STATISTICS IN MEDICINE
Statist. Med. **18**, 1119–1131 (1999)

MIXTURES OF PROPORTIONAL HAZARDS REGRESSION MODELS

ORI ROSEN[1] AND MARTIN TANNER[2]*

[1]*Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208, U.S.A.*
[2]*Department of Statistics, University of Pittsburgh, 139 University Place, Pittsburgh, PA 15260, U.S.A.*

SUMMARY

This paper presents a mixture model which combines features of the usual Cox proportional hazards model with those of a class of models, known as mixtures-of-experts. The resulting model is more flexible than the usual Cox model in the sense that the log hazard ratio is allowed to vary non-linearly as a function of the covariates. Thus it provides a flexible approach to both modelling survival data and model checking. The method is illustrated with simulated data, as well as with multiple myeloma data. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

This paper describes a generalization of the Cox proportional hazards model, utilizing ideas from a class of models, known as mixtures-of-experts (ME). ME, originally introduced by Jacobs *et al.*[1] assume that the process generating the data can be decomposed into a set of subprocesses defined on possibly overlapping regions of the covariate space. As an illustration, we consider the ME generalization of the usual linear regression model, $y_s = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_s + \varepsilon_s$, where $\varepsilon_s \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \sigma^2)$, $s = 1, \ldots, n$. The mixtures-of-experts model assumes that y_s is generated as follows. For each covariate $\boldsymbol{x}_s$

(i) a label r is chosen from a multinomial distribution with probability $g_r^{(s)} = P(r \,|\, \boldsymbol{x}_s, V)$, where $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{I-1}]$ is the $d \times (I-1)$ matrix of parameters underlying the multinomial distribution and d is the dimension of $\boldsymbol{x}_s$. The multinomial probabilities depend on $\boldsymbol{x}_s$, $s = 1, \ldots, n$, according to the generalized linear model

$$\log \frac{g_r^{(s)}}{g_I^{(s)}} = \boldsymbol{v}_r^{\mathrm{T}}\boldsymbol{x}_s, \quad r = 1, \ldots, I-1$$

where $\sum_{r=1}^{I} g_r^{(s)} = 1$;

* Correspondence to: Martin Tanner, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208, U.S.A.

Contract/grant sponsor: NIH
Contract/grant number: CA35464

Contract/grant sponsor: Mellon Foundation

CCC 0277–6715/99/091119–13$17.50
Copyright © 1999 John Wiley & Sons, Ltd.
Received February 1998
Accepted August 1998

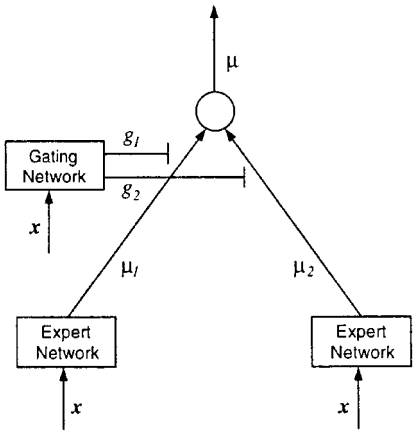

Figure 1. A mixtures-of-experts model (from Peng *et al.*[2])

(ii) for the label, r, a response $y^{(s)}$ is generated with probability

$$P(y_s \mid r, \boldsymbol{x}_s, \boldsymbol{\beta}_r, \sigma_r) = (2\pi\sigma_r^2)^{-1/2} \mathrm{e}^{-(y_s - \boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_s)^2/2\sigma_r^2}.$$

In the above illustration, the expected conditional value of the response y_s, denoted by $\mu_r^{(s)}$, is given by $\boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_s$. The total probability of generating y_s from $\boldsymbol{x}_s$ is given by the mixture density

$$P(y_s \mid \boldsymbol{x}^{(s)}, \Theta) = \sum_{r=1}^{I} P(r \mid \boldsymbol{x}_s, V) P(y_s \mid r, \boldsymbol{x}_s, \boldsymbol{\beta}_r, \sigma_r)$$

where Θ denotes all the parameters: $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{I-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_I, \sigma_1, \ldots, \sigma_I$. Assuming independently distributed data, the total probability of the data set $\chi = \{(\boldsymbol{x}_s, y_s)\}_{s=1}^n$ is the product of n such densities with the likelihood given by

$$L(\Theta \mid \chi) = \prod_{s=1}^{n} \sum_{r=1}^{I} g_r^{(s)} \frac{1}{\sqrt{(2\pi\sigma_r^2)}} \exp\left\{ -\frac{1}{2\sigma_r^2} (y_s - \boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_s)^2 \right\}.$$

Thus, the ME model can be viewed as a mixture model where the mixing weights and mixed distributions depend on the covariate $\boldsymbol{x}$.

ME can also be viewed as a tree-structured model. Figure 1 presents a graphical representation of ME in the case $I = 2$. The model consists of I modules, each referred to as an *expert*. These experts approximate the data within each region of the covariate space; expert r maps its input $\boldsymbol{x}$ to an output μ_r. It is assumed that different experts are appropriate in different regions of the covariate space. Thus, the model attempts to solve problems using a 'divide-and-conquer' strategy; complex tasks are decomposed into simpler subtasks. Consequently the model requires a module, referred to as a *gating network*, that identifies for an input $\boldsymbol{x}_s$ the expert or blend of experts whose output is most likely to approximate the corresponding response y_s. The gating network outputs are the probabilities g_r that weight the contributions of the various experts. The total output of the model is $\mu = \sum_{r=1}^{I} g_r \mu_r$. In general, it is assumed that the conditional distribution of y_s given $\boldsymbol{x}_s$ and the label r, is a member of the exponential family. Thus, in general,

Copyright © 1999 John Wiley & Sons, Ltd.

ME provide a richer class of models than ordinary generalized linear models. In the linear regression case, in particular, they can be viewed as an alternative to non-parametric regression.

