Nets and Trees in Survival Analysis

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

ripley@stats.ox.ac.uk
http://stats.ox.ac.uk/∼ripley

Paper at
http://stats.ox.ac.uk/pub/bdr/NNSM.ps.gz

Overview

- Classical approaches
  - *via* classification problems
  - *via* censored regression
- Tree-based survival analysis
- Neural networks primer
- Neural networks in survival analysis
- Two case studies

The Problem

Typical problem:

Data on a few hundred cancer patients, with survival time (either time to death or length of disease/relapse-free survival). Aims:

- test efficacy of proposed prognostic factors, and
- predict survival for future patients

The second has gained most attention, and there are now many studies suggesting that non-linear methods can outperform Cox regression and clinicians.

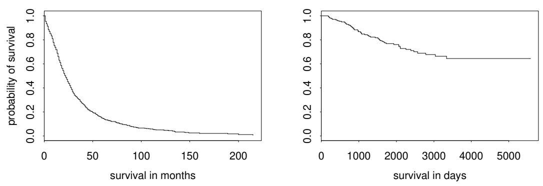
Views differ: contrast *Sunday Times* article (based on a *Lancet* paper) with conclusions of Schwarzer, Vach and Schumacher (1997):

‘The results of a search in the medical literature from 1991 to 1995 on applications of ANNs in oncology and some important common mistakes are reported. It is concluded that there is no evidence that ANNs can provide real progress in the field of diagnosis and prognosis in oncology.’

Case Studies

Advanced Breast Cancer

Survival in months (up to 18 years, but with a median of 23 months) from advanced breast cancer, supplied by Dr J.-P. Nakache. There are 981 patients and 12 explanatory features all of which are categorical. We randomly divided this into a training set of size 500 and a test set of size 476 and assessed the methods on predictions of survival for 24 months; only 3% of the patients did not have complete follow-up to that time.



Plots of the Kaplan-Meier estimates of survival curves for the full (left) breast cancer and (right) melanoma datasets.

Malignant Melanoma

The second dataset is of 205 patients with malignant melanoma following a radical operation, and has five explanatory features.

From Andersen, Borgan, Gill, Keiding (1993).

Long-term survival for 65% of patients, only 57 of the patients died from the melanoma during the study.

The survival distribution does not follow any of the standard distributions.

We assessed methods on their ability to predict survival to 2500 days, by which point 86 of the patients had incomplete follow-up; our analysis shows that we expect 82 of these to have survived for 2500 days.

Classical Approaches

Classification

(a) If we are interested in 2-year survival, dichotomize the survival time at 2 years. Then have a classification problem, for which we could use logistic discrimination:

$$\log \frac{P(\text{survival} \mid \boldsymbol{x})}{P(\text{death} \mid \boldsymbol{x})} = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Problem is *censoring*: some patients are not observed for the full 2 years. These are usually omitted: **this is invalid** and can seriously bias the results.

Ravdin & Clark (1992):

268 patients had known follow-up for 60 months
213 had died
Kaplan-Meier estimate of the survival probability was 50%.

Our melanoma example:

Of those patients with complete follow up to 10 years, 23 out of 80 survived, yet the Kaplan-Meier estimate of survival for this time is 64.5%.

(b) Polytomize time, into, say, death in the first year, second year, third year or survive for more than three years.

Can use multiple logistic regressions:

$$\log \frac{P(\text{class } j \mid \boldsymbol{x})}{P(\text{class } 1 \mid \boldsymbol{x})} = \eta_j(\boldsymbol{x}), \qquad j = 2, \ldots, k$$

and so

$$P(\text{class } j \mid \boldsymbol{x}) = \frac{e^{\eta_j}}{\sum_{c=1}^{k} e^{\eta_c}}, \qquad \eta_1 \equiv 0$$

or proportional odds models:

$$\log \frac{P(\text{class} \geq j \mid \boldsymbol{x})}{P(\text{class} < j \mid x)} = \phi_j + \eta(\boldsymbol{x})$$

('ordinal logistic regression').

