Learning and Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais (SBRN), Vol. 3, No. 2, pp. 50-60,2005

INVITED TUTORIAL

SURVIVAL ANALYSIS NEURAL NETWORKS

Pereira, B. de B. and Rao C. R.

Abstract – In this paper we review some recent advances on survival models which generalizes the Cox model. A review of neural networks used for survival data are presented

1. Survival Analysis Models

If *T* is a non-negative random variable representing the time to failure or death of an individual, we may specify the distribution of *T* any one of the probability density function *f(t)*, the cumulative distribution function *F(t)*, the survivor function *S(t)*, the harzard function *f(t)* or the cumulative hazard function *H(t)*. These are related by:

$$F(t) = \int_0^t f(u)\,du$$

$$f(t) = F'(t) = \frac{d}{dt} F(t)$$

$$S(t) = 1 - F(t)$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left[\ln S(t)\right] \qquad (1.1)$$

$$H(t) = \int_0^t h(u)\,du$$

$$h(t) = H'(t)$$

$$S(t) = \exp[-H(t)].$$

A distinct feature of survival data is the occurrence of incomplete observations. This feature is known as *censoring* which can arise because of time limits and other restrictions depending of the study.

There are different types of censoring.

- Right censoring occurs if the event is not observed before the pre-specified study-term or some competitive event (e.g. death by other cause) that causes interruption of the follow-up on the individual experimental unit.
- Left censoring happens if the starting point is located before the time of the beginning of the observation for the experimental unit (e.g. time of infection by HIV virus in a study of survival of AIDS patients).
- Interval censoring the exact time to the event is unknown but it is known that it falls in an interval I_i (e.g. when observations are grouped).

The aim is to estimate the previous functions from the observed survival and censoring times. This can be done either by assuming some parametric distribution for *T* or by using non-parametric methods. Parametric models of survival distributions can be fitted by maximum likelihood techniques. The usual non-parametric estimator for the survival function is the Kaplan-Meier estimate. When two or more group of patients are to be compared the log-rank or the Mantel-Hanszel tests are used.

General class of densities and the non-parametric procedures with estimation procedure are described in Kalbfleish and Prentice (2002).

Usually we gave covariates related to the survival time *T*. The relation can be linear $\left(\beta \underset{\sim}{x}\right)$ or non-liner $\left(g\left(j; \underset{\sim}{x}\right)\right)$. A general class of models relating survival time and covariates is studied in Louzada-Neto (1977, 1999). Here we describe the three most common particular cases of the Louzada-Neto model.

The first class of models is the *accelerated failure time* (AFT) models

$$\log T = -\beta \underset{\sim}{e}' \underset{\sim}{x} + W \tag{1.2}$$

where W is a random variable. Then exponentiation gives

$$T = \exp\left(-\beta' \underset{\sim}{x}\right) e^{w} \text{ or } T' = e^{w} = T \exp\left(\beta' \underset{\sim}{x}\right) \tag{1.3}$$

where T´ has hazard function h_0 that does not depend on β. If $h_j(t)$ is the hazard function for the j´th patient it follows that

$$h_j(t) = h_0(t \exp \beta' x) \exp \beta' x \tag{1.4}$$

The second class is the *proportional odds* (PO) where the regression is on the log-odds of survival, correspondence to a linear logistic model with "death" or not

$$\log \frac{S_j(t)}{1 - S_j(t)} = \beta \underset{\sim}{x} + \log \frac{S_0(t)}{1 - S_0(t)} \tag{1.5}$$

or

$$\frac{S_j(t)}{1 - S_j(t)} = \frac{S_0(t)}{1 - S_0(t)} \exp \beta \underset{\sim}{x} \tag{1.6}$$

The third class if the "proportional hazard" or Cox regression model (PH).

$$\log h_t(t) = \beta \underset{\sim}{x} + \log h_0(t) \tag{1.7}$$

$$h_j(t) = h_0(t) \exp \beta \underset{\sim}{x} \tag{1.8}$$

Ciampi and Etezadi-Amoli (1985) extended models (1.2) and (1.7) under a mixed model and not only extend these models but also puts the three models under one more general comprehensive model (Louzada-Neto and Pereira, 2000). See Figure 6.1.