Jordan and Jacobs[3] present an extension of ME, called hierarchical mixtures-of-experts (HME). Peng *et al.*[2] discuss full Bayesian inference in ME and HME with an application to speech recognition. A detailed description of Bayesian inference in ME and HME with applications to regression and classification is given in Jacobs *et al.*[4] A Bayesian approach to model selection in ME and HME is proposed by Jacobs *et al.*[5]

The paper by Kooperberg *et al.*[6] presents an approach to hazard regression modelling using splines. Other spline approaches to the non-parametric modelling of survival time on covariates include the work by Hastie and Tibshirani,[7] O'Sullivan,[8] Sleeper and Harrington[9] and Gray.[10] LeBlanc and Crowley[11] develop an approach motivated by the CART methodology of Breiman *et al.*[12]

The outline of the paper is as follows. In Section 2 we present our method, preceding it with necessary background on the Cox model. The proposed method is based on a mixture embedded within the usual Cox partial likelihood, resulting in a more flexible model. A method for selecting the number of components in this mixture is also described. Section 3 illustrates the method with simulated, as well as real data. Section 4 concludes with a discussion.

2. A MIXTURE OF COX EXPERTS

2.1. The model

Prior to presenting our method, we describe the familiar set-up of the Cox model, which will be used in the sequel. The proportional hazards model[13] assumes that

$$\lambda(t; x_1, \ldots, x_p) = \lambda_0(t) \exp\left(\sum_{i=1}^{p} \beta_i x_i\right) \tag{1}$$

where $\lambda(t; x_1, \ldots, x_p)$ is the hazard at time t, given covariate values $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$, and λ_0 is an unspecified baseline hazard function. The available data are of the form $(y_1, \boldsymbol{x}_1, \delta_1), \ldots, (y_n, \boldsymbol{x}_n, \delta_n)$, where the survival time $y_i, i = 1, \ldots, n$, is complete if $\delta_i = 1$ and censored if $\delta_i = 0$. The distinct failure times are denoted by $t_1 < t_2 < \cdots < t_k$ with d_i failures at $t_i, i = 1, \ldots, k$. The parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ are estimated without specification of $\lambda_0(t)$ by maximizing the partial likelihood

$$\prod_{i=1}^{k} \frac{\exp(\sum_{j \in \mathscr{D}(t_i)} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_j)}{(\sum_{l \in \mathscr{R}(t_i)} \mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_l})^{d_i}}. \tag{2}$$

In (2), $\mathscr{D}(t_i)$ is the set of indices of the failures at t_i, and $\mathscr{R}(t_i)$ is the set of indices of the subjects at risk at time t_i^-. This approximation for ties is due to Peto[14] and Breslow.[15] For a fuller discussion of ties, see Kalbfleisch and Prentice.[16] The construction of the partial likelihood can be justified by arguments given in Kalbfleisch and Prentice.[16]

One of the goals of fitting a survival model is to estimate the survivor function of a subject with covariate values $\boldsymbol{x}_0$. Let $\Lambda_0(t) = \int_0^t \lambda_0(u)\, \mathrm{d}u$ be the cummulative baseline hazard. Based on the

Copyright © 1999 John Wiley & Sons, Ltd.

maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$, Breslow[17] estimates $\Lambda_0(t)$, for t in $[t_i, t_{i+1})$, $i = 1, \ldots, k-1$, by

$$\hat{\Lambda}_0(t) = \sum_{j=1}^{i} \frac{d_j}{\sum_{l \in \mathscr{R}(t_j)} \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_l}}. \tag{3}$$

Using (3), the survivor function $S(t; \boldsymbol{x}_0)$ of a subject with covariate values $\boldsymbol{x}_0$ is then estimated by

$$\hat{S}(t; \boldsymbol{x}_0) = \exp(-\hat{\Lambda}_0(t) \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_0}). \tag{4}$$

Our approach to the construction of a partial likelihood for the ME generalization is to adapt a derivation of the original proportional hazards model, presented in Miller.[18] Given the I experts, the probability of a failure in $[t_i, t_i + \Delta]$, for small Δ, given $\mathscr{R}(t_i)$ is approximately

$$\sum_{l \in \mathscr{R}(t_i)} \sum_{r=1}^{I} g_r^{(l)} \mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}} \boldsymbol{x}_l} \lambda_0(t_i) \Delta \tag{5}$$

where $g_r^{(l)}$ depends on $\boldsymbol{x}_l$ according to the generalized linear model

$$\log \frac{g_r^{(l)}}{g_I^{(l)}} = \boldsymbol{v}_r^{\mathrm{T}} \boldsymbol{x}_l, \quad r = 1, \ldots, I-1 \tag{6}$$

and $\sum_{r=1}^{I} g_r^{(l)} = 1$. It then follows that the probability of a failure of a subject with covariates $\boldsymbol{x}_j$ at time t_i, given one failure in $\mathscr{R}(t_i)$ at time t_i is

$$\sum_{r=1}^{I} g_r^{(j)} \mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}} \boldsymbol{x}_j} \Bigg/ \sum_{l \in \mathscr{R}(t_i)} \sum_{r=1}^{I} g_r^{(l)} \mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}} \boldsymbol{x}_l}.$$

Taking the product of these conditional probabilities (with an adjustment for ties) gives the partial likelihood for the mixture of experts

$$\mathrm{PL}_g = \prod_{i=1}^{k} \prod_{j \in \mathscr{D}(t_i)} \frac{\sum_{r=1}^{I} g_r^{(j)} \mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}} \boldsymbol{x}_j}}{\sum_{l \in \mathscr{R}(t_i)} \sum_{r=1}^{I} g_r^{(l)} \mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}} \boldsymbol{x}_l}}. \tag{7}$$

As described in Section 1, the idea underlying ME is the 'divide-and-conquer' principle, according to which different experts are appropriate in different regions of the covariate space. In the application of ME to proportional hazards regression, this idea is reflected in equation (5). In this equation the probability of a failure is written as a mixture model, where the mixing weights $g_r^{(l)}$ depend on the covariate vector $\boldsymbol{x}_l$. Analogous to the tree structure in Figure 1, each expert provides an assessment of the conditional probability of a failure in the infinitesimal interval $[t_i, t_i + \Delta]$. These assessments are then merged by the gating network via the weighting scheme given in (6). Note that when $I = 1$, (7) reduces to the ordinary partial likelihood (2).