Also, conditional logistic regressions:

$$\log \frac{P(\text{class } j \mid \text{class} \geq j, \boldsymbol{x})}{P(\text{class} > j \mid \text{class} \geq j, \boldsymbol{x})} = \eta_j(\boldsymbol{x})$$

These can all take censoring into account to some extent, provided the log-likelihood L is adjusted to be the sum of log probability of known events. Thus the contribution to the L for a patient who was lost to follow up after 2 years is

$$\log\left\{P(\text{death in year 3} \mid \boldsymbol{x}) + P(\text{survive 3+ years} \mid \boldsymbol{x})\right\}$$

It is possible to fit conditional logistic regressions by standard software (Efron, 1988, JASA). Enter each patient once for each class for which the outcome is known. But special-purpose software is much more efficient.

Censored regression

The 'obvious' model is to regress $\log T$ on $\boldsymbol{x}$, an *accelerated-life* model.

$$\log T = \eta(\boldsymbol{x}) + \epsilon, \qquad T = T_0 e^{\eta(\boldsymbol{x})}$$

Standard choices for T_0 are Weibull, log-normal and log-logistic.

Other standard procedures are *proportional hazards* models:

$$h(t) = h_0(t) e^{\eta(\boldsymbol{x})}$$

and proportional odds models:

$$\frac{1 - F(t)}{F(t)} = e^{\eta(\boldsymbol{x})} \frac{1 - F_0(t)}{F_0(t)}$$

Using a Weibull regression gives a proportional hazards model whereas a log-logistic regression is a proportional odds model.

The Cox (1972) procedure is a proportional hazards model in which the baseline hazard h_0 is unspecified. The parameters in η are chosen by maximizing a partial likelihood, roughly [ties?]

$$\prod_{\text{deaths } j} \frac{e^{\eta_j}}{\sum_{R(t_j)} e^{\eta_i}}$$

where $R(t_j)$ is the 'at risk' set at time $T = t_j$.

The log partial likelihood is not a sum over patients. Breslow (1974) justifies the following alternating algorithm

1. Estimate the baseline cumulative hazard $H_0(t) = \int_0^t h_0(u)\, \mathrm{d}u$ by the Breslow estimator

$$\widehat{H}_0(t) = \sum_{\text{deaths before } t} \frac{e^{\eta_j}}{\sum_{R(t_j)} e^{\eta_i}}$$

2. Choose the parameters in η to maximize the log-likelihood for an exponential regression model for time run at rate $\widehat{H}_0$.

$$L(\eta) = \sum_{\text{deaths}} \eta_i - \sum_i e^{\eta_i} \widehat{H}_0(t_i)$$

With a good starting point, few iterations are needed. This only involves sums over patients.

Non-linear Survival Analysis

There are non-linear versions of all these procedures which replace the linear predictor

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

by a non-linear function of the covariates $\boldsymbol{x} = (x_i)$. [Some methods need a vector η.]

- Interactions, product terms, transformations of the covariates would be used in the classical approach (but are often not used in the comparisons).
- Smooth terms – so-called additive models. Both non-parametric smoothers and splines (with knot control, smoothing splines, tensor splines).
- Piecewise constant functions based on partitioning the sample space (by trees).
- Very general parametric families $\eta(\boldsymbol{x}; \theta)$—neural networks.

Tree-structured Analyses

Long history in classification problems, in engineering, machine learning, computer science and statistics (Breiman, Friedman, Olshen, Stone, 1984). Reviewed in Ripley (1996, chapter 7).

Idea is to recursively partition the dataset, at each stage choosing groups of homogeneous experience by splitting on a single feature. Then the leaves of the tree partition the sample space, and we give a single prediction (class, survival time, relative hazard, …) for each cell of the partition.

Too hard to know when to stop dividing; better to grow a large tree and then prune it. For classifications trees BFOS introduced a penalty for the tree size:

$$\text{number of errors} + \alpha \times \text{number of leaves}$$

and an algorithm to find the best subtree for all α (it is a nested family).

Then select α for the problem by cross-validation.

