STATISTICS IN MEDICINE
Statist. Med. 2001; **00**:1–23 *Prepared using* simauth.cls *[Version: 2000/06/29 v1.1]*

Improving Cox survival analysis with a neural-Bayesian approach

Bart Bakker, Tom Heskes, Jan Neijt and Bert Kappen

Theoretical Foundation SNN Laboratory
University of Nijmegen
Geert Grooteplein 21
6525 EZ Nijmegen
The Netherlands
{bartb,tom,bert}@mbfys.kun.nl
j.p.neijt@azu.nl

SUMMARY

In this article we show that traditional Cox survival analysis can be improved upon when supplemented with sensible priors and analysed within a neural Bayesian framework. We demonstrate that the Bayesian method gives more reliable predictions, in particular for relatively small data sets.

The obtained posterior (the probability distribution of network parameters given the data) which in itself is intractable, can be made accessible by several approximations. We review approximations by Hybrid Markov Chain Monte Carlo sampling, a variational method and the Laplace approximation. We argue that although each Bayesian approach circumvents the shortcomings of the original Cox analysis, and therefore yields better predictive results, in practice the use of variational methods or Laplace is preferable. Since Cox survival analysis is infamous for its poor results with (too) many inputs, we use the Bayesian posterior to estimate p-values on the inputs and to formulate an algorithm for backward elimination. We show that after removal of irrelevant inputs Bayesian methods still achieve significantly better results than classical Cox.

1. Introduction

The purpose of survival analysis for statistics in medicine is to estimate a patient's chances of survival as a function of time, given the available medical information at t_0, the time the patient is admitted to the study. A well-known way to conduct such an analysis is Cox's proportional hazards method [1]. In this method the hazard function $h(t;x)$, which estimates the probability density of death occurring at time t (given that the patient has survived upto that time), is a product of two independent parts. The first part is the proportional hazard, $h(x) = \exp(w^T x)$, which depends on patient information x only, the second part is a time-dependent baseline hazard $h_0(t)$. The most obvious weakness of this model is its vulnerability to overfitting on the training set, often resulting in poor predictions for future patients.

In the medical statistical community a considerable interest exists in the application of neural networks in survival analysis [2, 3, 4, 5, 6, 7, 8, 9]. Although many authors use the

Contract/grant sponsor: RWCP

 Revised 12 Juli 2001

neural network machinery to model *non*-proportional hazards, the majority of these models is still based on the original Cox method.

The weaknesses of the standard approach have been acknowledged. Solutions exist in the form of weight decay on the model parameters [5] and implementation of penalty terms [3]. However, an integral solution to the shortcomings of the model based on a solid theoretical background still does not exist.

Another important question in survival analysis is how to implement the effect of time in the model. Options are to view time as an input [5, 6] or to see survival analysis as a series of classification problems, where for each time interval the patients are divided into survivors and non-survivors [2, 3, 4].

In this article we hold on to the proportional hazards model since Cox analysis is still the standard in statistical medicine. Furthermore, as we will show in Section 7.2, a more complex model does not improve performance. We implement a discretised version of this model in the form of a multi-layered perceptron with one hidden unit and exponential transfer functions, as will be shown in Section 4. Each output of the network corresponds to an evaluation of the survivor function at a discrete point in time. The possible adverse effects of this discretisation are researched in Section 7.3.

Our neural interpretation suggests a Bayesian analysis to overcome the weaknesses of the standard approach. In Section 5 sensible priors are introduced which, in combination with the available data, lead to a posterior distribution on the model parameters. This posterior is intractable, but with sampling techniques such as Hybrid Markov Chain Monte Carlo (HMCMC) sampling (see e.g. [10]) we can sample from this posterior to obtain an ensemble of neural networks.

A disadvantage of (HMCMC) sampling is that the number of samples that needs to be drawn to describe the posterior properly grows strongly with the number of network parameters. This not only takes a lot of computation time, it also introduces approximation errors. Furthermore, it is difficult to determine when enough samples have been drawn. As an alternative we propose a form of "ensemble learning". This term was coined by Hinton and van Camp [11] and has been applied to multi-layered perceptrons and Radial Basis Function networks in [12, 13]. We approximate the posterior by minimising the Kullback-Leibler (KL) divergence between the exact posterior given by Bayes' formula and an approximating analytical distribution, varying only the parameters of the latter. We also implement a simpler version of this procedure, where instead of minimising a KL-divergence we make a Laplace approximation of the posterior. We compare all approaches by estimating the predictive qualities of the resulting posterior distributions. The results are summarised and compared with other (neural) approaches in the discussion.

In practice, medical experts do not work with probability distributions over model parameters directly: they rather use the concept of p-values, for example to express the relevance of patient characteristics. The Bayesian posterior distribution provides a direct way to calculate p-values for the inputs to the network (i.e. patient information). Another problem present in most survival analysis projects is the tendency to include many sources of medical data: rather than missing a possibly significant explanatory variable, medical experts generally prefer to consider a wide variety of possible inputs in the survival model. When the relevance of some of these variables cannot be demonstrated with the model at hand, it is desirable to eliminate these "irrelevant" inputs from the model, leaving only the inputs with a clear effect on the survival of the patient (see e.g. [14, 15, 16]). In Section 8 we propose a Bayesian algorithm

to select the relevant inputs of the model and use it to remove the irrelevant variables from the models. After this improvement of the survival model, both for the discretised Cox model and the Bayesian approach, we make our final comparison between the various approaches.

The proposed methods are tested on a medical database, which is described further in the next section.

2. Survival analysis on ovarian cancer patients

2.1. Survival analysis

In the present article we study the survival of ovarian cancer patients. Each patient is characterised by a selection of medical data, described in more detail later in this section. Monthly observations of the patients provide information on either the time of death (in months after the initial diagnosis) or survival. The direct task of the methods described in this article is to estimate the probability of survival for any patient and for any month, based on the medical data that is available at the onset of the study. The models providing such estimations will be used to evaluate the (observable) effect of each of the covariates (medical data) on the survival of a patient.

2.2. Patients

A database was constructed including 1023 patients from 4 studies, two studies from The Netherlands Joint Study Group for Ovarian Cancer and two from the Gynaecological Cancer Cooperative Group of the European Organisation for Research and Treatment of Cancer (EORTC). The Dutch studies were initiated in 1979 and 1981, respectively. The first study compared a combination of hexamethylmelamine, cyclophosphamide, methotrexate, and 5-fluorouracil (Hexa-CAF) with cyclophosphamide and hexamethylmelamine alternating with doxorubicin and a 5-day course of cisplatin (CHAP-5) in 186 patients with advanced epithelial ovarian carcinoma. In the study initiated in 1981, 191 eligible patients were enrolled and treated with either CHAP-5 or cyclophosphamide and cisplatin (CP), both administered intravenously on a single day at 3-week intervals. Protocol entry criteria, pretreatment staging, histology grading, the randomisation procedure, assessment and definitions of tumour response, evaluation and statistical methods were the same in both studies.