2.2. Inferential issues

The parameters of the mixture model are estimated by maximizing $l = \log \mathrm{PL}_g$. Approximate standard errors can be obtained utilizing the Hessian of l. The first and second derivatives of l are given in the Appendix.

Next, we use the Hessian to derive approximate pointwise confidence intervals for $\eta^{(s)} = \sum_{r=1}^{I} g_r^{(s)} \boldsymbol{\beta}_r^{\mathrm{T}} \boldsymbol{x}_s$. A graph of $\hat{\eta}^{(s)} = \sum_{r=1}^{I} \hat{g}_r^{(s)} \hat{\boldsymbol{\beta}}_r^{\mathrm{T}} \boldsymbol{x}_s$ versus x_s^i, the ith component of $\boldsymbol{x}_s$, fixing all the other components of $\boldsymbol{x}_s$, provides a useful diagnostic for the proportional hazards model.

Copyright © 1999 John Wiley & Sons, Ltd.

Non-linearity in this plot suggests a lack of fit of the proportional hazards model. The pointwise standard error of $\hat{\eta}^{(s)}$ is obtained via the delta method. Specifically, the gradient of $\eta^{(s)}$ is given by

$$\nabla^{\mathrm{T}}\eta^{(s)} = \left(\frac{\partial\eta^{(s)}}{\partial\boldsymbol{v}_1^{\mathrm{T}}}, \ldots, \frac{\partial\eta^{(s)}}{\partial\boldsymbol{v}_{I-1}^{\mathrm{T}}}, \frac{\partial\eta^{(s)}}{\partial\boldsymbol{\beta}_1^{\mathrm{T}}}, \ldots, \frac{\partial\eta^{(s)}}{\partial\boldsymbol{\beta}_I^{\mathrm{T}}}\right)$$

where $\partial\eta^{(s)}/\partial\boldsymbol{v}_l = g_l^{(s)}[(1-g_l^{(s)})\boldsymbol{\beta}_l^{\mathrm{T}}\boldsymbol{x}_s - \sum_{r\neq l} g_r^{(s)}\boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_s]\boldsymbol{x}_s'$, $l = 1, \ldots, I-1$, $\partial\eta^{(s)}/\partial\boldsymbol{\beta}_m = g_m^{(s)}\boldsymbol{x}_s$, $m = 1, \ldots, I$, and $\boldsymbol{x}_s' = (1, \boldsymbol{x}_s^{\mathrm{T}})^{\mathrm{T}}$. It then follows that approximate $(1-\alpha)$ level pointwise confidence intervals for $\eta^{(s)}$ are formed by

$$\hat{\eta}^{(s)} \pm z_{1-\alpha/2}(\nabla^{\mathrm{T}}\hat{\eta}^{(s)}(-\hat{H}_l)^{-1}\nabla\hat{\eta}^{(s)})^{1/2} \tag{8}$$

where H_l is the Hessian matrix of l, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution.

For estimating the baseline cumulative hazard and the survivor, we use weighted versions of formulae (3) and (4):

$$\hat{\Lambda}_0(t) = \sum_{j=1}^{i} \frac{d_j}{\sum_{l\in\mathcal{R}(t_j)}\sum_{r=1}^{I}\hat{g}_r^{(l)}\mathrm{e}^{\hat{\boldsymbol{\beta}}_r^{\mathrm{T}}\boldsymbol{x}_l}}.$$

$$\hat{S}(t;\boldsymbol{x}_0) = \exp(-\hat{\Lambda}_0(t)\sum_{r=1}^{I}\hat{g}_r^{(0)}\mathrm{e}^{\hat{\boldsymbol{\beta}}_r^{\mathrm{T}}\boldsymbol{x}_0}). \tag{9}$$

For the examples in Section 3, the function `nlminb` of S-PLUS was used to maximize l. In order to avoid local maxima for the examples in this paper, the maximization process was repeated 20 times with random starting values. Thus, the presented estimates represent the maximizer over the 20 maximizations. The use of multiple starting points is quite standard in such problems and not terribly onerous. For the examples in this paper, the algorithm converged fairly quickly, and, for the most part, the global maximum was not hard to find. For example, in the first example of Section 3, 14 out of the 20 maximizations resulted in the global maximum. However, as noted by the referee, multiple modes can be more of an issue as I is made larger. All computations were performed on a Sun Ultra 1 workstation.

2.3. Selecting the number of experts

An important issue in the specification of the mixture of proportional hazards models is the selection of the number of experts. In the context of mixtures of generalized linear models, this issue is addressed by Jacobs *et al.*[5] In the context of mixtures of marginal models, this issue is addressed by Rosen *et al.*[19] To develop an approach to selecting the number of experts in the mixture of proportional hazards models context, we again draw upon the motivating logic in Miller.[18]

According to expert r, the (unconditional) probability of a failure in the interval $[t_i, t_i+\Delta]$ is given by $g_r^{(l)}\mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_l}\lambda_0(t_i)\Delta$. The total (unconditional) probability of a failure in this infinitesimal interval according to all I experts is given by $\sum_{j=1}^{I} g_j^{(l)}\mathrm{e}^{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_l}\lambda_0(t_i)\Delta$. Thus, to motivate a measure of the value of expert r, we consider the ratio

$$\frac{g_r^{(l)}\mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_l}\lambda_0(t_i)\Delta}{\sum_{j=1}^{I} g_j^{(l)}\mathrm{e}^{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_l}\lambda_0(t_i)\Delta} = \frac{g_r^{(l)}\mathrm{e}^{\boldsymbol{\beta}_r^{\mathrm{T}}\boldsymbol{x}_l}}{\sum_{j=1}^{I} g_j^{(l)}\mathrm{e}^{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_l}} = h_r^{(l)}.$$

Copyright © 1999 John Wiley & Sons, Ltd.