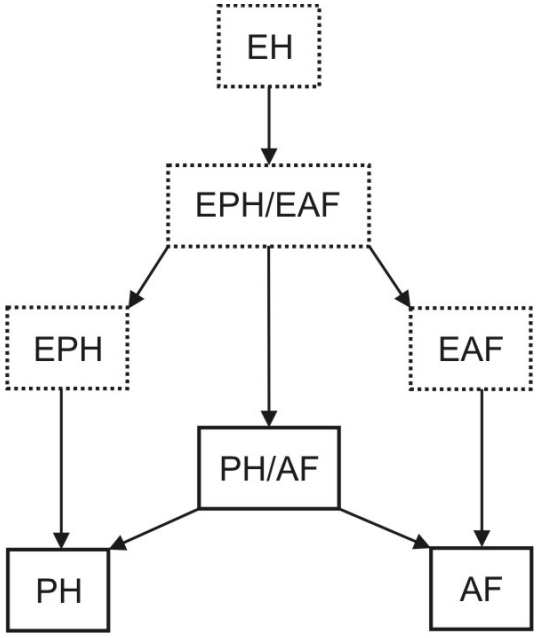

Figure1: Classes of regression model for survival data

2. Neural Networks Model For Survival Analysis

Ripley (1988) investigated seven neural networks in modeling breast cancer prognosis; her models were based on alternative implementation of models (1.2) to (1.8) allowing for censoring. There we outline the important results of the literature.

The accelerated failure time – AFT model is implemented using the architecture of regression network with the censored times estimated using some missing value method as in Xiang et al (2000).

For the Cox proportional hazard mode, Faraggi and Simon (1995) substitute the linear function βx_j by the output $f(x_j, \theta)$ of the neural network, that is

$$L_c(\theta) = \prod_{i\in} \frac{\exp\left\{\sum_{h=1}^{H} \alpha_h / \left[1+\exp\left(-w_h' x_i\right)\right]\right\}}{\sum_{j\in R_i} \exp\left\{\sum_{h=1}^{H} \alpha_h / \left[1+\exp\left(-w_h' x_i\right)\right]\right\}} \tag{1.9}$$

and estimations are obtained by maximum likelihood through Newton-Raphson.

The corresponding network is shown in Figure 2.

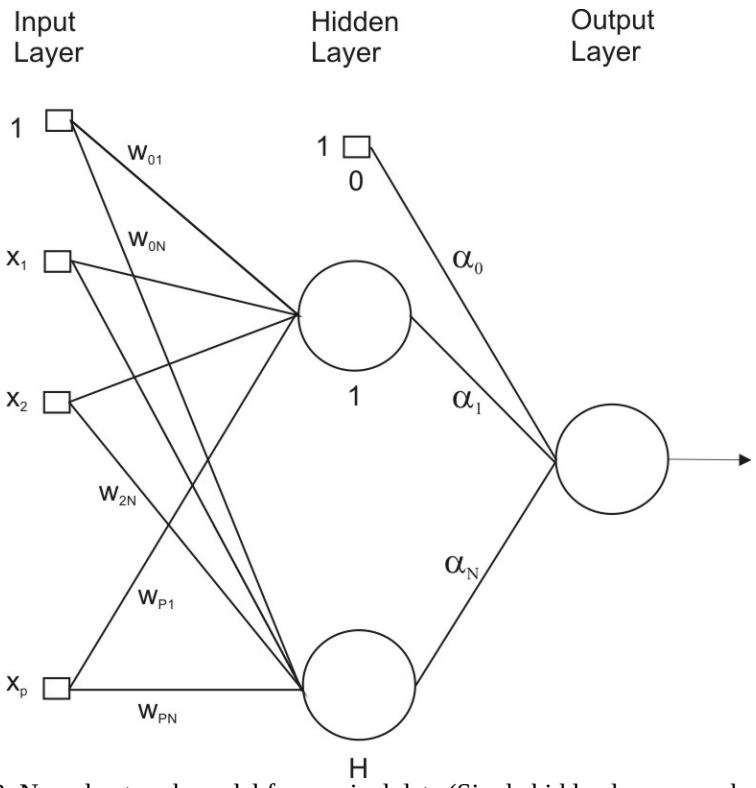

Figure 2: Neural network model for survival data (Single hidden layer neural network)

As an example (Faraggi and Simon, 1995) consider the data related to 506 patients with prostate cancer in stage 3 and 4. The covariates are: stage, age, weight, treatment (0.2; 1 or 5 mg of DES and placebo).