The EORTC studies compared CHAP-5 and a combination of cyclophosphamide, hexamethylmelamine alternating with doxorubicin and carboplatin, CHAC-1 (EORTC study no. 55836, 1984), or addressed the question about the efficacy of intervention surgery for patients treated with CP (EORTC study no. 55865, 1987).

In the present article we excluded the 94 patients treated with Hexa-CAF, since their medical profile differed largely from the other patients. Of the remaining 929 patients, 246 were censored (i.e. their time of death is not (yet) known). The median observation time of the 929 patients is 653 days.

2.3. Missing Data.

To handle missing values we distinguished categorical and continuous prognostic variables. For categorical variables we added an additional category representing patients with missing data.

For continuous variables we took the average value for that characteristic. Estimating missing values based on the joint probability of missing and known variables did not alter the outcome.

2.4. Transformation of data.

In the Dutch studies the performance status was registered according to the Karnofsky scale. We reclassified them according to the scale used by the Zubrod-ECOG-WHO. Patients with a Karnofsky rating of 100% were classified as ECOG 0, those with 90 or 80% as 1, those with 70 or 60% as 2 and patients with a Karnofsky rating less than 60% were classified as 3. When the Broder's grading system was used, we registered a Broder's grade 1 as a well differentiated tumour, Broder's grade 2 as moderately differentiated and Broder's grades 3 and 4 as poorly differentiated. The dose-intensity of each drug was expressed as mg/m^2/wk administered during the first treatment cycle.

2.5. Variables.

Thirty one variables have been processed by the network: tumour size before and after initial surgery, cell type (serous, endometrioid, mucinous, clear cell, undifferentiated, unclassified, other), grade, weight (kg), length, body surface (m^2), age, treatment, carboplatin dose in cycle 1 divided by the area under the curve (AUC), dose in cycle 1 of doxorubicin, cyclophosphamide, hexamethylmelamine, cisplatin, carboplatin, dose intensity in cycle 1 (mg/m^2/wk), number of sites before and after surgery, the presence of ascites, FIGO stage, performance status according to WHO criteria, hemoglobin (mmol/L), leucocytes (10^9/L), lowest leucocytes (10^9/L), renal clearance (mL/min), number of thrombocytes (10^6/L), lowest number of thrombocytes (10^6/L), serum creatinine (mmol/L), bilirubin (μmol/L).

In order to compare these factors with each other, they were rescaled. For each factor the mean and standard deviation were determined and the mean subtracted from the input values. The result was divided by the standard deviation in such a way that each of the characteristics had a mean equal to zero and a standard deviation equal to one.

3. Cox survival analysis

Given the hazard function $h(t;x) = \exp(w^T x)h_0(t)$ the survivor function $F(t;x)$ indicating the probability to survive upto time t can be formulated as

$$F(t;x) = \exp\left[-\int_0^t dt'\ h(t';x)\right] . \tag{1}$$

The probability density $f(t;x)$ for a patient to expire at time t is then given by

$$f(t;x) = -\frac{\partial F(t;x)}{\partial t} = h(t;x)F(t;x) .$$

The likelihood function $P(D|w,h_0)$, expressing the probability to observe the data in database D given the model parameters w and the specific choice for the baseline hazard h_0, then immediately follows as

$$P(D|w,h_0) = \prod_{\nu\in\text{uncensored}} f(t^\nu;x^\nu) \prod_{\mu\in\text{censored}} F(t^\mu;x^\mu) . \tag{2}$$

The first product is over the patients of whom the time of death is known. An element in the second product specifies the estimated probability of censored patient μ to be alive at time t^μ, the time this patient was taken out of the study. Since in this case the time of death is not known this is the strongest prediction that can be verified.

In classical Cox analysis the model parameters are estimated through optimisation of the likelihood $P(D|w, h_0)$ with respect to w and h_0. An advantage of Cox analysis is that the optimal parameters of the proportional and the time-dependent hazard can be found sequentially. The optimal choice for the parameters w (w^{ML}) depends only on the ordering of the times of death of the patients (see [1]); all other time-dependent information is modelled in the function $h_0(t)$.

Disadvantages of this approach are the tendency of the hazard to become highly non-smooth and the danger of strongly over-fitting the data. In this article we will eliminate these disadvantages of Cox analysis through the introduction of sensible priors (Section 5). We will show (Section 7 and further) that by doing so we obtain a stable, reliable model, which still preserves the elegance and simplicity of the original method.

4. A discretised model

The first step in our approach of survival analysis is to define a framework of parameters that specifies the hazard function. The parameters for the proportional hazard, $h(x) = \exp(w^T x)$, have already been defined. However, the baseline hazard h_0 in classical Cox analysis is a free, continuous function of time. To find the optimal choice for this function one must either choose a specific functional form for the baseline hazard and optimise the corresponding function parameters, or discretise $h_0(t)$ over time. Since the latter option provides more freedom for the form of $h_0(t)$ we will use a discretised version of survival analysis, estimating values for $h_0(t_k)$ for specific discrete points in time. Between two points in time, say t_{k-1} and t_k, we take the hazard function $h(t; x)$ to be constant, i.e.

$$h(t; x) = h(t_k; x), \qquad t_{k-1} < t \leq t_k,$$

where we take all time intervals (t_{k-1}, t_k) to have the same size, determined by the length of the study and the desired number of discrete model outputs. In this way, although the baseline hazard is not inferred directly from the data for each of the infinitely many points in time, it is not restricted to a specific functional form either. In Section 7.3 we will show that for the right number of discrete outputs the model will not suffer from the inaccuracy brought into it by discretisation.

After division of the time span in discrete intervals of length Δt the survivor function reads:

$$F(t; x) = \exp\left[-\sum_{i=1}^{k-1} \Delta t \; h(t_i, x) - (t - t_{k-1}) h(t_k, x)\right], t_{k-1} < t \leq t_k,$$

changing Equation (2) to

$$P(D|w, h_0) = \prod_{\nu \in \text{uncensored}} h(t_k^\nu; x^\nu) F(t^\nu; x^\nu) \prod_{\mu \in \text{censored}} F(t^\mu; x^\mu), \tag{3}$$

where

$$t_{k-1}^\nu < t^\nu \leq t_k^\nu, t_{k-1}^\mu < t^\mu \leq t_k^\mu.$$

Optimisation of the discrete likelihood function proceeds in much the same way as in the classical Cox method, resulting in what we will call the maximum likelihood (ML) Cox solution.

The discrete survivor function can be represented by a multi-layered perceptron with exponential transfer functions, as can be seen in Figure 1 (concentrate for the moment only on the solid lines). The input x consists of the elements of patient information, such as the type of medication administered or the presence of specific symptoms. The weights w in the first layer (input to hidden) correspond to the parameters w in the proportional hazard function $\exp(w^T x)$ which is the output of the hidden unit. When the weights v in the second layer (hidden to output) are equated to

$$v_i = -\int_0^{t_i} dt'\ h_0(t'),$$

i.e. minus the integral upto time t_i of the base-line hazard, the output of the network at neuron i reads

$$F_i(x) = \exp\left[-\int_0^{t_i} dt'\ h_0(t')\exp(w^T x)\right],$$

which, as in Equation (1), equals the probability for a patient with characteristics x to survive upto time t_i.