Deviance approach

Several people suggested viewing this as a model, and choosing each split to maximize the log-likelihood then pruning by a penalized deviance (which is AIC-like, and not so good an idea as error rate).

In regression trees use normal likelihood.

In Poisson trees, use Poisson GLM, that is explain counts by covariates (Therneau).

Exponential survival models are equivalent to Poisson GLMs; have 1 event or zero if censored, for each patient.

Several papers have derived exponential survival models from scratch in this framework.

Other ideas

Therneau, Grambsch & Fleming (1990) suggested using regression trees on the martingale residuals from a simple Cox model.

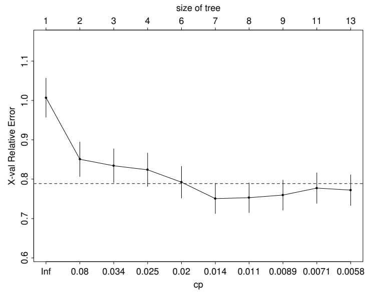
Lots of *ad hoc* splitting and pruning rules in the literature.

The alternating algorithm works well.

Breast Cancer Example

Used a heavily modified and corrected version of the RPARTS package by Anderson and Therneau. Earlier version of this dataset.

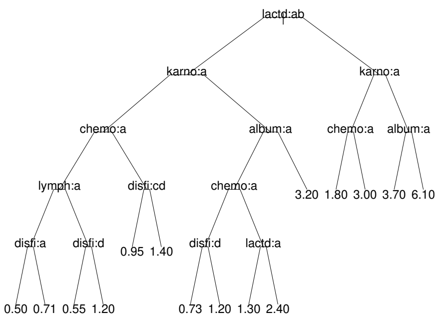
	CP	nsplit	rel error	xerror	xstd
1	0.15643	0	1.000	1.003	0.0493
2	0.04056	1	0.844	0.850	0.0440
3	0.02871	2	0.803	0.818	0.0423
4	0.02166	3	0.774	0.808	0.0411
5	0.01828	5	0.732	0.789	0.0403
6	0.01117	6	0.713	0.736	0.0376
7	0.01011	7	0.702	0.734	0.0375
8	0.00778	8	0.692	0.733	0.0378
9	0.00650	10	0.677	0.750	0.0395
10	0.00524	13	0.658	0.748	0.0397
11	0.00407	15	0.647	0.741	0.0399



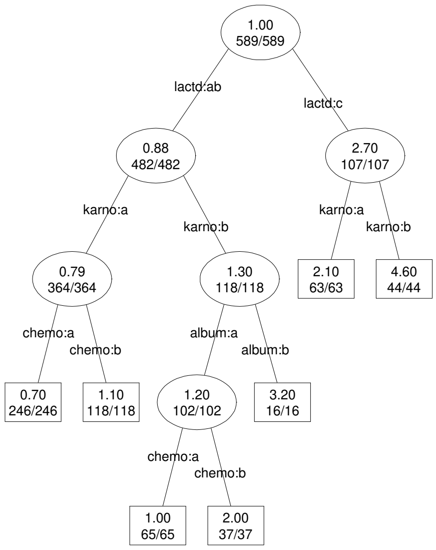

```
> print(breast.rp, cp=0.015)
node), split, n, deviance, yval
      * denotes terminal node

 1) root 589 580 1.00
   2) lactd:<180,180-380 482 400 0.88
     4) karno:<60 364 250 0.79
       8) chemo:no 246 160 0.70 *
       9) chemo:yes 118  75 1.10 *
     5) karno:>60 118 130 1.30
      10) album:<30 102 100 1.20
        20) chemo:no 65  68 1.00 *
        21) chemo:yes 37  22 2.00 *
      11) album:>30 16  13 3.20 *
   3) lactd:380+ 107  96 2.70
     6) karno:<60 63  50 2.10 *
     7) karno:>60 44  29 4.60 *
```

Numbers at leaves are relative hazards.