The results are given in the tables 1, 2, 3 below:

(a) First-order PH model 4 (parameters);
(b) Second-order (interactions) PH model (10 parameters);
(c) Neural network model with two hidden nodes (12 parameters);
(d) Neural network model with three hidden nodes (18 parameters).

Table 1 – Summary statistics for the factors included in the models

	Complete Data	**Training Set**	**Validation Set**
Sample size	475	238	237
Stage 3	47.5%	47.6%	47.4%
Stage 4	52.5%	52.4%	52.6%
Median age	73 years	73 years	73 years
Median wight	98-0	97-0	99-0
Treatment: Low	49.9%	48.3%	51.5%
Treatment: High	50.1%	51.7%	48.5%
Median survival	33 months	33 months	34 months
% censoring	28.8%	29.8%	27.7%

Table 2 – Log-likelihood and c statistics for first-order, second-order and neural network proportional hazards models

		Training Data		Test Data	
Model	**Number of Parameters**	**Log lik**	**c**	**Log lik**	**c**
First order PH	4	-814.3	0.608	-831.0	0.607
Second-order PH	10	-805.6	0.648	-834.8	0.580
Neural network H = 2	12	-801.2	0.646	-834.5	0.6000
Neural network H = 3	18	-794.9	0.661	-860.0	0.582

Table 3 – Estimation of the main effects and higher order interactions using 2^4 factorial design contrasts and the predictions obtained from the different models

Effects	**PH 1st order**	**PH 2nd order**	**Neural Network H = 2**	**Neural Network H = 3**
Stage	0.300	0.325	0.451	0.450
Rx*	-0.130	-0.323	-0.198	-0.260
Age	0.323	0.248	0.219	0.278
Weight	-0.249	0.315	-0.302	-0.581
Stage x Rx	0	-0.032	-0.404	-0.655
Stage x Age	0	0.238	-0.330	-0.415
State x Wt*	0	-0.025	-0.032	-0.109
Rx x Age	0	0.256	0.513	0.484
Rx x Wt	0	-0.026	-0.025	0.051
Stage x Rx x Age	0	0.213	0.360	0.475
Stage x Rx x Wt	0	-0.024	0.026	0.345
Stage x Age x Wt	0	0.006	-0.024	0.271
Rx x Age x Wt	0	-0.363	0.006	-0.363
State x Ex x Age x Wt	0	0.069	0.028	-0.128
		0.293		
		-0.195		
		0		
		0		
		0		
		0		
		0		

* Rx = Treatment
Wt – Weight

Implementation of the proportional odds and proportional hazard were implemented also by Liestol et al (1994) and Biganzoli at al (1998).

Liestol, Anderson (1994) used a neural network for Cox's model with covariates in the form.

Let T be a random survival time variable, and I_k the interval $t_{k-1} < t < t_k$, k = 1, …, K where $0 < t_0 < t_1 < \ldots < t_k < \infty$.

The model can be specified by the conditional probabilities.

$$P\left(T \in I_k \middle| T > t_{k-1}, x\right) = \frac{1}{1 + \exp\left(- \beta_{0k} - \sum_{i=1} \beta_{ik}\ x_i\right)} \tag{1.10}$$

for K = 1, …, K.

The corresponding neural network is the multinomial logistic network with k outputs.