Given this network structure, it is easy to extend the model beyond proportional hazards. The addition of more hidden units to the second layer (dashed lines) may well provide the possibility to model complex input-output relations, which cannot be found in the proportional hazard approach.

5. Bayesian inference

A discretised version of the Cox proportional hazards method has been defined in Sections 3 and 4. In this section we describe a method to find the optimal choice for the model parameters without overfitting on the training data. We propose a Bayesian approach in which a probability distribution over all possible values of the model parameters will be constructed. This distribution will not only depend on the (medical) data, but also on prior knowledge about the nature of the problem. This prior knowledge is expressed in probability distributions of the model parameters, which are called *priors*. Using Bayes' formula, the priors and the data likelihood can be combined in the posterior distribution that describes the probability of any choice for the model parameters w and v.

The first prior

$$P(v|\gamma) \propto \exp\left[-\frac{\gamma}{2}\sum_{ij} g(|i-j|)[v_i - v_j]^2\right] \propto \exp\left[-\frac{\gamma}{2}v^T\Gamma v\right],$$

where $\Gamma_{ij} = -g(|i-j|), \Gamma_{ii} = \sum_{j\neq i} g(|i-j|)$ and we choose $g(x) = \mathrm{e}^{-x^2/\tau}$, prevents the hazard from becoming too sharp as a function of time. In the classical Cox approach the hazard over a specific time interval is directly proportional to the number of casualties in that interval. Although this may give a good representation of the part of the data the model is trained on, it provides poor generalisation and yields highly non-smooth functions which are not intuitively

plausible. Since $P(v|\gamma)$ assigns the highest likelihood to a hazard function which is constant in time, it smoothes out the hazard function and introduces a preference for survivor functions which decay exponentially.

The effect of this prior is visualised in Figure 2. The hazard function in the ML Cox approach is a jagged function, due to the limited information in the database. After imposing the prior $P(v|\gamma)$ this function becomes much more smooth. This smoother function is not only more plausible *a priori* but, as will be shown in Section 7, it also has a large positive effect on the predictive qualities of the model.

The second prior

$$P(w|\lambda) \propto \exp\left[-\frac{\lambda}{2} w^T \Lambda w\right] \ , \quad \text{where} \quad \Lambda = \frac{1}{\#\ patients} \sum_{\mu} x^{\mu} x^{\mu T} \ ,$$

prevents large activities of hidden units (high values for the proportional hazard), i.e. prefers small weights. This prior corresponds to a ridge-type estimator, as discussed in [17]. Incorporation of the covariance matrix Λ makes this preference independent of a (linear) scaling of the inputs x. The effect of imposing this prior on the parameters w is shown in Figure 3. It can be seen that in the unrestrained case (left panel) it occurs quite often that the proportional hazard of one patient is upto ten times larger than it is for another patient, indicating that the model may be overfitting on the training data. After imposing a Bayesian prior on the model these differences become much more reasonable.

The hyperparameters γ and λ express the confidence one has in the knowledge expressed in the two prior distributions. Since we do not want to specify the exact values of γ and λ, we introduce gamma distributions (defined as $P(\lambda|\sigma,\tau) \propto \lambda^{\sigma-1}\exp(-\tau\lambda))$ $P(\lambda)$ and $P(\gamma)$ for the hyperparameters. The expectation value and the variance of λ are calculated as $\frac{\sigma}{\tau}$ and $\frac{\sigma}{\tau^2}$, respectively. Choosing values for these two ratios enables us to specify our believes about the model in more detail. We use the visualisation of the effect of our first prior (Figure 2) to judge which γ yields acceptable (smooth) hazard functions. In estimating the expected value for λ we can use Figure 3, or we can consider what patient to patient variation in diagnosis may still be acceptable. Using this decision, λ can be set to assign the desired *a priori* probability of the parameters w. The variance will be kept relatively high to obtain very wide hyperpriors.

The posterior distribution of the parameters w and v and hyperparameters λ and γ given the data follows from Bayes' formula:

$$P(w,v,\lambda,\gamma|D) = \frac{P(D|w,v)P(w|\lambda)P(v|\gamma)P(\lambda)P(\gamma)}{P(D)} \ , \tag{4}$$

with $P(D|w,v)$ the likelihood as in (3) and $P(D)$ an irrelevant normalising constant. The probability density of any conceivable choice for the model parameters w and v is given by the posterior $P(w,v|D)$, which follows by integrating out the hyperparameters λ and γ.

6. Approximation of the posterior

6.1. Three methods

In theory the expression for the posterior probability distribution of the model parameters is all that is needed to make predictions for new patients. First however, the posterior distribution

of model parameters and hyperparameters (Equation (4)) needs to be transformed into a distribution of the model parameters alone. To this end we must perform the integral over the hyperparameters λ and γ. Since this cannot be done analytically, we have to make an approximation of the true posterior.

Fortunately, an ample arsenal of methods to approximate $P(w, v|D)$ is available. These methods have become popular tools in the neural networks community. In this section, we will apply and evaluate three such methods: Hybrid Markov Chain Monte Carlo Sampling, a variational approach and the Laplace approximation.

6.2. Hybrid Markov Chain Monte Carlo Sampling

Since the posterior $P(w, v, \lambda, \gamma|D)$ is not a simple analytic function of the model parameters, it is hard to draw samples from it. Therefore, we will use both Gibbs sampling and HMCMC to make it tractable. In Gibbs sampling, one starts by taking random values for the hyperparameters, say $\tilde{\lambda}$ and $\tilde{\gamma}$. Next, $P(w, v|\tilde{\lambda}, \tilde{\gamma}, D)$ is obtained from

$$\begin{aligned} P(w, v|\tilde{\lambda}, \tilde{\gamma}, D) &= \frac{P(D|w, v, \tilde{\lambda}, \tilde{\gamma})P(w, v|\tilde{\lambda}, \tilde{\gamma})}{P(D|\tilde{\lambda}, \tilde{\gamma})} \\ &\propto P(D|w, v)P(w, v|\tilde{\lambda}, \tilde{\gamma}) , \end{aligned}$$

the product of the likelihood and the priors, with fixed values for the hyperparameters. Now, in turns samples are drawn from $P(w, v|\tilde{\lambda}, \tilde{\gamma}, D)$ yielding $\tilde{w}$ and $\tilde{v}$, and from $P(\lambda, \gamma|\tilde{w}, \tilde{v})$,

$$\begin{aligned} P(\lambda, \gamma|\tilde{w}, \tilde{v}) &= \frac{P(\tilde{w}, \tilde{v}|\lambda, \gamma)P(\lambda, \gamma)}{\int d\lambda d\gamma P(\tilde{w}, \tilde{w}|\lambda, \gamma)} \\ &\propto P(\tilde{w}|\lambda)P(\tilde{v}|\gamma)P(\lambda)P(\gamma) , \end{aligned}$$

which is itself a gamma distribution. Since $P(w, v|\tilde{\lambda}, \tilde{\gamma}, D)$ does not have the form of one of the standard distributions (normal, gamma) we use HMCMC to sample from it.