Clearly, if expert r plays little or no role in modelling the response surface in the region of $\boldsymbol{x}_l$, then $h_r^{(l)}$ will be negligible. However, if expert r takes the lead role in modelling the response surface in the region of $\boldsymbol{x}_l$, then $h_r^{(l)}$ may approach unity. To examine the overall value or worth of expert r, we then average these $h_r^{(l)}$'s over all the distinct failure times, that is, we consider the worth index (w_r) of expert r as $(1/k)\sum_{l=1}^{k} h_r^{(l)}$. If the worth indices for each of the experts are all of similar magnitudes, then this suggests that the current model may be too small, and that a model with additional experts should be considered. Alternatively, if the worth index for an expert is small relative to that of other experts, then there is evidence to suggest that this expert can be pruned from the model. A useful more formal criterion suggested by Jacobs *et al.*[5] to determine the number of experts is the minimum number of experts with the largest worth indices for which the sum of their worth indices exceeds some critical value κ, that is

$$\min\left\{J : \sum_{r=1}^{J} w_{(r)} \geqslant \kappa\right\} \tag{10}$$

where $w_{(1)} \geqslant w_{(2)}$. … All other experts are pruned from the model. Jacobs *et al.*[5] take $\kappa = 0{\cdot}8$. Although the 0·8 cut-off criterion is arbitrary, based on empirical work, Jacobs *et al.*[5] have found this value to work well in practice. For example, suppose that the worth indices for the four experts that comprise a model have the values 0·4, 0·3, 0·2 and 0·1. According to (10), this suggests that three experts are required ($0{\cdot}4 + 0{\cdot}3 + 0{\cdot}2 = 0{\cdot}9$), and one expert can be pruned. Care, however, has to be taken; if an expert is used for a specific type of data which occurs relatively infrequently in the data set, then dropping the expert may not be optimal. If the expert worth indices are similar in magnitude, for example, they are all within 10 per cent of $1/I$, where I is the number of experts, then one may wish to add experts to the model.

3. EXAMPLES

We illustrate the mixture model with a simulated data set, as well as using multiple myeloma data found in Krall *et al.*[20] and analysed by Collett.[21]

3.1. Simulated data

The data considered here were simulated from the Weibull model

$$\log T_i = |x_i| + 0{\cdot}2\varepsilon_i, \quad i = 1, \ldots, 200 \tag{11}$$

where T_i is the survival time of the ith subject, x_i is the corresponding covariate, taking equally spaced values on a grid between -1 and 1, and the ε_i's are i.i.d. Gumbel variates. The censoring times were simulated from an exponential random variable, resulting in 10 per cent non-informative censoring. Model (11) implies that the survival times follow a Weibull distribution with hazard function

$$\lambda(t; x) = 5t^4 \, \mathrm{e}^{-5|x|}. \tag{12}$$

Note that this model can also be written as

$$\lambda(t; x) = 5t^4(g_1 \mathrm{e}^{5x} + g_2 \mathrm{e}^{-5x})$$

Copyright © 1999 John Wiley & Sons, Ltd.

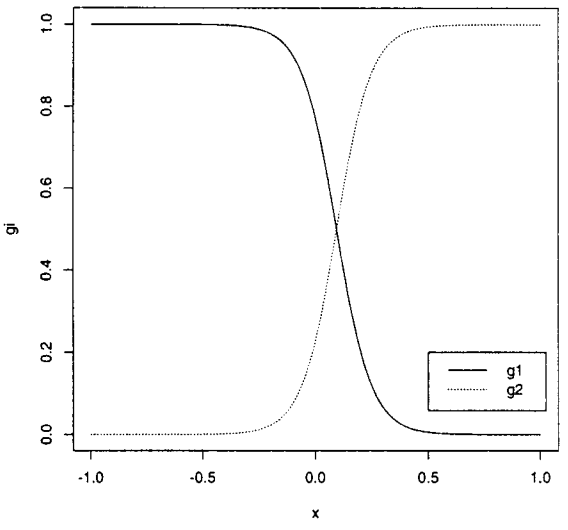

Figure 2. Mixing probabilities versus covariate values

where

$$g_1 = \begin{cases} 1 & x < 0 \\ 0 & x \geqslant 0 \end{cases}$$

and $g_2 = 1 - g_1$. Note also that the true log hazard ratio for this model, is $-5|x|$. Equations (1) and (12) imply that for a given x, the underlying true β satisfies $\beta = -5\ \mathrm{sgn}(x)$.

Model (7) with two experts ($I = 2$) was fitted to these data, resulting in the estimates $\hat{\boldsymbol{\nu}}_1^{\mathrm{T}} = (\hat{v}_{10}, \hat{v}_{11}) = (1{\cdot}22, -12{\cdot}86)$, $\hat{\beta}_1 = 4{\cdot}78$ and $\hat{\beta}_2 = -5{\cdot}12$, with standard errors $\mathrm{SE}(\hat{v}_{10}) = 0{\cdot}79$, $\mathrm{SE}(\hat{v}_{11}) = 2{\cdot}47$, $\mathrm{SE}(\hat{\beta}_1) = 0{\cdot}48$ and $\mathrm{SE}(\hat{\beta}_2) = 0{\cdot}69$.

Figure 2 shows how the covariate space was segmented between the two experts. In particular, the horizontal axis gives the covariate values and the vertical axis gives the values of the mixing probabilities $\hat{g}_1^{(s)} = \exp(1{\cdot}22 - 12{\cdot}86x_s)/(1 + \exp(1{\cdot}22 - 12{\cdot}86x_s))$ and $\hat{g}_2^{(s)} = 1 - \hat{g}_1^{(s)}$, $s = 1, \ldots, 200$. While $\hat{g}_1^{(s)}$ is large for $x_s < 0$, $\hat{g}_2^{(s)}$ is large for $x_s \geqslant 0$. The mixing probabilities thus dictate that expert 2 assumes the main responsibility for modelling the response corresponding to $x_s \geqslant 0$, whereas for $x_s < 0$ expert 1 is the primary expert.

To assess whether two experts are sufficient in this case, we used the technique described in Section 2.3. In particular, a model with three experts was fit to the data, resulting in the worth indices 0·20, 0·21 and 0·59. Criterion (10), thereby, suggests that a model with two experts is sufficient. In fact, the fitted log hazard ratio using three experts is quite similar to the fit using two experts, and appealing to Ockham's Razor we prefer the two-expert model.