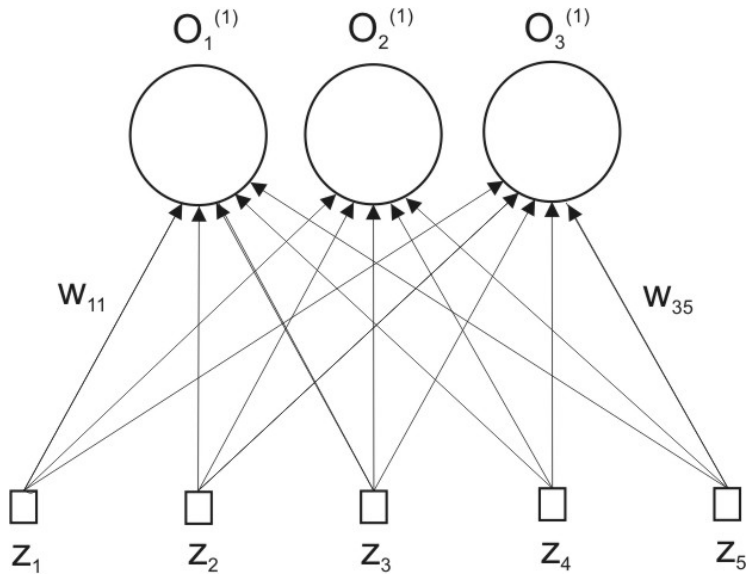

Figure 3: Log odds survival network

The output 0_k in the k^{th} output neuron corresponds to the conditional probability of dying in the interval I_k.

Data for the individual *n* consist of the regressor x^n and the vector $\left(y_1^n, \cdots, y^n\right)$ where y_k^n is the indicator of individual *n*, y in I_k and $k_n \leq K$ is the number of intervals where *n* is observed. Thus $y_1^n, \cdots, y^n\ k_{n-1}$ are all 0 and $y^n\ k_n = 1$ of *n* dies in I_{kn} and

$$0_k = f\left(x, w\right) = \Lambda\left(\beta_{0k} + \sum_{i-1}^{1} \beta_{ik}\ x_i\right) \tag{1.11}$$

and the function to optimize

$$E^*\left(w\right) = \sum_{h-1}^{N} \sum_{k=1}^{K_n} - \log\left(1 - \left|y_k^n - f\left(x^n, w\right)\right|\right) \tag{1.12}$$

and $w = \left(\beta_{01}, \cdots, \beta_{0k}, \cdots, \beta_{Ik}\right)$ and under the hypothesis of proportional rates make the restriction $\beta_{1j} = \beta_2 = \beta_{3j} = \beta_{4j} = \cdots = \beta_j$. Other implementations can be seen in Biganzoli et al (1998).

An immediate generalization would be to substitute the linearity for non-linearity on the regressors adding a hidden layer as in the figure 4.

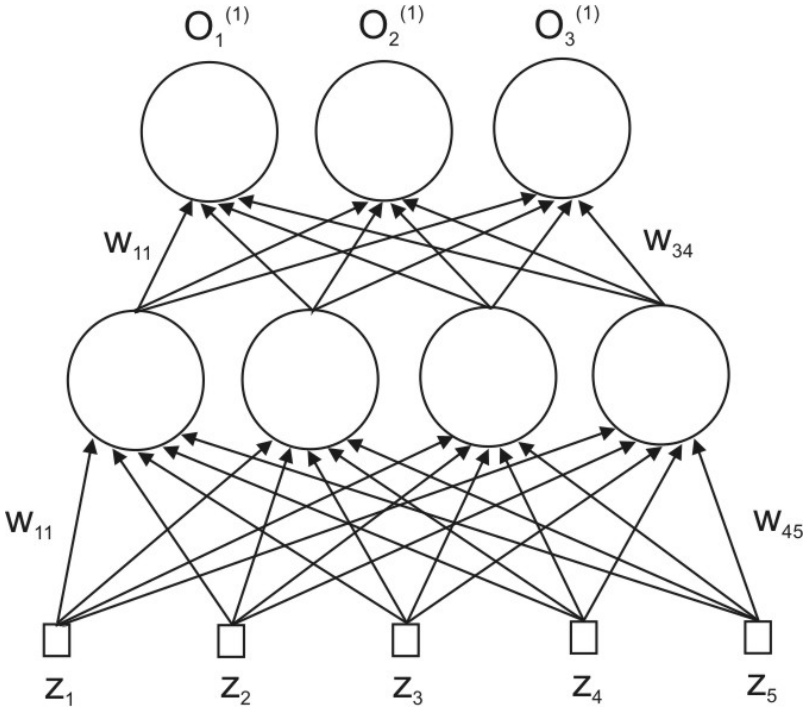

Figure 4– Non-linear survival network (Two-layer feed-forward neural nets, showing the notation for nodes and weights: input nodes (•), output (and hidden) nodes (0), covariates Z_j, connection wrights w_{ij}, output values $0_1^{(2)}$

.