Hybrid Markov Chain Monte Carlo sampling [18] is a well-known method for sampling from complex distributions. The first step in HMCMC is to express $P(q)$, the distribution we wish to sample from, as

$$P(q) = \exp(-E(q)),$$

where in the case of the model discussed in this article, q represents the parameters $\{w, v\}$. Further we define the canonical distribution

$$P(p) = \exp(-K(p)),$$

where

$$K(p) = \sum_{i=1}^{n} \frac{p_i^2}{2m_i},$$

so that each p_i is normal distributed around zero with variance m_i. n is the dimension of p, which is equal to the dimension of q. The joint distribution $P(q, p)$ is defined as

$$P(q, p) = \exp(-E(q)) \exp(-K(p)) .$$

Now we sample from the joint distribution $P(q,p)$. We first take a random choice for q and draw a sample p from the normal distribution $K(p)$. Next, this sample is transformed by applying the following dynamics:

$$\frac{dq_i}{d\tau} = \frac{p_i}{m_i},$$

and

$$\frac{dp_i}{d\tau} = -\frac{\partial E}{\partial q_i}, \tag{5}$$

where τ serves as a simulated "time" parameter. It can be shown that under these rules the probability $P(q,p)$ remains the same. Therefore, the simulation of the evolution of q and p in "time" provides a new and -if a sufficient number of steps in "time" have been taken- independent sample (q^*,p^*), which has the same probability as the old sample (q,p). Since q and p are independent under the distribution $P(q,p)$, this also yields a new sample q^* of $P(q)$. Drawing a new sample from $P(p)$ completes the new sample (q,p), on which the dynamics (5) can be applied again, etc. etc. This way, neglecting p, samples are drawn effectively from $P(q)$.

However, because the applied "time evolution" proceeds in finite steps, the probability $P(q,p)$ may not be conserved exactly. Therefore, before accepting a new sample q^* the ratio of the joint probability of the current sample $\{p^*,q^*\}$ and the previous sample $\{p,q\}$,

$$R_M = \frac{P(q^*,p^*)}{P(q,p)}\,,$$

must be considered. If $R_M > 1$ the new sample is accepted always, otherwise it is accepted with probability R_M. It can be shown [18] that in this way, in spite of the computational inaccuracy introduced in (5), one still samples from $P(q)$. The first few samples may suffer from initialisation effects and therefore cannot be trusted to give a good representation of the posterior. Therefore, after sampling we will discard the first 100 samples.

The advantage of this method is that large jumps in "q-space" may be taken (there can be large differences between subsequent samples of q), yielding a series of independent samples, which is not the case in standard Markov chain sampling. Further, if the time steps in (5) are chosen small enough very few samples will be rejected.

Adding extra hidden units does not change the format of this sampling procedure. It does however cause the number of parameters, and therefore the number of samples that needs to be drawn, to increase strongly.

6.3. Variational approach

Another approximation of $P(w,v|D)$ can be made by fitting an analytical distribution to the exact posterior. An advantage of this approach is that we obtain a simple expression for the posterior. Following Barber and Bishop [12] we approximate the joint posterior distribution of weights and hyperparameters $P(w,v,\lambda,\gamma|D)$ by a factorised distribution of the form

$$P^*(w,v,\lambda,\gamma) = Q(w,v)R(\lambda)S(\gamma)\,,$$

with $Q(w,v) = \mathcal{N}(\hat{w},\hat{v},C)$, a gaussian distribution with mean $\hat{w},\hat{v}$, and $R(\lambda)$ and $S(\gamma)$ for the moment unspecified. We assume that there is no interaction between w and v, so the covariance matrix C can be written as

$$C = \begin{pmatrix} C_{ww} & \emptyset \\ \emptyset & C_{vv} \end{pmatrix}\,,$$

which yields a significant reduction in the number of free parameters. To reduce further the number of free parameters in the covariance matrix, instead of using the full covariance matrix C_{vv} we take a constrained matrix C^*_{vv} specified by

$$[C^*_{vv}]_{ij} = k_i \ \exp(-\Delta_{ij}) \ k_j$$

where k is an unconstrained vector and

$$\Delta_{ij} = |i - j|,$$

which we found to be a good approximation of the real covariance matrix. The strongly reduced number of free parameters in C_{vv} (now equal to the number of discrete model outputs) makes it possible to use a fine discretisation in time, i.e. a large number of model outputs. The covariance matrix C_{ww} is left unconstrained.

As a distance measure between $P(w, v, \lambda, \gamma|D)$ and $P^*(w, v, \lambda, \gamma)$ we use the Kullback-Leibler divergence

$$KL[Q, R, S] = \int dw dv d\lambda d\gamma \ Q(w, v)R(\lambda)S(\gamma) \log \left[\frac{Q(w, v)R(\lambda)S(\gamma)}{P(w, v, \lambda, \gamma|D)}\right] \tag{6}$$

$$\equiv \langle \log [Q(w, v)R(\lambda)S(\gamma)] \rangle_{Q,R,S} - \langle \log [P(D|w, v)P(w|\lambda)P(v|\gamma)P(\lambda)P(\gamma)] \rangle_{Q,R,S} \ ,$$

where in the second line we substituted Equation (4). The goal is to find the normal distribution $Q(w, v)$ and additional distributions $R(\lambda)$ and $S(\gamma)$ that minimise this distance.

The Kullback-Leibler divergence depends both on the choice of parameters $\{\hat{w}, \hat{v}, C\}$ and on the distributions $R(\lambda)$ and $S(\gamma)$. Let us first suppose that $R(\lambda)$ and $S(\gamma)$ are given. The terms in (6) depending on $Q(w, v)$ and thus on the variational parameters $\{\hat{w}, \hat{v}, C\}$ are

$$KL[Q] = \langle \log Q(w, v) \rangle_Q - \langle \log P(D|w, v) \rangle_Q - \langle \log P(w|\lambda) \rangle_{Q,R} - \langle \log P(v|\gamma) \rangle_{Q,S} \ .$$