Figure 3 shows: (a) the true log hazard ratio $-5|x|$; (b) $\hat{\eta} = \sum_{r=1}^{2} \hat{g}_r \hat{\beta}_r x$; (c) $\hat{\beta}^{\mathrm{PH}} x$, based on the proportional hazards model, where $\hat{\beta}^{\mathrm{PH}} = -0{\cdot}0204$; (d) Hastie and Tibshirani's estimate (see details below); and (e) approximate 95 per cent pointwise confidence intervals for η, based on the mixture model. The mixture estimate $\hat{\eta}$ is seen to recover the true log hazard ratio. The

Copyright © 1999 John Wiley & Sons, Ltd.

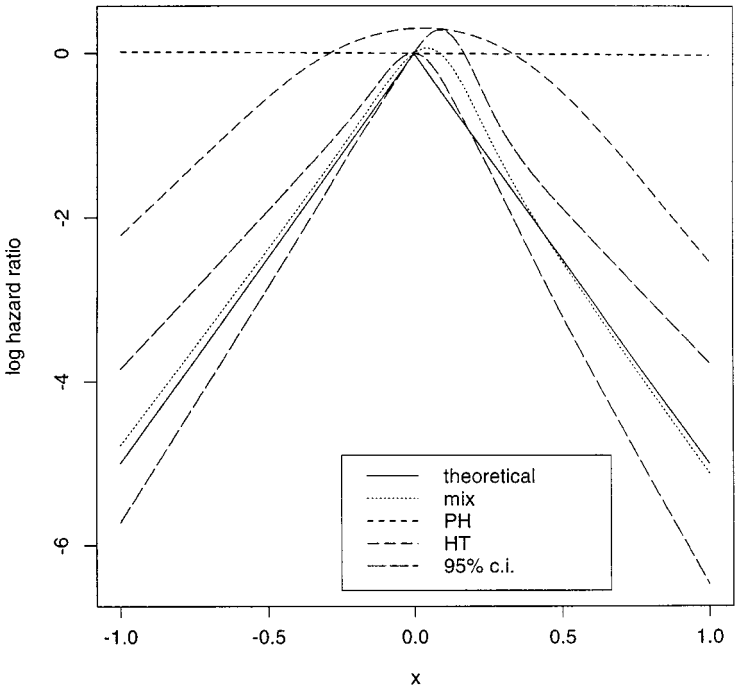

Figure 3. Theoretical log hazard ratio, mixture estimate of log hazard ratio, proportional hazards estimate of log hazard ratio, Hastie and Tibshirani's estimate and approximate 95 per cent pointwise confidence bands for the log hazard ratio

proportional hazards estimate clearly mismodels the data over the whole range of x. Of course, had the data analyst known that the log hazard ratio followed the absolute value of x, then it would have been appropriate to fit this form of the model. We assume here that the true structure is unknown to the data analyst. Moreover, while a careful residual analysis may uncover this structure, the point of this example is to illustrate that the mixture of Cox experts model can automatically detect this non-linear (and non-differentiable) structure. Hastie and Tibshirani's estimate was computed using statlib's S-PLUS function `coxgam()`, which implements the additive proportional hazards model described in Hastie and Tibshirani.[7] As can be seen, this estimate is also biased. Note that the true log hazard ratio is non-differentiable at $x = 0$, whereas the additive proportional hazards model assumes that the log hazard ratio is a sum of smooth functions, each differentiable and having an absolutely continuous first derivative.[7] Non-differentiable response surfaces arise in change-point models, which play an important role in cancer epidemiology.[22] The confidence bands based on (8) contain the log hazard ratio over most of the covariate range.

Figure 4 depicts the theoretical survivor function $S(t; x) = \exp(-t^5 e^{-5|x|})$, computed at $x = 0$, along with estimates based on the proportional hazards model and the mixture model ((4) and (9), respectively). It is evident that the survivor estimate based on the proportional hazards model is severely biased, whereas the bias of the mixture estimate is much smaller.

Copyright © 1999 John Wiley & Sons, Ltd.

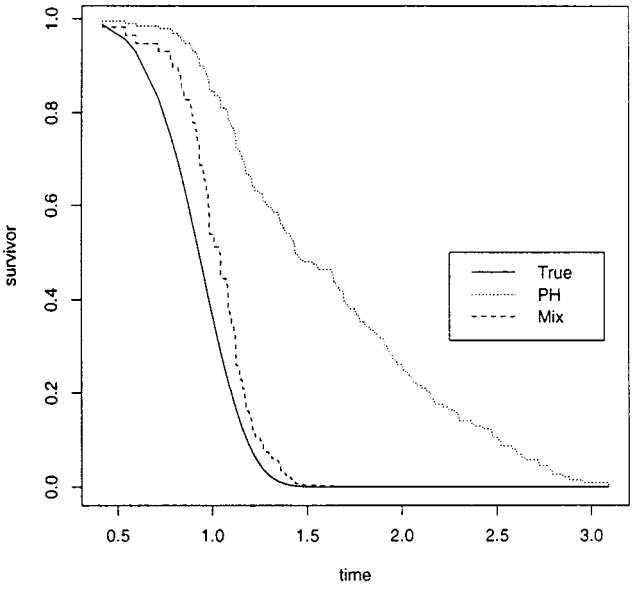

Figure 4. Theoretical survivor function and estimates

3.2. Multiple myeloma data

The multiple myeloma data relate to 48 patients at ages 50–80. Some of these patients were still alive when the study was completed, so their survival times are right-censored. A number of covariates were recorded for each patient. We fit a mixture of two experts ($I = 2$) with the two covariates, serum haemoglobin (H) and blood urea nitrogen (B). Collett[21] found the proportional hazards model containing these two covariates to be the most satisfactory among several models.

Figure 5 shows, on the left-hand side, plots of $\hat{\eta}_H$ against H, where B is fixed at its sample lower quartile, median and upper quartile, resulting in three of the displayed lines. Also shown is a graph of $\hat{\beta}_H^{\mathrm{PH}} H + \hat{\beta}_B^{\mathrm{PH}} B$, based on the proportional hazards model, where B is fixed at its sample upper quartile. On the right, the analogous quantities are plotted against B, where H is fixed at its sample quartiles. The point estimates in Figure 5 suggest that η may not be entirely additive in H and B, and that η may have non-linearities. A definitive conclusion, however, requires the use of confidence bands. Figure 6 shows for each of the two covariates: (a) one of the three lines from Figure 5, based on the mixture model, computed when the other covariate is fixed at its sample upper quartile; (b) approximate 95 per cent level confidence bands; (c) the estimate based on the standard proportional hazards model. The confidence bands on the left-hand side plot show that the bend in $\hat{\eta}$ is probably due to chance variation (note the small number of observations to the left of $H = 7{\cdot}5$). The same is also true for the other two lines (not shown). Our analysis confirms that one expert, that is, the proportional hazards model, yields a satisfactory fit to the data. Thus, the ME approach not only provides a flexible scheme to model survival data, it also assists in model checking.