An example from Liestol et al (1994) used the data from 205 patients with melanoma of which 53 died, 8 covariates were included).

Several networks were studied and the negative of the likelihood (interpreted as prediction error) is given in the table 4 below:

Table 4 – Cross validation of models for survival with malignant melanoma. Column 1. Linear model; 2. Linear model with weight-decay; 3. Linear model with a penalty term for non-proportional hazards; 4. Non-linear model with proportional hazards; 5. Non-linear model with a penalty term for non-proportional hazards; 6. Non-linear model with proportional hazards in first and second interval and in third and fourth intervalo; 7. Non-linear model with non-proportional hazards

	1	**2**	**3**	**4**	**5**	**6**	**7**
Prediction error	17	170.	168.	18	16	16	1
Change	2.6	7 - 1.9	6 - 4.0	1.3 8.7	7.0 - 5.6	8.0 - 4.6	70.2 - 2.4

The main results for non-linear models with two hidden nodes were:

- Proportional hazard models produced inferior predictions, decreasing the test log-likelihood of a two hidden node model by 8.7 (column 4) when using the standard weight decay, even more if no weight decay was used.
- Again the test log-likelihood was obtained by using moderately non-proportional models. Adding a penalty term to the likelihood of a non-proportional model or assuming proportionality over the two first and last time intervals improved the test log-likelihood by similar amounts (5.6 in the former case (column 5) and 4.6 in the latter (column 6)). Using no restrictions on the weights except weight decay gave slightly inferior results (column 7, improvement 2.4).

In summary, for this small data set the improvements that could be obtained compared to the simple linear models were moderate. Most of the gain could be obtained by adding suitable penalty terms to the likelihood of a linear but non-proportional model.

An example from Biganzoli et al (1998) is the application neural networks in the data sets of Efron´s brain and neck cancer and Kalbfleish and Prentice lung cancer using the network architecture of figure 5. The results of survival curve fits follows in figures 6 and 7.