Except for the second term involving the data likelihood, all terms are straightforward. Neglecting irrelevant constants we have

$$\langle \log Q(w, v) \rangle_Q = -\frac{1}{2} \log \det C \ \text{ and } \ \langle \log P(w|\lambda) \rangle_{Q,R} = -\frac{\bar{\lambda}}{2} \left[\hat{w}^T \Lambda \hat{w} + \text{Tr } (C_{ww}\Lambda)\right] \ ,$$

with $\bar{\lambda} \equiv \langle \lambda \rangle_R$, and a similar expression for $\langle \log P(v|\gamma) \rangle_{Q,S}$. The likelihood term $\langle \log P(D|w, v) \rangle_Q$ can again be decomposed into two terms (see (3)): a term involving only uncensored patients and a term to which all patients contribute. Both contributions can be computed analytically. The uncensored patients ν yield

$$\left\langle \log \left[e^{v_k} e^{w^T x^\nu} \right] \right\rangle_Q = \hat{v}_k + \hat{w}^T x^\nu, t_{k-1} < t^\nu \leq t_k \ , \tag{7}$$

and all patients μ contribute terms of the form

$$-\left\langle e^{v_i} e^{w^T x^\mu} \right\rangle_Q = -\exp\left[\hat{w}^T x^\mu + \hat{v}_i + \frac{1}{2} x^{\mu T} C_{ww} x^\mu + \frac{1}{2} C_{v_i v_i}\right] \ , \tag{8}$$

where we have used the equality

$$\int dy \ P(y) e^{y^T z} = \exp\left[m^T z + \frac{1}{2} z^T \Sigma^2 z\right] \quad \text{with} \quad P(y) = \mathcal{N}(m, \Sigma^2) \ ,$$

in which we substitute $\{w, v\}$ for y and the vector $\{x, [\ldots, 0, 0, 1, 0, 0, \ldots]\}$, with 1 at the position of i, for z.

Summarising, given the values for $\bar{\lambda}$ and $\bar{\gamma}$, the variational parameters $\{\hat{w}, \hat{v}, C\}$ of $Q(w, v)$ can be found by minimising the error function $KL[Q]$, for example using a conjugate gradient method.

Now suppose that $Q(w, v)$ is known and we would like to optimise for $R(\lambda)$. The terms in (6) which depend on $R(\lambda)$ are

$$KL[R] = \langle \log R(\lambda) \rangle_R - \langle \log P(\lambda) \rangle_R - \langle \log P(w|\lambda) \rangle_{Q,R} \ .$$

It is easy to show that, with a gaussian prior $P(w|\lambda)$ and a gamma distribution for $P(\lambda)$, the optimal $R(\lambda)$ is also gamma distributed (see e.g. [12] for details). The procedure for $S(\gamma)$ is completely equivalent.

The approximate posterior $P^*(w, v, \lambda, \gamma)$ can now be found by iterating the following two steps.

- Minimise $KL[R, S]$ to find $R(\lambda)$ and $S(\gamma)$ and calculate the expectation values $\bar{\lambda}$ and $\bar{\gamma}$ given these distributions.
- Substitute $\bar{\lambda}$ and $\bar{\gamma}$ in $KL[Q]$ and minimise this expression.

Since both steps decrease the value of the total divergence $KL[Q, R, S]$, this iterative procedure will converge to an (at least locally) optimal distribution $P^*(w, v, \lambda, \gamma)$. This is actually similar to the HMCMC sampling procedure described in Section 6.1, where now instead of sampling from $P(\lambda, \gamma|\bar{w}, \bar{v}, D)$ and $P(w, v|\bar{\lambda}, \bar{\gamma}, D)$, we obtain analytical expressions for both distributions (which, for $P(\lambda, \gamma|w, v, D)$ is merely a matter of writing down the correct expressions) and calculate the expectation values for $\{\lambda, \gamma\}$ and $\{w, v\}$ respectively.

Note that, due to the exponential transfer functions and our particular choice of parametrisation and corresponding priors, all integrals in $KL[Q, R, S]$ can be done analytically. Therefore, although the specification of the terms of the Kullback-Leibler divergence may look rather complicated, the actual approximation consists of no more than a simple minimisation process. This makes the application of ensemble learning to survival analysis especially attractive, in contrast with applications to neural networks with sigmoidal transfer functions [12], where numerical evaluations are unavoidable. However, with more than one hidden unit either the term (7) involving only uncensored patients, or, with a different parametrisation and choice of priors, the contributions (8) involving all patients would require numerical integration scaling with the number of added hidden units.

6.4. Laplace approximation

The procedure of Section 6.3 can be simplified by replacing the minimisation of the Kullback-Leibler divergence by a Laplace approximation. We again take a normal distribution $\mathcal{N}(\hat{w}, \hat{v}, C)$ for $Q(w, v)$ with

$$\{\hat{w}, \hat{v}\} = \underset{\{w,v\}}{\operatorname{argmax}} P(w, v|\lambda, \gamma, D)$$

and C the Hessian of $-\log(P(w, v|\lambda, \gamma, D))$. As in Section 6.1 λ and γ are kept fixed on the current $\bar{\lambda}$ and $\bar{\gamma}$. Note that in this case no restrictions are imposed on C. The other parts of this approximation are equal to the variational procedure described in Section 6.3. This method to approximate the posterior corresponds to the "evidence framework" introduced in [19].

7. Predictive qualities: an evaluation

7.1. Model comparison

The approximations of the posterior, described in Section 6, enable us to demonstrate the improvement that is gained by the introduction of Bayesian priors. Since we have proposed three different methods to approximate the posterior, we will not only compare the Bayesian approach to survival analysis to the ML Cox method, but we will also find out which of the three approximations yields the best predictions.

To test the different types of models in this article we employ a database of 929 ovarian cancer patients of whom, apart from their medical information at the time of entrance to the study, either their time of death is known, or the last date that they were observed to be alive. For test purposes this database is randomly divided into a training set and a test set. To estimate the predictive qualities of the survival model obtained from the approximations of the posterior as defined by the two priors and the likelihood based on the data in the training set (Equation (4)) we compare it to the maximum likelihood fit of the Cox model on the same set. As an error measure we take

$$\mathcal{E} = -\log \mathcal{L}(D_v), \tag{9}$$

where $\mathcal{L}(D_v)$ is the likelihood of the data in the validation set, estimated either by the maximum likelihood Cox model or the solution obtained by Bayesian methods. The likelihood estimated by the ML solution is calculated simply by inserting $\{w^{\mathrm{ML}}, v^{\mathrm{ML}}\}$ for the model parameters $\{w, v\}$. In the Bayesian approach we use

$$\mathcal{L}(D_v) = \prod_{\nu \in \text{uncensored}} \langle f(t^\nu; x^\nu) \rangle \prod_{\mu \in \text{censored}} \langle F(t^\mu; x^\mu) \rangle \ ,$$

where $\langle \ldots \rangle$ denotes the expectation value over the posterior. In the sampling approach this expectation value is calculated through averaging over all samples after initialisation (see Section 6.2). In both the variational approach and the Laplace approximation ideally we would use the obtained normal distributions to calculate the expectation value analytically. However, due to the specific form of the likelihood this is not possible. Instead we sample from the normal distributions and take the average over the collected samples, just as in the HMCMC approach.

To get a clear indication of the relative strengths of both methods we also compute the ML solution on the test set, and use (9) to calculate a "minimum error" $\mathcal{E}_{\min}$. The relative error used in our comparisons now reads

$$\mathcal{E}_{\mathrm{rel}} = \frac{\mathcal{E} - \mathcal{E}_{\min}}{\mathcal{E}_{\mathrm{cox}} - \mathcal{E}_{\min}}, \tag{10}$$

where $\mathcal{E}_{\mathrm{cox}}$ is the test error obtained from the ML Cox method. So, if one of the Bayesian methods is just as good as the Cox method it would have a relative error of 1.

