Science in China Series A: Mathematics
Jun., 2009, Vol. 52, No. 6, 1131–1133
www.scichina.com math.scichina.com
www.springerlink.com

A note on Optimal weights and variable selections for multivariate survival data

Zhao Sihai Dave & LI Yi

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, MA 02115, USA
(email: szhao@fas.harvard.edu, yili@jimmy.harvard.edu)

Fan et al. are to be congratulated for this important contribution to the analysis of multivariate failure time data. They have provided three regression parameter estimators for multiple covariates in the marginal hazard model. Using the weighted estimating equation approach, they proposed sets of weights to:

1) Minimize the componentwise variance of each parameter estimate,

2) Minimize the sum of the variances of each parameter estimate, or

3) Minimize each element of the covariance matrix.

They showed that each of their proposed estimators can consistently outperform estimates derived using the working independence model.

In this short note, we show that in the presence of high-dimensional covariates Fan et al.'s ideas can be combined with those of [1] to achieve these optimal estimates along with simultaneous variable selection. That is, our interest lies in controlling the variances of the estimates of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ associated with high dimensional covariates, while simultaneously selecting the "important" covariates in order to construct a parsimonious model. Here, we consider p as a large but fixed constant, as opposed to [1] which considered the situation where p may increase with the sample size n.

The key idea is to add a penalty function $p_{\lambda_j}(|\beta_j|)$ to Fan et al.'s weighted partial likelihood function (10), which leads to the following penalized likelihood function

$$l_W^P(\beta) = \sum_{j=1}^{J} w_j l_j(\beta) - n \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|). \tag{1}$$

Denote by $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}}$ the true value of $\boldsymbol{\beta}$ and suppose, without loss of generality, that $\beta_{0k} \neq 0$, $k \leqslant s$ and $\beta_{0k} = 0$, $k > s$ for some $s \leqslant p$. Let $\boldsymbol{\beta}_{01} = (\beta_{01}, \ldots, \beta_{0s})^{\mathrm{T}}$. Finally, let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)^{\mathrm{T}}$ be the solution that maximizes (1) such that $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_1, \ldots, \widehat{\beta}_s)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}_2 = (\widehat{\beta}_{s+1}, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$.

With an appropriate penalty function, our estimator $\widehat{\boldsymbol{\beta}}$ may enjoy the oracle property. That is, the procedure should select the true model with probability tending to 1 and, given the true model, the coefficient estimates should asymptotically behave like maximum (partial) likelihood

This work was supported by U. S. National Cancer Institute (Grant No. R01 CA95747)
Jun., 2009, Vol. 52, No. 6, 1131–1133 DOI: 10.1007/s11425-009-0134-3

estimators. More specifically, consider

$$a_n = \max_{1\leqslant j\leqslant s}\{|p'_{\lambda_{jn}}(|\beta_{j0}|)| : \beta_{j0}\neq 0\},$$
$$b_n = \max_{1\leqslant j\leqslant s}\{|p''_{\lambda_{jn}}(|\beta_{j0}|)| : \beta_{j0}\neq 0\}.$$

Here, since λ_j depends on n, we write it as λ_{jn}. Along the lines of Theorem 2 of [1], we can show that if the penalty function is such that $a_n = O(n^{-\frac{1}{2}})$ and $b_n \to 0$, and in addition $\lambda_{jn}\to 0$ and $\lambda_{jn}\sqrt{n}\to 0$,

$$\begin{aligned}&\widehat{\boldsymbol{\beta}}_2 = 0 \text{ with probability approaching 1, and}\\ &n^{1/2}(\mathbf{A}_{W11}+\boldsymbol{\Sigma}_{11})\{\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}+(\mathbf{A}_{W11}+\boldsymbol{\Sigma}_{11})^{-1}\mathbf{b}_n\}\to_d N(0,\mathbf{V}_{W11}),\end{aligned}\tag{2}$$

where

- $\mathbf{A}_{W11}$ is the first $s\times s$ submatrix of $\mathbf{A}_W(\boldsymbol{\beta})=\sum_{j=1}^J w_j\boldsymbol{\Sigma}_j(\boldsymbol{\beta})$,
- $\boldsymbol{\Sigma}_{11}$ is the first $s\times s$ submatrix of $\mathrm{diag}\{p''_{\lambda_{1n}}(|\beta_{01}|)\mathrm{sgn}(\beta_{01}),\ \ldots,\ p''_{\lambda_{pn}}(|\beta_{0p}|)\mathrm{sgn}(\beta_{0p})\}$,
- $\mathbf{V}_{W11}$ is the first $s\times s$ submatrix of $\boldsymbol{\Sigma}_W(\boldsymbol{\beta})$, which was defined in Fan et al.'s (11), and
- $\mathbf{b}_n=(p'_{\lambda_{1n}}(|\beta_{01}|),\ldots,p'_{\lambda_{sn}}(|\beta_{0s}|))^{\mathrm{T}}$.