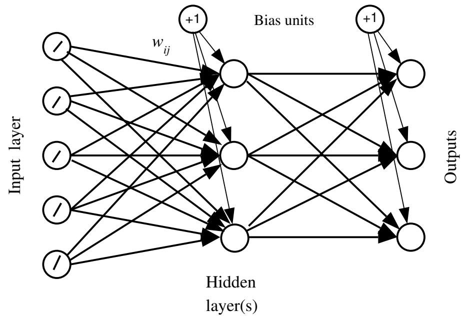


Another example is given in the case studies. For the malignant melanoma data only a single split is suggested.



The numbers m/n mean m uncensored out of n patients reaching that node.

Neural Networks Primer



A generic feed-forward neural network.

In mathematics:

$$y_k = \phi_o\Big(\alpha_k + \sum_j w_{jk}\,\phi_h\Big(\alpha_j + \sum_i w_{ij}\,x_i\Big)\Big)$$

where

- ϕ_h is usually the logistic $\phi_h(x) = e^x/(1+e^x)$
- ϕ_o is linear, logistic or threshold
- the 'biases' α_i can be replaced by weights to $+1$
- the number of *hidden units* can vary
- sometimes weights are set to zero or equal.

Include 'skip-layer' connections from input to output to allow linear part:

$$y_k = \phi_o\Big(\alpha_k + \sum_i w_{ik} x_i + \sum_j w_{jk}\,\phi_j\Big(\alpha_j + \sum_i w_{ij}\,x_i\Big)\Big)$$

Statisticians may like to think as this as a regression of logistic regressions plus the original variables.

Universal approximation

FFNNs provide functions $g\colon \mathbb{R}^p \to \mathbb{R}^k$, of 'complexity' N and some parameters θ. Both are *universal approximators* in the sense that the family of functions is dense in the topologies of uniform convergence on compacta and in $L_q(\mu)$ for any probability measure μ on $\mathbb{R}^p$, and any q.

There are even results on the rate of convergence in $L_2(\mu)$, $O(1/\sqrt{N})$, but they apply to quite restricted classes of target functions.

In other words, they can in principle model anything (noise as well as signal).

Training a FFNN

Fitting a FFNN is normally done by least-squares or maximum likelihood, although unconventional optimization algorithms are used.

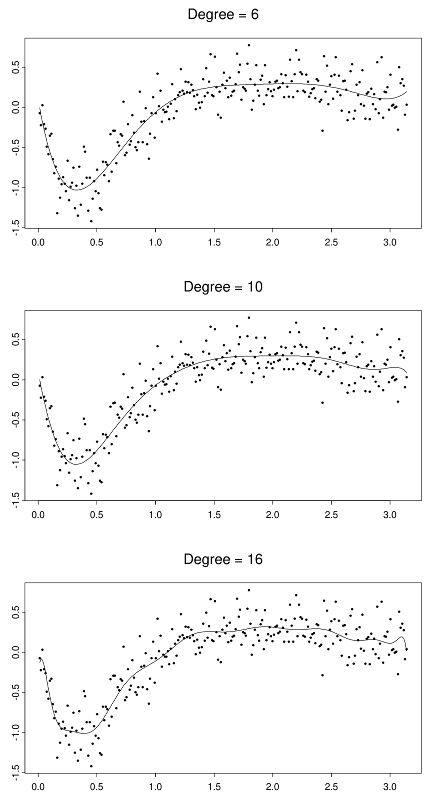
What's different about neural networks, compared to curve-fitting as done in statistics?

