

SUPPLEMENTAL APPENDIX

Nguyen,D.V. and Rocke,D.M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*.

SUPPLEMENTAL APPENDIX A

Algorithm for PLS Optimization

The following PLS algorithm can be used to obtain the PLS weights and other related quantities in PLS. We note that PLS is not invariant to scaling of the variables. Wold *et al.* (1984) recommend that variables be standardized in the absence of prior information about the importance of the variables, similar to principal components analysis. Thus, the input data sets are standardized to mean 0 and variance 1: $x_{ij} \leftarrow (x_{ij} - m_j^x)/s_j^x$ and $y_{ij} \leftarrow (y_{ij} - m_j^y)/s_j^y$. The test data is standardized using training data: $x_{ij}^* \leftarrow (x_{ij}^* - m_j^x)/s_j^x$. Here, m_j^x and s_j^x denote the sample mean and standard deviation calculated from the training data.

PLS Algorithm

1. Input training data pair (\mathbf{X}, \mathbf{Y}) and test data \mathbf{X}^* .
2. Input the number of PLS components, K_P .
3. Set convergence criterion ϵ , say $\epsilon = 1\text{E-}12$ and set $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{Y}_1 = \mathbf{Y}$.
4. FOR $k = 1$ to K_P DO
 - Set \mathbf{u} to be the first column of \mathbf{Y}_k and initialize δ .
 - a. WHILE $\delta > \epsilon$ DO
 - $\mathbf{w} = \mathbf{X}_k' \mathbf{u} / \mathbf{u}' \mathbf{u}$ and scale \mathbf{w} to unit length.
 - $\mathbf{t} = \mathbf{X}_k \mathbf{w}$. PLS components.
 - $\mathbf{c} = \mathbf{Y}_k' \mathbf{t} / \mathbf{t}' \mathbf{t}$ and scale \mathbf{c} to unit length.
 - $\mathbf{u} = \mathbf{Y}_k \mathbf{c}$.
 - $\delta = (\mathbf{w} - \mathbf{w}_{\text{prev}})'(\mathbf{w} - \mathbf{w}_{\text{prev}})$, \mathbf{w}_{prev} is the previous value of \mathbf{w} .
 - END
 - b. Compute and save relevant quantities:
 - $\mathbf{c}_k = \mathbf{c}$.
 - $\mathbf{p}_k = \mathbf{X}_k' \mathbf{t} / (\mathbf{t}' \mathbf{t})$ and scale \mathbf{p}_k to unit length.
 - $\mathbf{t}_k = \mathbf{t} c_p$, $c_p = (\mathbf{p}_k' \mathbf{p}_k)^{0.5}$.
 - $\mathbf{w}_k = \mathbf{w} c_p$.
 - $b_k = \mathbf{u}' \mathbf{t} / (\mathbf{t}' \mathbf{t})$.
 - Residual Matrices: $\mathbf{X}_{k+1} = \mathbf{X}_k - \mathbf{t}_k \mathbf{p}_k'$, $\mathbf{Y}_{k+1} = \mathbf{Y}_k - b_k \mathbf{t}_k \mathbf{c}_k'$.
 - END
5. Compute test components based on training information from 5.
 - Set $\mathbf{X}_1^* = \mathbf{X}^*$ and compute $\mathbf{t}_1^* = \mathbf{X}_1^* \mathbf{w}_1$. Subsequent test components are computed as $\mathbf{t}_k^* = \mathbf{X}_k^* \mathbf{w}_k$ where $\mathbf{X}_k^* = \mathbf{X}_{k-1}^* - \mathbf{t}_{k-1}^* \mathbf{p}_{k-1}'$, for $k = 1, \dots, K_P - 1$.

SUPPLEMENTAL APPENDIX B

Proportional Hazard Regression

The proportional hazard (PH) regression model is

$$h(t; \mathbf{x}_i; \boldsymbol{\beta}) = h_0(t) \exp(\mathbf{x}_i' \boldsymbol{\beta}) \quad (1)$$

where h is the hazard function associated with covariate \mathbf{x}_i and h_0 is an unspecified baseline hazard function. The hazard function $h(x)$ is defined as

$$h(x) = \lim_{\Delta x \rightarrow 0} P(x \leq X < x + \Delta x | X \geq x) / \Delta x \quad (2)$$

where $X \sim f(x)$. For X continuous, we have the relation $h(x) = f(x)/S(x)$, where $S(x) = P(X > x)$ is the survival function. The survival function under the PH model is

$$S(t; \mathbf{x}_i; \boldsymbol{\beta}) = [S_0(t)]^{\exp(\mathbf{x}'_i \boldsymbol{\beta})}, \quad (3)$$

where $S_0(t)$ is the baseline survival function.