Copyright © 1999 John Wiley & Sons, Ltd.

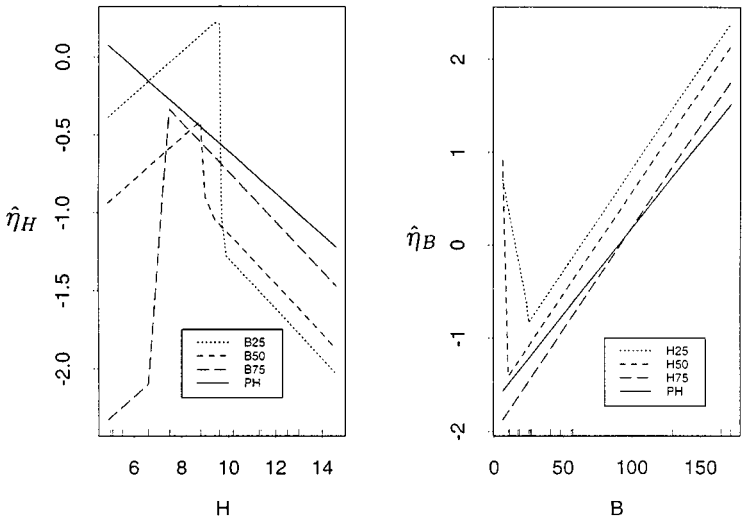

Figure 5. $\hat{\eta}_H$ versus H, on the left-hand side, where B is fixed at its sample quartiles. On the right, $\hat{\eta}_B$ versus B, where H is fixed at its sample quartiles. Also shown are $\hat{\beta}_H^{\mathrm{PH}}\mathrm{H} + \hat{\beta}_B^{\mathrm{PH}}B$, where H and B are fixed in turn at their upper quartile

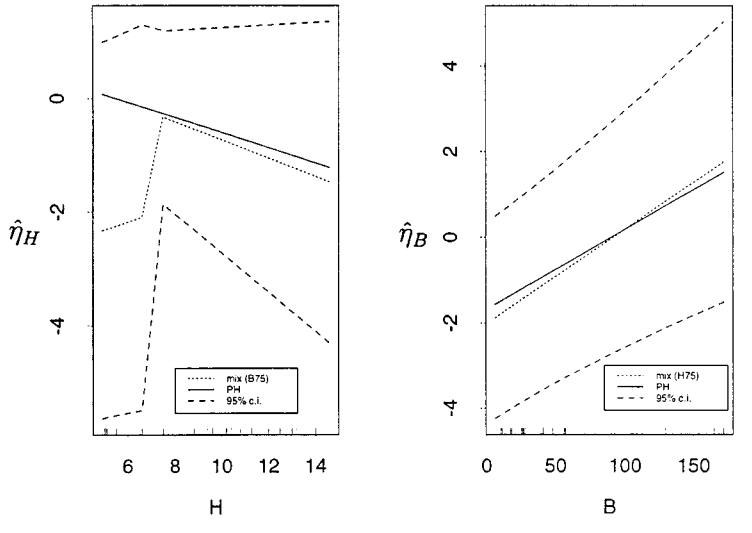

Figure 6. Approximate 95 per cent level confidence intervals, based on the mixture model, superimposed on $\hat{\eta}_H$ and $\hat{\beta}_{\mathrm{H}}^{\mathrm{PH}}H + \hat{\beta}_B^{\mathrm{PH}}\mathrm{B}$ versus H on the left-hand side, and the analogous quantities versus B, on the right. The covariates are fixed in turn at their upper quartile

Figure 7 shows estimates of the survivor function, based on the proportional hazards model and the mixture model, computed at the sample means $\bar{H}$ and $\bar{B}$. There is a close agreement between the two estimates. The discrepancies between the two estimation methods, apparent in Figure 5 and Figure 6, have only a negligible effect on the corresponding survivor estimates.

Copyright © 1999 John Wiley & Sons, Ltd.

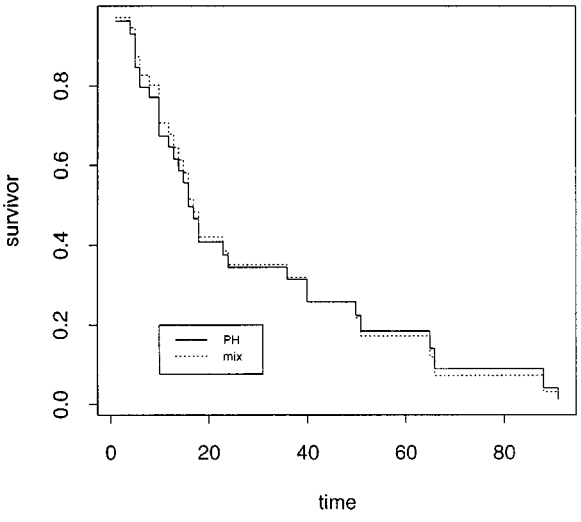

Figure 7. Survivor estimates

4. CONCLUSION

Since any statistical model can be considered an 'expert', the mixtures-of-experts paradigm provides a generic approach for relaxing the parametric assumptions in the systematic component of *any* statistical model. Jacobs *et al.*[4] consider the 'non-parameterization' of the generalized linear model. Rosen *et al.*[19] consider the marginal model. This paper presents a mixture model which combines features of the usual proportional hazards model and those of mixtures-of-experts. In particular, the proportional-hazards assumption is retained, but the underlying log hazard-ratio can take a more general form than the linear combination of predictors inherent in the usual proportional hazards model. Thus, the mixtures-of-experts model provides a flexible approach to modelling survival data and to model checking.