Therefore, by (2) we have that the asymptotic covariance of $\sqrt{n}\widehat{\boldsymbol{\beta}}_1$ is

$$\boldsymbol{\Sigma}_W^{*P}=(\mathbf{A}_{W11}+\boldsymbol{\Sigma}_{11})^{-1}\mathbf{V}_{W11}(\mathbf{A}_{W11}+\boldsymbol{\Sigma}_{11}),\tag{3}$$

where $\boldsymbol{\Sigma}_W^{*P}$ is the penalized version of $\mathrm{var}(\widehat{\boldsymbol{\beta}}_W)$ defined in section 2.3 of Fan et al. One example of an appropriate $p_{\lambda_j}(\theta)$ is the smoothly clipped absolute deviation penalty of [2], where

$$p'_\lambda(\theta)=\lambda I(\theta\leqslant\lambda)+\frac{(a\lambda-\theta)_+}{a-1}I(\theta>\lambda).$$

More examples can be found in [3].

We can now follow Fan et al. and define optimality criteria that will allow us to simultaneously estimate the optimal weights $\mathbf{w}=(w_1,\ldots,w_J)$. Minimizing the component-wise variance may not be ideal because minimizing the variance of the $\widehat{\beta}_j, j>s$ is irrelevant if the true $\beta_{0j}=0$. Minimizing the variance of any arbitrary linear function of the parameter estimates is also not always feasible, as was explained in Subsection 3.3 of Fan et al. Hence, we focus on minimizing the total variance:

$$\min_{\mathbf{w}}\mathrm{tr}(\boldsymbol{\Sigma}_W^{*P}),$$

analogous to (14). Following the derivation in Subsection 3.2, we assume that $\boldsymbol{\Sigma}_j(\beta)\approx b_j\boldsymbol{\Gamma}$ for some $\boldsymbol{\Gamma}$. Thus, if we constrain $\sum_{j=1}^J w_jb_j=1$, we are to minimize

$$\mathrm{tr}\left(\sum_{j=1}^J w_j^2[\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1}\boldsymbol{\Sigma}_{j11}[\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1}+\sum_{k\neq l}^J w_kw_l[\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1}\mathbf{D}_{kl11}[\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1}\right)$$

over $\mathbf{w}$, where $\boldsymbol{\Gamma}_{11}$ and $\mathbf{D}_{kl11}$ are the first $s\times s$ submatrices of $\boldsymbol{\Gamma}$ and $\mathbf{D}_{kl}(\beta)$, respectively. If $\mathbf{H}^P$ is a symmetric matrix with diagonal elements $\mathrm{tr}([\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1}\boldsymbol{\Sigma}_{j11}[\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1})$ and off-diagonal elements $\mathrm{tr}([\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1}\mathbf{D}_{kl11}[\boldsymbol{\Gamma}_{11}+\boldsymbol{\Sigma}_{11}]^{-1})$, the solution is given by

$$\mathbf{w}=(\mathbf{H}^P)^{-1}\mathbf{b}/\mathbf{b}^{\mathrm{T}}(\mathbf{H}^P)^{-1}\mathbf{b},\tag{4}$$

where $\mathbf{b} = (b_1, \ldots, b_J)^{\mathrm{T}}$. $\boldsymbol{\Gamma}_{11}$ can be estimated by $\boldsymbol{\Gamma}_{11} = \frac{1}{J}\sum_{j=1}^{J} \widehat{\boldsymbol{\Sigma}}_{11}$, and suggestions for possible choices for $\mathbf{b}$ were given in Subsection 3.2.

Finally, we can choose the parameters λ_j by iteratively minimizing the generalized cross-validation statistic of Cai et al. and solving for the optimal weights w_j. Let

$$\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}) = \mathrm{diag}(p'_{\lambda_1}(|\widehat{\beta}_1|/|\widehat{\beta}_1|), \ldots, p'_{\lambda_p}(|\widehat{\beta}_p|)/|\widehat{\beta}_p|)),$$

$$e(\lambda_1, \ldots, \lambda_p) = \mathrm{tr}\left\{\left[\sum_{j=1}^{J} w_j l''_j(\widehat{\beta}) - \boldsymbol{\Sigma}_\lambda\right]^{-1}\left[\sum_{j=1}^{J} w_j l''_j(\widehat{\beta})\right]\right\}.$$

If $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$, we choose

$$\boldsymbol{\lambda} = \mathrm{argmin}_{\boldsymbol{\lambda}}\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{-\sum_{j=1}^{J} w_j l_j(\widehat{\beta})}{n[1 - e(\boldsymbol{\lambda})/n]^2}. \tag{5}$$

We can first assume the working independence model to choose the initial λ_j with (5), then use those λ_j to solve for the w_j with (4). We can then iterate between (5) and (4) until λ_j and w_j converge.

To conclude, we have suggested a way in which the work of Fan et al. can be extended to perform variable selection on models for multivariate survival times with high dimensional covariates, while simultaneously providing optimally efficient parameter estimates for the selected covariates. Our future direction lies in utilizing empirical date and simulations to evaluate the accuracy and stability of the variable selection process, as well as the performance of the estimates derived using these optimal weights.

References

1 Cai J, Fan J, Li R, et al. Variable selection for multivariate failure time data. *Biometrika*, **92**: 303–316 (2005)

2 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, **96**: 1348–1360 (2001)

3 Johnson B A, Li D Y, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. *J Amer Statist Assoc*, **103**: 672–680 (2008)