The results (Figure 4c) show that the errors in any of the approximations to the Bayesian posterior are significantly ($p \approx 1 \times 10^{-5}$) smaller than the error in the classical Cox approach. A closer look at the difference between the three Bayesian approaches reveals that the error in the variational approach is slightly (but significantly) larger than the error in the sampling approach. The Laplace approximation, which takes about as much computation time as the variational approach, does not perform significantly better or worse.

The size of the database that we have access to (929 patients) is much larger than is common in survival analysis. Most medical databases in this field consist of 100-200 patients. To show the effect of the size of the available data on the quality of the model we trained the model on smaller parts (120 and 200 patients) of our database, using the same approaches as in the previous section. The results can be seen in Figures 4a and b, where for comparison we have used the values for $\mathcal{E}_{\rm cox}$ and $\mathcal{E}_{\rm min}$ obtained on the training set of 600 patterns to scale the errors corresponding to the smaller databases as well. For lower and lower numbers of training patterns the error in the classical Cox method increases dramatically. The error in the Bayesian approaches also increases slightly, but due to the effect of the sensible priors the Bayesian method is much more stable under decrease of the number of training patterns.

This effect is not very surprising: for the commonly observed database in the field of survival analysis overfitting is a very serious problem due to the small number of patients. Here, enormous improvements can be made by applying Bayesian priors. For larger and larger databases the overfitting problem grows smaller and smaller until finally, in the limit of an infinite number of patients, ML Cox and the Bayesian approach would coincide. Note however that even on a training set of 600 patterns the Bayesian approach yields significantly better results than the ML Cox method.

The comparison made here, with the complete model considering all inputs, may however not be a totally "fair" one. In the medical statistical community it is well-known that Cox analysis with a full set of inputs strongly suffers from overfitting. A standard procedure to deal with the overfitting problem (and thus the error) in Cox analysis is to reduce the number of inputs to the model. Therefore, in Section 8 we will propose a backward elimination procedure, and we will compare the reduced Cox model to the Bayesian approach again in Section 9.

7.2. Non-proportional hazards

In Section 4 we showed that, with some minor adjustments, it is possible to extend the model beyond Cox proportional hazard. To test whether a more complicated model would indeed be more able to fit the data we added one hidden unit (see Figure 1). We found that the overfitting problem already present in the ML Cox model with one hidden unit increased strongly after adding a second hidden unit (doubling the number of free model parameters), yielding a relative error which vastly exceeded the error in the simple model. The error in the sampling approach did not change significantly with the addition of an extra hidden unit. In the remainder of this article we will concentrate on the version with one hidden unit.

7.3. The effect of discretisation

To demonstrate the effect of the discretisation of the baseline hazard we implemented the ML Cox method and the variational approach on network configurations with numbers of outputs ranging between 5 and 1000 (at this point further discretisation was meaningless since no time interval contained more than one patient). The results are shown in Figure 5, where the error (9) is plotted against the number of outputs. For both the ML Cox method and the Bayesian approach at first an improvement due to finer discretisation can be observed, followed by a deterioration due to overfitting on the training data when the number of model (output) parameters grows. For the Bayesian approach the optimal model has a larger number of output parameters since the imposed priors reduce the overfitting problem. From these results it is clear that the model does not suffer from the inaccuracy that is introduced by discretisation.

In fact, a free, continuous model, corresponding to a very large number of outputs in Figure 5, will produce very poor results. Therefore, for the comparison of the models' predictive qualities we use 10 outputs in the ML Cox model and 50 in each of the Bayesian models, corresponding to the observed optima in Figure 5.

8. Bayesian backward elimination

In Section 7 we mentioned that the comparison between ML Cox and the Bayesian approach was not completely fair. Therefore, in this section we propose a method to reduce the number of input parameters. After eliminating "irrelevant" inputs from the model the ML Cox method will be less impaired by overfitting problems. After this improvement, we will be able to make a "fair" comparison between the Bayesian approach and the ML Cox method (Section 9).

In medicine, the calculation of p-values is a well-known tool for statistical analysis. A p-value is defined as follows: consider a null-hypothesis H_0, which is rejected if a certain test statistic T exceeds the critical value T_c. Now, the p-value of a specific measurement t of T is calculated as $P(T > t|H_0)$, the probability of finding a value for T which exceeds t, given that the null-hypothesis is true [20]. The p-value gives an indication of the conflict between the null-hypothesis and the observed data.

To determine the relevance of one of our models input weights, e.g. w_k, we define a null-hypothesis that states that the "true" value of w_k is zero. This hypothesis would be rejected if the measured value of w_k (the maximum likelihood value w_k^{ML} for classical Cox, the expectation value $\langle w_k \rangle$ for the Bayesian approach) would exceed some critical value w_c. After we derive $P(w_k > w_k^{\mathrm{ML}}|w_k = 0)$ from the approximated posterior (where for ML Cox we take a Laplace approximation based on the likelihood function (3)), it is possible to calculate the p-value for each of the inputs of the model.

Upon constructing the posterior and calculating the p-values of the input weights of the model nearly all weights were deemed irrelevant ($p > 0.05$). This does not necessarily mean that we have created a useless model, it merely states that the function of any input can be taken over by the other inputs if it would be removed. This is a common problem in survival analysis: rather than missing a source of data which might be relevant, medical scientists consider a large variety of possibly interesting sources of medical data. However, since the model is trained on a limited database, not only the desired underlying "true" relations between the data are modelled, but also random effects and anomalies present in the database. To discern between relevant and irrelevant parameters, c.q. inputs, we will propose and apply a backward elimination procedure.

Backward elimination is a procedure in which one by one all irrelevant model parameters are removed, leaving a completely "meaningful" model. In each step of the procedure the least relevant model parameter is found and eliminated. The procedure should continue as long as the decrease in the error due to over-fitting outweighs the reduced functionality of the model. Note that backward elimination is a suboptimal heuristic: in principle all possible subsets of parameters should be considered. However, it has been shown to give close to optimal results in many cases [21].

Two elements need to be defined for backward elimination: a criterion to decide which of the model parameters should be eliminated in each step and an indicator to say when to stop removing parameters. The first criterion may depend, for example, on the size of the

parameters or on some estimation of the increase in training error. The stopping criterion can be established by cross-validation on a separate validation set, or e.g. by Akaike's Information Criterion. In our case we could use the concept of p-values, removing in each step the parameter with the highest p-value, and stopping when all parameters are relevant ($p < 0.05$). However, many respectable statisticians criticise the concept of p-values ([22],[23]); p-values just do not coincide with the characteristic we are interested in: the probability that a certain model parameter is irrelevant or, in a broader sense, the probability of the model given the observed data. In fact, a p-value generally overstates the evidence against the null-hypothesis and thus gives a wrong impression of the relevance of model parameters to researchers without a broad experience working with p-values. We tend to agree with the critics in the field, and therefore use the *Bayes factor* [19, 22] instead. The Bayes factor reads

$$BF = \frac{P(D|H_0)}{P(D|H_1)}, \tag{11}$$

where the hypotheses H_0 and H_1 are specified by

$$\begin{aligned} H_0 &: |w_D| < \epsilon, \qquad \epsilon \to 0 \\ H_1 &: |w_D| > \epsilon, \qquad \epsilon \to 0 \end{aligned}$$

with w_D the parameter we consider removing. In other words: the null-hypothesis states that w_D is actually zero and can be removed, the alternative hypothesis H_1 that w_D is relevant and should stay (the mathematically less critical reader can read $w_D = 0$ for $|w_D| < \epsilon$ and $w_D \neq 0$ for $|w_D| > \epsilon$). The Bayes factor is an expression indicating which of both hypotheses best explains the data D.