Maximizing Partial Likelihood and Newton-Raphson

The PH model (1) is $h(t; \mathbf{x}_i; \boldsymbol{\beta}) = h_0(t) \exp(\mathbf{x}'_i \boldsymbol{\beta})$. The log partial likelihood (Cox, 1972) with no tied survival times) is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^D \mathbf{x}'_{(i)} \boldsymbol{\beta} - \sum_{i=1}^D \log \left[\sum_{j \in R(t_{(i)})} \exp(\mathbf{x}'_j \boldsymbol{\beta}) \right]$$

where $t_{(1)} < \dots < t_{(D)}$ are the ordered survival times with corresponding covariates $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(D)}$ and $R(t_{(i)})$ is the risk set consisting of all individuals with survival or censored times just prior to (i.e., \geq) time $t_{(i)}$.

For computational purposes we express the log partial likelihood in terms of all the observations $i = 1, \dots, N$. Let $\mathbf{D}_{N \times N}$ be the risk indicator matrix with ij th element $d_{ij} = I(t_j \geq t_i)$ and $\mathbf{d}'_i = (d_{i1}, \dots, d_{iN})$ be the i th row of \mathbf{D} . Let $\boldsymbol{\delta}' = (\delta_1, \dots, \delta_N)$ be the vector of censoring indicators, $\delta_i = I(t_i = \min(y_i, z_i))$. The log partial likelihood given above is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^N \delta_i \left[\mathbf{x}'_i \boldsymbol{\beta} - \log \left(\sum_{j=1}^N d_{ij} \exp(\mathbf{x}'_j \boldsymbol{\beta}) \right) \right].$$

Let $\mathbf{u}_{N \times 1}$ be a vector with element $u_j = \exp(\mathbf{x}'_j \boldsymbol{\beta})$ and $\mathbf{a}_{N \times 1} = \mathbf{D}\mathbf{u}$. Also, define \mathbf{w}_i to be an $N \times 1$ vector with element $w_{ij} = (1/a_i) u_j d_{ij}$ (i.e., $\mathbf{w}_i = (1/a_i) \mathbf{u} * \mathbf{d}_i$ where the $*$ denotes elementwise multiplication) and $\mathbf{W}_i = \text{diag}\{\mathbf{w}_i\}$ be the $N \times N$ diagonal matrix with diagonal elements as elements of \mathbf{w}_i . With these notations the log partial likelihood $l(\boldsymbol{\beta})$, the score vector $S(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$, and the information matrix $I(\boldsymbol{\beta}) = \partial S(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}'$ can be expressed, respectively, as

$$\begin{aligned} l(\boldsymbol{\beta}) &= \boldsymbol{\beta}' \mathbf{X}' \boldsymbol{\delta} - \boldsymbol{\delta}' \mathbf{a}, \\ S(\boldsymbol{\beta}) &= \sum_{i=1}^N \delta_i (\mathbf{x}_i - \mathbf{X}' \mathbf{w}_i) \text{ and} \\ I(\boldsymbol{\beta}) &= \sum_{i=1}^N \delta_i [\mathbf{X}' \mathbf{W}_i \mathbf{X} - (\mathbf{X}' \mathbf{w}_i)(\mathbf{w}'_i \mathbf{X})]. \end{aligned}$$

Note that the dependence on β is in \mathbf{a} , \mathbf{w} , and \mathbf{W} through $u = \exp(\mathbf{x}'\beta)$. These forms facilitate computations in the Newton-Raphson algorithm to compute (update) β , given by $\beta^{(s+1)} = \beta^{(s)} + I^{-1}(\beta^{(s)})S(\beta^{(s)})$. If the Newton-Raphson algorithm converges, then the vector of coefficients is denoted $\hat{\beta}$ and it is the maximum partial likelihood estimate (MPLE) of β .

Estimating the Survival Distribution for the PH Model

The baseline survival distribution estimate for the PH regression model is based on the product limit estimate (Kalbfleisch and Prentice, 1973). It is based on ML method and conditional on the MPLE of β from the PH model. The MLE estimate $\hat{\alpha}_i$ of α_i is obtained numerically from

$$\sum_{k \in F_i} \frac{\hat{u}_k}{1 - \hat{\alpha}_i^{\hat{u}_k}} = \sum_{l \in R(t_{(i)})} \hat{u}_l$$

where $\hat{u}_k = \exp(\mathbf{x}'_k \hat{\beta})$, F_i is the set of individuals failing at time $t_{(i)}$, and $R(t_{(i)})$ is the risk set at time $t_{(i)}$. In the case where there are no tied survival times, the set F_i contains only one individual and the solution to the above equation can be solved analytically as

$$\hat{\alpha}_i = \left[1 - \left(\hat{u}_i / \sum_{l \in R(t_{(i)})} \hat{u}_l \right) \right]^{\hat{u}_i^{-1}}.$$

The estimate of the baseline survival function is

$$\hat{S}_0(t) = \prod_{t_{(i)} \leq t} \hat{\alpha}_i.$$

REFERENCES

- Kalbfleisch, J.D., and Prentice, R.L. (1973) Marginal likelihoods based on Cox's regression and like model. *Biometrika*, **60**, 267–78.
- Wold, S., Ruhe, A., Wold, H., and Dunn, W.J. (1984) The collinearity problem in linear regression: the partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, **5**, 735–743.