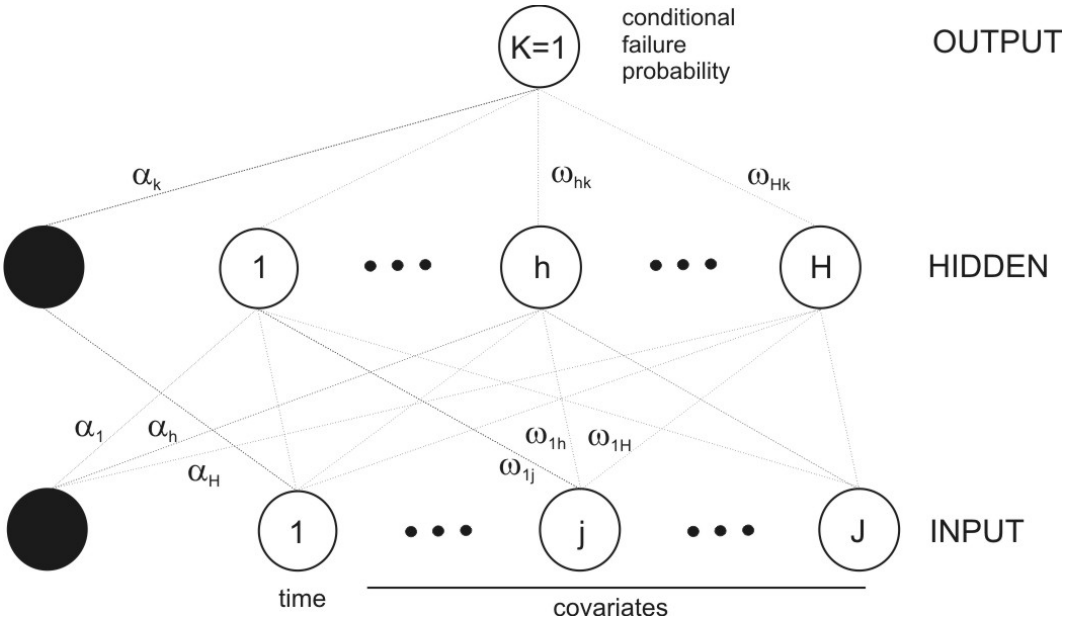

Figure 5 – Feed forward neural network model for partial logistic regression (PLANN). The units (nodes) are represented by circles and the connections between units are represented by dashed lines. The input layer has *J* units for time α and covariates plus one … unit (0). The hidden layer has *H* units plus the … unit (0). A single output unit (K = 1) compute conditional failure probability x_1 and x_2 are the weights for the connections of the … unit with the hidden and output unit w_a and $w_{...}$ are the weights for the connections between input and hidden units and hidden and output unit, respectively.

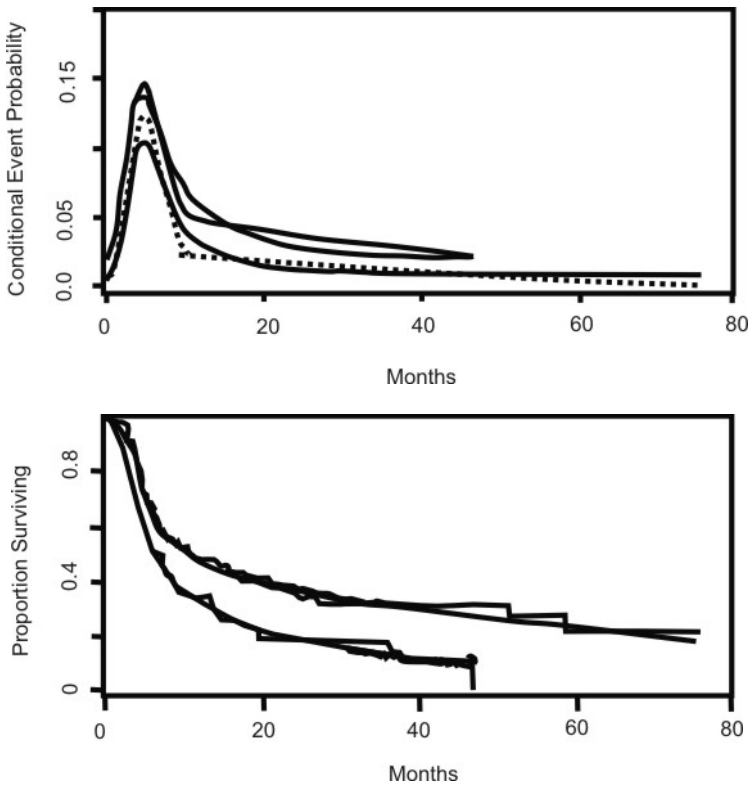

Figure 6 – Head and neck cancer trial((a) estimates of conditional failure probability obtained with a sub optional PLANN model (solid line) and the cubic-linear spline proposed by Efron[13] (dashed lines); (b) corresponding survival function and Kaplan-Meyer estimates