Main differences are:

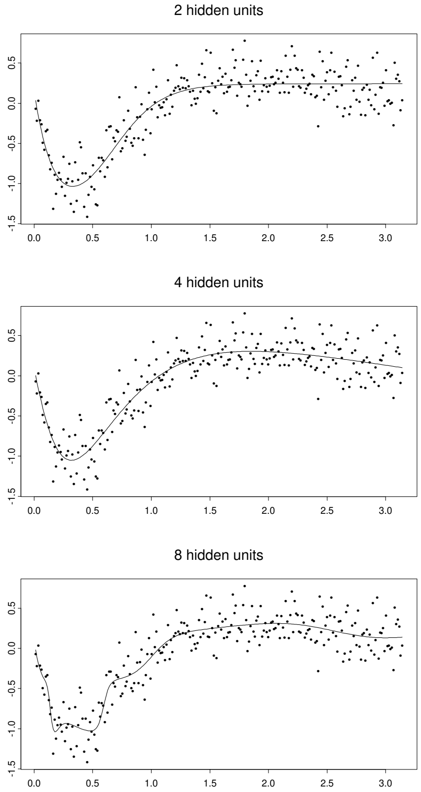
- Multiple local minima are common (especially for large models).
- Over-fitting is controlled by non-likelihood fitting.
- Sequence of possible architectures is not necessarily nested.

Consequence:

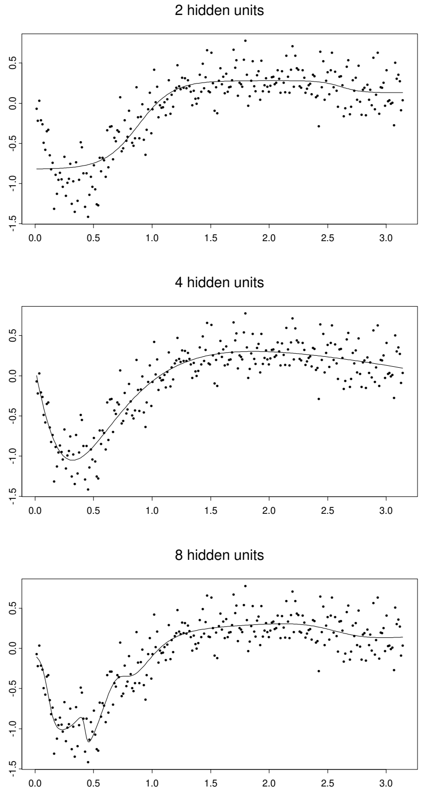
Everyone fitting polynomials will get the same answer, but almost no two people fitting neural networks will. (*Exception:* networks without hidden layers.)



Fits with polynomials.



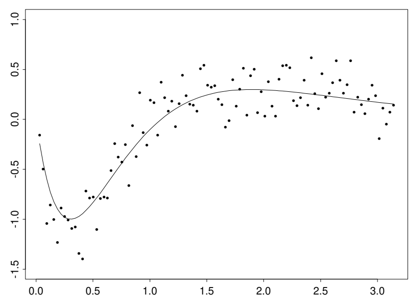
Fits with feed-forward neural networks.



More fits with feed-forward neural networks.

Demonstration example

Taken from Wahba & Wold (1975) where it illustrates smoothing splines.

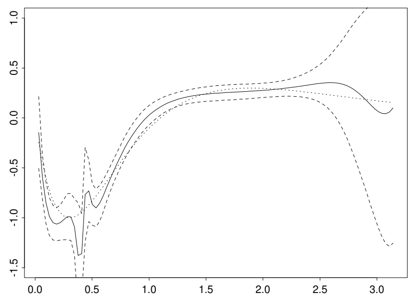


100 data points from a curve-fitting problem, with the true curve.

Throughout we assume that the true noise variance, $(0.2)^2$, is known.

Regression Neural Networks

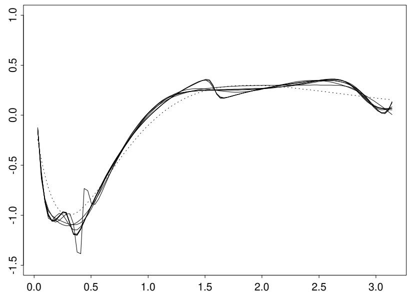
Model selection amounts to choosing the number of hidden units.



Curve fitted by 1-8-1 neural network, with ± 2 standard error bands and true curve (dotted).

The error limits come from standard regression theory, and are based on local linearization.