Data analysts have at their disposal an extensive armamentarium for the analysis of censored data. The methodology of Hastie and Tibshirani[7] can address data with underlying smooth additive functions. The approach of LeBlanc and Crowley[11] can approximate piecewise constant functions that may not be additive. The methods of Kooperberg *et al.*[6] fall somewhere in between. In the examples considered in this paper, we see that the mixtures-of-experts approach can be used to model non-differentiable surfaces in a possibly non-additive manner. Jiang and Tanner[23] present formal regularity conditions for which a mixtures-of-experts model applied in the generalized linear model context is consistent, as well as present theoretical convergence rates. Under investigation are parallel results for mixtures of Cox experts.

APPENDIX: FIRST AND SECOND DERIVATIVES

We first define the following quantities ($\boldsymbol{x}' = (1, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}}$):

$$c_j = \sum_{r=1}^{I} g_r^{(j)} \exp(\boldsymbol{\beta}_r^{\mathrm{T}} \boldsymbol{x}_j)$$

Copyright © 1999 John Wiley & Sons, Ltd.

$$\boldsymbol{a}_{mj} = \frac{\partial c_j}{\partial \boldsymbol{\beta}_m} = g_m^{(j)} \exp(\boldsymbol{\beta}_m^{\mathrm{T}} \boldsymbol{x}_j) \boldsymbol{x}_j$$

$$A_{mj} = \frac{\partial \boldsymbol{a}_{mj}}{\partial \boldsymbol{\beta}_m} = g_m^{(j)} \exp(\boldsymbol{\beta}_m^{\mathrm{T}} \boldsymbol{x}_j) \boldsymbol{x}_j \boldsymbol{x}_j^{\mathrm{T}}$$

$$\boldsymbol{b}_{mj} = \frac{\partial c_j}{\partial \boldsymbol{v}_m} = g_m^{(j)} [(1 - g_m^{(j)}) \exp(\boldsymbol{\beta}_m^{\mathrm{T}} \boldsymbol{x}_j) - \sum_{i \neq m} g_i^{(j)} \exp(\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{x}_j)] \boldsymbol{x}_j'$$

$$B_{mj} = \frac{\partial \boldsymbol{b}_{mj}}{\partial \boldsymbol{v}_m} = g_m^{(j)} (1 - 2g_m^{(j)}) [(1 - g_m^{(j)}) \exp(\boldsymbol{\beta}_m^{\mathrm{T}} \boldsymbol{x}_j) - \sum_{i \neq m} g_i^{(j)} \exp(\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{x}_j)] \boldsymbol{x}_j' \boldsymbol{x}_j'^{\mathrm{T}}$$

$$D_{mnj} = \frac{\partial \boldsymbol{b}_{mj}}{\partial \boldsymbol{v}_n} = g_m^{(j)} g_n^{(j)} \Bigg[(2g_m^{(j)} - 1) \exp(\boldsymbol{\beta}_m^{\mathrm{T}} \boldsymbol{x}_j) - (1 - g_n^{(j)}) \exp(\boldsymbol{\beta}_n^{\mathrm{T}} \boldsymbol{x}_j) + \sum_{i \neq m} g_i^{(j)} \exp(\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{x}_j) + \sum_{i \neq m,n} g_i^{(j)} \exp(\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{x}_j) \Bigg] \boldsymbol{x}_j' \boldsymbol{x}_j'^{\mathrm{T}}$$

$$F_{mj} = g_m^{(j)} (1 - g_m^{(j)}) \exp(\boldsymbol{\beta}_m^{\mathrm{T}} \boldsymbol{x}_j) \boldsymbol{x}_j' \boldsymbol{x}_j^{\mathrm{T}}$$

$$G_{mnj} = g_n^{(j)} g_m^{(j)} \exp(\boldsymbol{\beta}_n^{\mathrm{T}} \boldsymbol{x}_j) \boldsymbol{x}_j \boldsymbol{x}_j'^{\mathrm{T}}.$$

The first and second derivatives of $l = \log \mathrm{PL}_g$ are given as functions of the above expressions:

$$\frac{\partial l}{\partial \boldsymbol{v}_m} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{\boldsymbol{b}_{mj}}{c_j} - \frac{\sum_{l \in \mathscr{R}_i} \boldsymbol{b}_{ml}}{\sum_{l \in \mathscr{R}_i} c_l} \right\}$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}_m} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{\boldsymbol{a}_{mj}}{c_j} - \frac{\sum_{l \in \mathscr{R}_i} \boldsymbol{a}_{ml}}{\sum_{l \in \mathscr{R}_i} c_l} \right\}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^{\mathrm{T}}} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{c_j A_{mj} - \boldsymbol{a}_{mj} \boldsymbol{a}_{mj}^{\mathrm{T}}}{c_j^2} - \frac{(\sum_{l \in \mathscr{R}_i} c_l)(\sum_{l \in \mathscr{R}_i} A_{ml}) - (\sum_{l \in \mathscr{R}_i} \boldsymbol{a}_{ml})(\sum_{l \in \mathscr{R}_i} \boldsymbol{a}_{ml})^{\mathrm{T}}}{(\sum_{l \in \mathscr{R}_i} c_l)^2} \right\}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_n^{\mathrm{T}}} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{-\boldsymbol{a}_{mj} \boldsymbol{a}_{nj}^{\mathrm{T}}}{c_j^2} + \frac{(\sum_{l \in \mathscr{R}_i} \boldsymbol{a}_{nl})(\sum_{l \in \mathscr{R}_i} \boldsymbol{a}_{ml})^{\mathrm{T}}}{(\sum_{l \in \mathscr{R}_i} c_l)^2} \right\}, \quad m \neq n$$