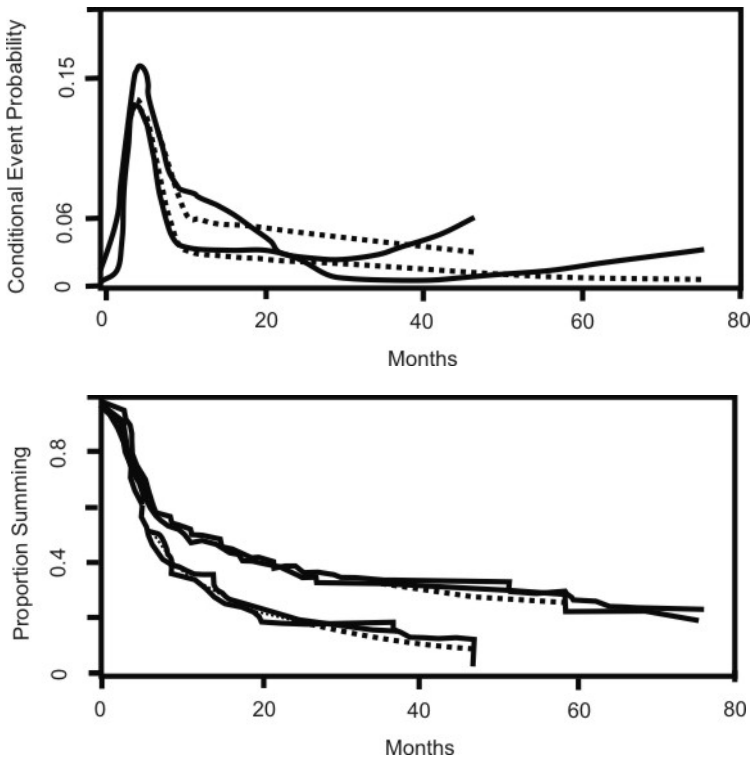

Figure 7 – Head and neck cancer trial ((a) estimates of conditional failure probability obtained with the best PLANN configuration (solid line) and the cubic-linear spline proposed by Efron[13] (dashed lines); (b) corresponding survival function and Kaplan-Meyer estimates)

A further reference is Bakker et al (2005) who have used a neural-Bayesian approach to fit Cox survival model using MCMC and an exponential activation function. Other applications can be seen in Lapuerta et al (1995), Ohno-Machado et al (1995) and Mariani at al (1997).

References

[1] Bakker, B., Heskes, T., Neijt, J., Kappen, B. (2003). *Improving Lox Survival Analysis with a Neural-Bayesian Approach Statistics in Medicine* (in press).
[2] Biganzoli, E., Baracchi, P., Marubini, E. (2002). A general framework for neural network models on censored survival data. *Neural Networks*, 15, 209-218.
[3] Biganzoli, E., Baracchi, P., Mariani, L., Marubini, E. (1998). Fee-forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17, 1169-1186.
[4] Ciampi A and Etezali_Amoli,J - (1985) A general model for testing the proportional hazard and the accelerated failure time hypothesis in the analysis of censored survival data with covariates. Communications in Statistics A 14:651-667.
[5] Faraggi, D., Simon, R. (1995b). The maximum likelihood neural network as a statistical classification model. *J. of Statistical Planning and Inference*, 46, 93-104.
[6] Faraggi, D., LeBlanc, M., Crowley, J. (2001). Understanding neural networks using regression trees: An application to multiply myeloma survival data. *Statistics in Medicine*, 20, 2965-2976.
[7] Faraggi, D., Simon, R. (1995a). A neural network model for survival data. *Statistics in Medicine*, 14, 73-82.
[8] Faraggi, D., Simon, R., Yaskil, E., Kramar, A. (1997). Bayesian neural network models for censored data. *Biometrical Journal*, 39, 519-532.
[9] Groves, D. J., Smye, S. W., Kinsey, S. E., Richards, S. M., Chessells, J. M., Eden, O. B., Basley, C. C. (1999). A comparison of Cox regression and neural networks for tisk stratification in cases of acute lymphoblastic leukemia in children. *Neural Computing and Applications*, 8, 257-264.
[10] Kalbfleish, J. D., Prentice, R. L. (2002). *The Statistical Analysis of Failure Data*. Wiley, 2nd Ed.
[11] Lapuerta ,P. Azen,S.P and LaBree,L. (1995)Use of neural networks in predicting the risk of coronary arthery disease. Computers and Biomedical Research,28,38-52. Liestol, K., Anderson, P. K., Anderson, U. (1994). Survival analysis and neural nets. *Statistical in Medicine*, 13, 1189-1200.
[12] Louzada-Neto, F. (1997). Extended hazard regression moel for rehability and survival analysis. *Lifetime Data Analysis*, 3, 367-381.
[13] Louzada-Neto, F. (1999). Modeling life data: A graphical approach. *Applied Stochastic Models in Business and Industry*, 15, 123-129.