Writing the Bayes factor as

$$BF = \lim_{\epsilon \to 0} \frac{P(D||w_D| < \epsilon)}{\int_{|w_D|>\epsilon} dw_D \; d\lambda \; P(D|w_D) \; P(w_D|\lambda) \; P(\lambda)} \,, \tag{12}$$

we recognise that, since for $\epsilon \to 0$ the integral over $|w_D| > \epsilon$ is equal to the integral over all w_D, the lower term in (12) yields $P(D)$. Applying Bayes' rule to the upper term, we obtain

$$BF = \lim_{\epsilon \to 0} \frac{P(|w_D| < \epsilon|D)}{P(|w_D| < \epsilon)} \,. \tag{13}$$

The Bayes factor (calculated explicitly in Appendix I) provides us with a simple backward elimination algorithm:

- Calculate $\frac{P(D|H_0)}{P(D|H_1)}$ for each candidate w_D
- Select the candidate with the highest ratio: if it exceeds one, eliminate the parameter, else stop
- Retrain the reduced network
- Repeat the previous three steps until no irrelevant parameters are left.

Due to the retraining after each step, this can be a time-costly exercise. However, instead of being retrained the posterior can be re-estimated after each parameter removal through the calculation of $P(w_R|w_D = 0, H_0)$ (see Appendix II). Note that this expression is just another

way of writing "the distribution with $\hat{w}_D = 0$ which is closest to the original distribution, $P(w|H_0)$". This method has been implemented in [24] and is closely related to a backward elimination algorithm based on the Bayesian evidence framework proposed by MacKay [19].

Backward elimination in the classical Cox approach can be executed using for example a technique that in the neural network community is referred to as "Optimal Brain Surgeon" [25]. Here we express the difference between two models with parameters $q_1 = \{w_1, v_1\}$ and $q_2 = \{w_2, v_2\}$ as

$$D = \frac{1}{2}(q_1 - q_2)^T H (q_1 - q_2) \; ,$$

where for H we take the Hessian of the log-likelihood, estimated in $\{w^{\mathrm{ML}}, v^{\mathrm{ML}}\}$. In each step of the elimination process, for each parameter w_k we calculate the distance between the current model and the model without w_k that is closest to the original model. The reduced model with the smallest distance D is selected and the corresponding parameter w_k eliminated.

9. Properties of the reduced network

In Section 8 we have defined two backward elimination algorithms, one for the Bayesian approach, one for the ML Cox method. In this section we will apply these algorithms to the model parameters found in Section 7, after which we will make the final comparison between the Bayesian approach and the ML Cox method.

Figure 6 demonstrates the effect of the backward elimination process on the predictive power of the model. It can be seen that this procedure indeed has a wholesome effect: in both the ML Cox method and the Bayesian approach the test error decreases when the least relevant inputs are removed. The decrease of test error in the Cox method is larger than in the variational approach, since in the latter most of the overfitting problem has already been eliminated by the Bayesian priors; in the Cox method it still has to be removed through elimination of irrelevant variables. However, even after reduction the Bayesian approach still yields significantly better results than ML Cox. The effect of reducing the size of the database, described for the complete model in Section 7.1, does not change for the reduced model: for smaller databases the error in the ML Cox model increases significantly, whereas the Bayesian method remains much more stable. The variational approach and the Laplace approximation yield similar results, although the Laplace approximation tends to yield slightly better predictions at some points, including the optimal number of parameters.

It can also be seen (Figure 6, lower panel) that the Bayes factor gives a good estimation of the point where further reduction is no longer desirable: on average, the lowest test error is realised when 4 inputs are left in the network. The Bayes factor drops below one and stops the process when 2-4 inputs are left, which is clearly within the minimal test error range.

Although backward elimination is considered a rather instable procedure (see e.g. [26]), in the Bayesian approach the individual results do not vary dramatically: backward elimination on different training sets generally yielded the same set of remaining parameters. Reducing the size of the database did not have any significant effect on the choice of remaining parameters. The number of times each input remained after the elimination process is indicated in Figure 7. In the Bayesian approach three inputs (patient's performance, leucocytes and the number of tumours after surgery) are found to be extremely relevant. In the ML Cox method the selection of remaining parameters varies strongly between different training sets.

10. Alternative methods

Many other neural approaches to survival analysis have been proposed. One approach is to implement a series of classification problems, one for each discrete point in time [2, 3, 4]. For each of the problems the task of the network is to discriminate between patients who survive upto the corresponding time, and patients who do not. In this setting it may be hard to ensure the consistency of the model (i.e. if a patient is predicted to survive upto some time t a model corresponding to an earlier time should predict the same). Another problem in this approach is the treatment of censored patients.

An alternative to splitting the survival analysis problem into multiple classification problems is to use time as an input [5, 6]. An argument in favour of this method is that time can be treated as a continuous parameter, yielding higher accuracy. We found however (Section 7.3) that a finer discretisation of our hazard has no significant positive effect on the models accuracy. A disadvantage of using time as an input is that time, although it plays a completely different role in survival analysis, is treated on the same footing as explanatory variables such as patient characteristics.

An argument against the use of proportional hazards is that a model with a higher degree of complexity may describe the survival process better. We showed however (Section 7.2) that adding more hidden units had no significant effect on the Bayesian approach (HMCMC) and made the ML Cox method perform worse, even on a large database. These results are in agreement with the earlier findings of [2].

In the existing literature, the weak points of Cox analysis we have considered in this article, overfitting of the proportional hazard and irregularity of the baseline hazard, have been acknowledged. Solutions have been proposed by Biganzoli [5], who added weight decay on the input covariates w, while the problem of irregularity of the outputs has been approached by Liestöl [3], who imposed a penalty term on the difference between the baseline hazard parameters. Another often used approach to solve the latter problem is to use cubic splines methods [7, 8, 9].

11. Discussion

We have demonstrated that survival analysis can benefit strongly from a Bayesian approach, in particular for small data sets which are typically encountered in practice. The resulting method remains as clear and easy to use as the original Cox model, yet at the same time it is more robust and reliable. Of the several approximations of the Bayesian posterior, both the variational approach and the Laplace approximation are more desirable than sampling, because of the convenient properties of the normal distribution, as demonstrated in its application to backward elimination.