$$\frac{\partial^2 l}{\partial \boldsymbol{v}_m \partial \boldsymbol{v}_m^{\mathrm{T}}} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{c_j B_{mj} - \boldsymbol{b}_{mj} \boldsymbol{b}_{mj}^{\mathrm{T}}}{c_j^2} - \frac{(\sum_{l \in \mathscr{R}_i} c_l) \sum_{l \in \mathscr{R}_i} B_{ml} - (\sum_{l \in \mathscr{R}_i} \boldsymbol{b}_{ml})^{\mathrm{T}} (\sum_{l \in \mathscr{R}_i} \boldsymbol{b}_{ml})^{\mathrm{T}}}{(\sum_{l \in \mathscr{R}_i} c_l)^2} \right\}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{v}_m \partial \boldsymbol{v}_n^{\mathrm{T}}} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{c_j D_{mnj} - \boldsymbol{b}_{mj} \boldsymbol{b}_{nj}^{\mathrm{T}}}{c_j^2} - \frac{(\sum_{l \in \mathscr{R}_i} c_l) \sum_{l \in \mathscr{R}_i} D_{mnl} - (\sum_{l \in \mathscr{R}_i} \boldsymbol{b}_{ml})(\sum_{l \in \mathscr{R}_i} \boldsymbol{b}_{nl})^{\mathrm{T}}}{(\sum_{l \in \mathscr{R}_i} c_l)^2} \right\}, \quad m \neq n$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{v}_m^{\mathrm{T}}} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{c_j F_{mj} - \boldsymbol{a}_{mj} \boldsymbol{b}_{mj}^{\mathrm{T}}}{c_j^2} - \frac{(\sum_{l \in \mathscr{R}_i} c_l) \sum_{l \in \mathscr{R}_i} F_{ml} - (\sum_{l \in \mathscr{R}_i} \boldsymbol{a}_{ml})(\sum_{l \in \mathscr{R}_i} \boldsymbol{b}_{ml})^{\mathrm{T}}}{(\sum_{l \in \mathscr{R}_i} c_l)^2} \right\}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{v}_m \partial \boldsymbol{\beta}_n^{\mathrm{T}}} = \sum_{i=1}^{k} \sum_{j \in \mathscr{D}_i} \left\{ \frac{-c_j G_{mnj} - g_n^{(j)} \boldsymbol{a}_{mj} \boldsymbol{b}_{mj}^{\mathrm{T}}}{c_j^2} - \frac{(\sum_{l \in \mathscr{R}_i} c_l)(-\sum_{l \in \mathscr{R}_i} G_{mnl}) - (\sum_{l \in \mathscr{R}_i} \boldsymbol{a}_{nl})(\sum_{l \in \mathscr{R}_i} \boldsymbol{b}_{ml})^{\mathrm{T}}}{(\sum_{l \in \mathscr{R}_i} c_l)^2} \right\}, \ m \neq n.$$

Copyright © 1999 John Wiley & Sons, Ltd.

ACKNOWLEDGEMENTS

O. Rosen was supported by a postdoctoral fellowship grant from the Mellon Foundation. M. Tanner was supported by NIH grant CA35464. The authors wish to thank the editor and the referees for their comments and suggestions which greatly improved this paper.

REFERENCES

1. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. 'Adaptive mixtures of local experts', *Neural Computation*, **3**, 79–87 (1991).
2. Peng, F., Jacobs, R. A. and Tanner, M. A. 'Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition', *Journal of the American Statistical Association*, **91**, 953–960 (1996).
3. Jordan, M. I. and Jacobs, R. A. 'Hierarchical mixtures of experts and the EM algorithm', *Neural Computation*, **6**, 181–214 (1994).
4. Jacobs, R. A., Tanner, M. A. and Peng, F. 'Bayesian inference for hierarchical mixtures-of experts with applications to regression and classification', *Statistical Methods in Medical Research*, **5**, 375–390 (1996).
5. Jacobs, R. A., Peng, F. and Tanner, M. A. 'A Bayesian approach to model selection in hierarchical mixture-of-experts architectures', *Neural Networks*, **10**, 231–241 (1997).
6. Kooperberg, C., Stone, C. and Truong, Y. K. 'Hazard regression', *Journal of the American Statistical Association*, **90**, 78–94 (1995).
7. Hastie, T. and Tibshirani, R. *Generalized Additive Models*, Chapman and Hall, London, 1990.
8. O'Sullivan, F. 'Nonparametric estimation of relative risk using splines and cross-validation', *SIAM Journal of Scientific and Statistical Computing*, **9**, 531–542 (1988).
9. Sleeper, L. and Harrington, D. 'Regression splines in the Cox model with application to covariate effects in liver disease', *Journal of the American Statistical Association*, **85**, 941–949 (1990).
10. Gray, R. 'Flexible methods for analyzing survival data using splines with application to breast cancer prognosis', *Journal of the American Statistical Association*, **87**, 942–951 (1992).
11. LeBlanc, M. and Crowley, J. 'Relative risk trees for censored data', *Biometrics*, **48**, 411–425 (1992).
12. Breiman, L., Friedman, J., Olshen, R. and Stone, C. *Classification and Regression Trees*, Wadsworth, Belmont, California, 1984.
13. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
14. Peto, R. 'Contribution to the discussion of Cox', *Journal of the Royal Statistical Society, Series B*, **34**, 205–207 (1972).
15. Breslow, N. E. 'Covariance analysis of censored survival data', *Biometrics*, **30**, 89–99 (1974).
16. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
17. Breslow, N. E. 'Contribution to the discussion of Cox', *Journal of the Royal Statistical Society, Series B*, **34**, 216–217 (1972).
18. Miller, R. G. *Survival Analysis*, Wiley, New York, 1981.
19. Rosen, O., Jiang, W. and Tanner, M. 'Mixtures of marginal models', in press (1999).
20. Krall, J. M., Uthoff, V. A. and Harley, J. B. 'A step-up procedure for selecting variables associated with survival', *Biometrics*, **31**, 49–57 (1975).
21. Collett, D. *Modelling Survival Data in Medical Research*, Chapman and Hall, London, 1994.
22. Cronin, K., Slate, E. H., Turnbull, B. W. and Wells, M. T. 'Using the Gibbs sampler to detect changepoints: application to PSA as a longitudinal marker of prostate cancer', *in* Sall, J. and Lehman, A. (eds), *Computing Science and Statistics*, Interface Foundation of North America, 1994, pp. 314–318.
23. Jiang, W. and Tanner, M. A. 'Hierarchical mixtures-of-experts for exponential family regression models with generalized linear mean functions: A survey of approximation and consistency results', *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 296–303 (1998).

Copyright © 1999 John Wiley & Sons, Ltd.