[14] Louzada-Neto., F., Pereira, B. B. (2000). Modelos em análise de sobrevivência. *Cadernos de Saúde Coletiva*, 8(1) 9-26.
[15] Mariani,L. Coradin,D. Biganzoli,E. Boracchi, P.Marubini,E.Pilloti, S. Salvatori,R.Silvestrini,R, Veronesi,V. Zucali,R Rilke,F.(1997)Prognostic factors for ,etachronous contralateral breast cancer :a comparisomn of the linear Cox regression model and its artificial neural networks extension. Breast Cancer Research and Treatment 44, 167-178. Ohno-Machado, L. (1997). A comparison of Cox proportional hazard and artificial neural network models for medical prognoses. *Computers in Biology and Medicine*, 27, 55-65.
[16] Ohno-Machado,L. Walker, M.G.and Musen, M.A.(1995) Hierarchical neural networks for survival analysis. Stanford Medical Information Report 94-0542 In The Eight Word Congress on Medical Information Vancouver BC Canada
[17] Ripley, B. D., Ripley, R. M. (1998). Neural networks as statistical methods in survival analysis. In *Artificial Neural Networks: Prospects for Medicine*. Eds. R. Dybowski and C. Gant, Landes Biosciences Publishing.
[18] Ripley, R. M. (1998). Neural network models for breast cancer prognoses. *D. Phill. Thesis*. Department of Engineering Sciences, University of Oxford (www.stats.ox.ac.uk/ruth/index.html).
[19] Xiang, A., Lapuerta, P., Ryutov, ..., Buckley, J., Azen, S. (2000). Comparison of the performance of neural networks methods and Cox regression for censored survival data. *Computational Statistics and Data Analysis*, 34, 243-257.

Basilio de Braganca Pereira (on the right) is Full Professor at the Faculty of Medicine (since 1998) and also at the Engineering Graduate Program (COPPE), since 1970, both at the Federal University of Rio de Janeiro UFRJ. He has a B.Sc in Statistics (Escola Nacional de Ciências Estatísticas (1965-1968) and M.Sc.and PhD. degrees in Statistics (Imperial College Of Science Technology And Medicine, University of London (1972-1976)),and M.Sc. (1970) and L.D. (1989) in Operational Research (COPPE/UFRJ), was a Research Visitor at the Penn State University (2003-2004). Prof. Pereira is the coordinator of the Statistical research consulting group at The University Hospital of UFRJ. His main interests are Statistical Methods and their applications in Medicine and Engineering. He and Professor C.R. Rao are co authors of the book "Data Mining with Neural Networks: A Guide for Statisticians", in press.

Calyampudi R. Rao (on the left) is Emeritus Holder of the Eberly Family Chair in Statistics and director of the Center for Multivariate Analysis. He has received thirty honorary doctoral degrees from universities in seventeen countries on six continents.

One of the world's top five statisticians, Rao is recognized internationally as a pioneer who laid the foundation of modern statistics. Technical terms bearing his name appear in all standard textbooks on statistics, including such terms as the Cramer-Rao Inequality, Rao-Blackwellization, Fisher-Rao Theorem, Rao Distance, and Rao's Score test.

In 2002 Rao was honored by President George W. Bush with the National Medal of Science, the highest award given to an American scientist for lifetime achievement in fields of scientific research. He has been honored by the government of India with the Padma Vibhushan award in 2001—the country's second-highest civilian honor—for outstanding contributions to science, engineering, and statistics.

Rao earned his Ph.D. and Sc.D. degrees in 1948 at Cambridge University in England. He came to the United States 1978 after serving as director of the Indian Statistical Institute, where he had held various research and administrative positions since 1944. In 1982 he established the Center for Multivariate Analysis at the University of Pittsburgh, where he continues as adjunct professor. Rao joined the Penn State faculty in 1988. He has authored or co-authored 14 books—some of which have been translated into several languages—and more than 300 research papers published in scientific journals. He has supervised the doctoral research of approximately 50 students who have trained another 250 doctoral students themselves.