One problem is multiple local minima:



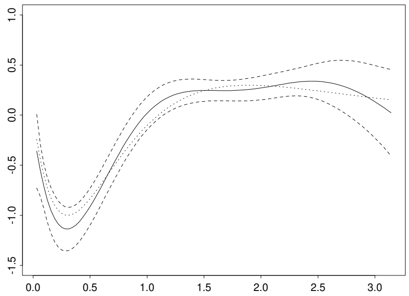
1-8-1 neural networks with different starting weights.

We clearly have over-fitting. Many ways to avoid this. Perhaps the best is to use *regularization* as used for smoothing splines. Somewhat easier for neural networks to use *weight decay*:

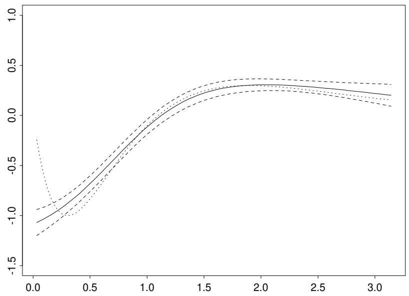
$$\sum_{i=1}^{n}[y_i - f(x_i;\mathbf{w})]^2 + \lambda \sum_{ij} w_{ij}^2$$

with the inputs and outputs suitably scaled.

Effect of Weight Decay

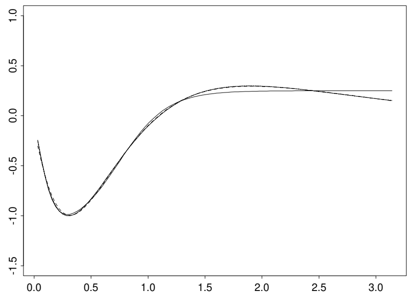


Curve fitted by 1-8-1 neural network, with ± 2 standard error bands and true curve $\lambda = 10^{-3}$ (above) and $\lambda = 0.1$ (below).



Bias/variance Compromises

We now have two ways to control the fit, the number of hidden units and λ, the degree of regularization. The true curve is not exactly representable by our model, so we have some bias.



Fitting to the true curve with 2 and 3 hidden units, and 8 units with $\lambda = 10^{-3}$.

Choosing fewer hidden units leads to more bias, as does adding a regularizer. But it also reduces their variability, so reduces the mean-square error.

Neural Networks for Survival Analysis

Neural networks give us a flexible family $\eta(\boldsymbol{x};\theta)$. We wrote S code to interface to my `nnet` package to allow us to alter the use of η and the fit criterion very quickly.

We also implemented the direct maximization of the log-partial likelihood for proportional hazards model, for comparison. (This was suggested earlier by Faraggi & Simon, 1995). The conditional logistic regression model was suggested by Liestøl, Andersen & Andersen (1994); the other polytomized models have been widely used.

Results on Case Studies

sensitivity is the probability of correctly predicting death,

specificity is the probability of correctly predicting survival,

accuracy is the percentage of correct predictions.

linear methods:

selection of the input variables by AIC

neural networks:

number of hidden units and the amount of weight decay chosen by 10-fold CV within the training set

Advanced Breast Cancer

training set of size 500

test set of size 476 (omitting 5 censored)

no censoring

Advanced Breast Cancer

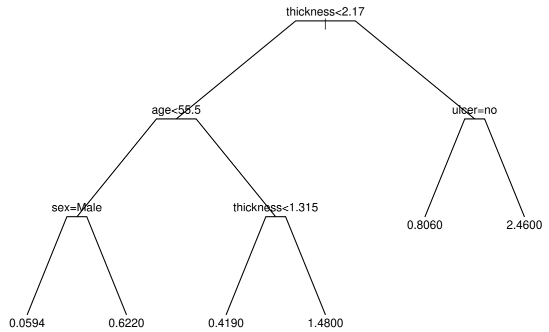
method	linear			neural net		
	specificity	sensitivity	accuracy	specificity	sensitivity	accuracy
binary classification	73	62	67	72	64	68
1-year periods	72	63	68	72	65	68
proportional odds				71	62	66
regression	66	68	67	63	71	67
proportional hazards	70	62	66	71	62	66
Weibull survival	72	58	64	72	61	66
log-logistic survival	70	66	67	68	66	67

Further investigations showed that the bias towards death of the regression models is due to the exclusion of six cases with incomplete follow-up to 24 months (which were also excluded for the binary classifications).

