. 6,* 131–138 (1997).
14. Fan, J. and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc., 96(a),* 1348–1360 (2001a).
15. Fan, J. and Li, R., Variable selection for Cox's proportional hazards model and frailty model. *Institute of Statistic Mimeo Series #2372, Dept. Statistics*, Univ. North Carolina, Chapel Hill (2001b).
16. George, E. and McCulloch, R., Variable selection via Gibbs sampling, *Journal of the American Statistical Association, 88,* 884-889 (1993).
17. Goeman, J. J., L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, *52,* 70–84 (2010).
18. Huster, W.J., Brookmeyer, R. and Self, S.G., Modelling paired survival data with covariates, *Biometrics, 45,* 145-156 (1989).
19. Kanagarajan, P., Partha Nandi, and Lokeshmaran, A., Prevalence of Cataract Blindness in a Rural Puducherry, *Indian Medical Gazatte*, pp 348-352 (2011).
20. Liang, K. Y., Self, S.G. and Chang, Y.C., Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society*, B, 55, 441-453 (1993).
21. Lin, D.Y., MULCOX: a computer program for the Cox regression analysis of multiple failure time variables. *Computer Methods and Programs in Biomedicine, 32,* 125-135 (1990).
22. Lin, D.Y., MULCOX2: a general computer program for the Cox regression analysis of multiple failure time variables. Computer

Methods and Programs in Biomedicine, *40*, 279-293 (1993).

23. Morris, C.N., Norton, E.C. and Zhou, X.H. Parametric duration analysis of nursing home usage. *In case studies in Biometry, Wiley*, New York, 231-248 (1994).
24. Nielsen, G.G., Gill, R.D., Andersen, P.K. and Sorensen, T.I.A., A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics*, *19*, 25-43 (1992).
25. Parner, E., Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist. 26*, 183–214 (1998).
26. Schwarz G. E., Estimating the dimension of a model. *Annals of Statistics*, *6(2)*, 461-464 (1978).
27. Tibshirani, R., Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, *Series B*,58, 267-288 (1996).
28. Tibshirani, R., The lasso method for variable selection in the Cox model. *Statist Med*, *16*, 385–95 (1997).
29. Volinsky, C.T. and Raftery, A.E., Bayesian information criterion for censored survival models. Biometrics, *56*, 256-262 (2000).
30. Wahba, G., A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, 13, 1378–1402 (1985).