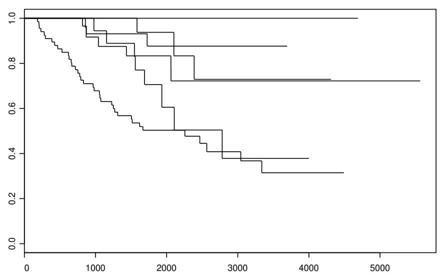
Malignant Melanoma

Because there was heavy censoring, assessment on just those patients with complete follow-up to 2500 days would be seriously biased. We used a tree-based analysis to divide the dataset into six groups with homogeneous survival experience, fitted Kaplan-Meier survival curves to each groups, and used these to estimate the probability that the patient would have survived from the end of observed follow up to 2500 days. (This probability was often one, and never less than 0.45.) These patients were then entered into the test set with both possible outcomes, weighted by the estimated probabilities.

We used 5-fold cross-validation to assess the models, that is we randomly divided the dataset into 5 parts and for each fitted to the remaining four parts and predicted survival on the single part.



Tree used to split the Melanoma data into six groups. At each node the label indicates the condition to go down the left branch, and the numbers are the hazards for the groups relative to the whole dataset.



Kaplan-Meier plots of survival in the six groups.

Malignant Melanoma

method	linear specificity	linear sensitivity	linear accuracy	neural net specificity	neural net sensitivity	neural net accuracy
binary classification	45.1	40.1	43.7	63.4	60.7	62.6
full training set	90.9	13.2	69.0	90.9	16.5	69.9
4-class	88.1	21.5	69.3	92.2	14.8	70.4
proportional odds	90.4	23.9	71.6	91.0	18.7	70.6
proportional hazards	84.3	32.8	69.8	87.6	34.3	72.6
Weibull survival	87.0	25.8	69.8	87.0	24.1	69.3
log-logistic survival	86.4	36.2	72.3	84.2	34.3	70.1

Results (%) from 5-fold cross-validation of the melanoma example. The second row of binary classification is using the estimated probabilities as targets for the patients with incomplete follow-up to 2500 days; these patients are completely ignored in the first line.

4-class problem had classes as 0–1500, 1500–2000, 2000–2500 and 2500– days, chosen by looking at the pattern of censoring times.

Despite the use of nested cross-validation (so that evaluating each neural network method involved $5\times 5\times 12$ fits) the total computation time was less than an hour on a laptop PC.

Generally small differences between the methods , even though the Weibull and log-logistic distributions cannot model long-term survival.

The large differences between sensitivity and specificity is not really surprising given that only about 28.2% of patients die within 2500 days. Thus we would achieve a higher accuracy than *all* of the methods by declaring all patients to survive.

The underlying difficulty is that is hard to find prognostic patterns, and the dominance of survival leads to predicted probabilities of death of individual patients which are above 28% but do not reach 50%.

Malignant Melanoma, with losses

Cost of failing to spot a death as twice that of incorrectly predicting death. Predict death if the probability of survival is less than 2/3.

method	linear specificity	linear sensitivity	linear loss	neural net specificity	neural net sensitivity	neural net loss
binary classification	25.1	56.6	160.5	39.6	74.7	118.1
full training set	75.5	46.5	98.0	73.5	55.3	90.8
4-class	74.3	50.5	95.0	76.2	46.6	96.8
proportional odds	74.8	46.6	98.8	76.2	51.8	90.8
proportional hazards	79.6	58.7	88.4	71.7	64.6	82.6
Weibull survival	73.5	53.5	92.8	72.4	64.6	81.6
log-logistic survival	72.1	53.5	94.8	69.0	66.4	84.5

Only the larger differences (those between the first 'binary classification' line and the rest) are statistically significant when assessed by paired t-tests.