We chose to hold on to the Cox proportional hazard model, which is still the standard in the field of survival analysis. The problems that are connected to the proportional hazards method are elegantly solved by embedding proportional hazards in a Bayesian framework. By choosing the right priors, we eliminated the problem of overfitting on the training data, resulting for example in a less irregular baseline hazard and obtained better survival predictions. This improvement was particularly clear for relatively small databases.

Within the Bayesian treatment we reviewed three methods to approximate the posterior:

HMCMC sampling, the variational approach and the Laplace approximation. The results showed that sampling outperformed the other two methods in predictive qualities, while there was no significant difference between variational methods and Laplace. However, HMCMC sampling not only takes more time, but is also cumbersome to use in practice: backward elimination for example, can be done directly when the posterior is described by an approximating normal distribution, yet is not easy to do in the sampling approach. In future research, the recently developed reversible jump Markov chain Monte Carlo method [27] may prove to be a more competitive alternative.

Since it is well known that the Cox proportional hazards method produces poor results with large numbers of inputs, we designed an elegant backward elimination procedure based on the obtained approximations to the Bayesian posterior. Although removal of irrelevant inputs indeed greatly improved the ML Cox method, its predictive qualities still did not exceed those of the Bayesian approach with a full set of inputs. Comparison between the reduced ML Cox model and the reduced Bayesian model again proved the latter to yield significantly better results. Furthermore, the selection of "relevant" inputs was shown to be much more stable under the Bayesian approach than for the ML Cox method.

The Bayesian framework which has been presented in this article provides a solid basis for survival analysis, yet there is still room for further research. For example, although our extension of the model to non-proportional hazards did not yield a significant improvement, another more complicated model might yield different results. Another opening is the fact that on our database the variational approach and the Laplace approximation achieved similar results, even though the variational approach is more sophisticated. It would be interesting to see whether, on a more complicated database, this equality still holds.

APPENDIX

I. Calculation of the Bayes factor

The lower term in (11) can be obtained from the prior on w:

$$\begin{aligned} P(|w_D| < \epsilon | H_0) &= \int d\lambda \; P(|w_D| < \epsilon | \lambda, H_0) P(\lambda | H_0) \\ &= \int d\lambda \int dw_R \int_{\hat{w}_D - \epsilon}^{\hat{w}_D + \epsilon} dw_D \left[\frac{\lambda^n |\Lambda|}{(2\pi)^n} \right]^{1/2} \exp\left[-\frac{\lambda}{2} (w - \tilde{w})^T \Lambda (w - \tilde{w}) \right] \Gamma(\lambda | \sigma, \tau) \\ &= 2\epsilon \left[\frac{|\Lambda|}{(2\pi)^{n_D} |\Lambda_{RR}|} \right]^{1/2} \frac{\tau^\sigma}{\Gamma(\sigma)} \frac{\Gamma(\hat{\sigma})}{\hat{\tau}^{\hat{\sigma}}} \end{aligned}$$

where n_D is the number of removed parameters,

$$\tilde{\Lambda} = (\Lambda_{DD} - \Lambda_{RD}^T \Lambda_{RR}^{-1} \Lambda_{RD})$$

and

$$\tilde{\sigma} = \frac{n_D}{2} + \sigma; \qquad \tilde{\tau} = \frac{1}{2} \tilde{w}_D^T \tilde{\Lambda} \tilde{w}_D + \tau.$$

To calculate the upper term, we use the posterior computed using the variational approach described in Section 6.3:

$$\begin{aligned} P(|w_D| < \epsilon | D, H_0) &= \int d\lambda \int dw_R \int_{\hat{w}_D - \epsilon}^{\hat{w}_D + \epsilon} dw_D P(w, \lambda | D, H_0) \\ &= 2\epsilon \left[\frac{|Q^{-1}|}{(2\pi)_D^n |Q_{RR}^{-1}|} \right] \exp(-\frac{1}{2} \hat{w}_D^T (Q_{DD} - Q_{RD}^T Q_{RR}^{-1} Q_{RD}) \hat{w}_D) \end{aligned}$$

II. Recalculation of the posterior

The posterior probability of the remaining parameters w_R after deletion of the parameters w_D reads:

$$\begin{aligned} P(w_R | w_D = 0, H_0) &= \int d\lambda \frac{P(w_R, w_D = 0, \lambda | D)}{P(w_D = 0 | D)} \\ &= \frac{\mathcal{N}(w_R, w_D = 0 | \hat{w}, \hat{v}, \Sigma)}{\int d\lambda P(w_D = 0 | D, \lambda) P(\lambda)} \\ &= \mathcal{N}(w_R | \hat{w}', \hat{v}, \Sigma'), \end{aligned}$$

with

$$\hat{w}' = \hat{w} + (\Sigma_{ww,RR}^{-1})^{-1} \Sigma_{ww,RD}^{-1} \hat{w}_D$$

and

$$\Sigma'_{ww} = (\Sigma_{ww,RR}^{-1})^{-1}, \Sigma'_{vv} = \Sigma_{vv},$$

where

$$\Sigma_{ww}^{-1} = \begin{pmatrix} \Sigma_{ww,DD}^{-1} & \Sigma_{ww,DR}^{-1} \\ \Sigma_{ww,RD}^{-1} & \Sigma_{ww,RR}^{-1} \end{pmatrix}.$$

ACKNOWLEDGEMENT

We would like to thank David Barber for stimulating comments and discussions.

REFERENCES

1. D. Cox and D. Oakes. *Analysis of Survival Data*. Chapman Hall, London, 1984.
2. R. Ripley, A. Harris, and L. Tarassenko. Neural network models for breast cancer prognosis. *Neural Computing & Applications*, 7:367–375, 1998.
3. K. Liestol, P.K. Andersen, and U. Andersen. Survival analysis and neural nets. *Statistics in Medicine*, 13:1189–1200, 1994.
4. W. Street. A neural network model for prognostic prediction. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 540–546, Madison, Wisconsin, 1998.
5. E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17:1169–1186, 1998.
6. M. De Laurentiis and P.M. Ravdin. A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Letters*, 77:127–138, 1994.
7. K. Hess. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine*, 13:1045–1062, 1994.
8. N. Intrator and C. Kooperberg. Trees and splines in survival analysis. *Statistical Methods in Medical Research*, 4:237–261, 1995.
9. J. Herndon and F. Harrell. The restricted cubic spline as baseline hazard in the proportional hazards model with step-function time-dependent covariates. *Statistics in Medicine*, 14:2119–2129, 1995.
10. R. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, New York, 1996.
11. G. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the 6th Annual Workshop on Computational Learning Theory*, pages 5–13, New York, 1993.
12. D. Barber and C. Bishop. Ensemble learning for multi-layer networks. In *Advances in Neural Information Processing Systems 10*, pages 395–401, Cambridge, 1997. MIT Press.
13. D. Barber and B. Schottky. Radial Basis Functions: a Bayesian treatment. In *Advances in Neural Information Processing Systems 10*, pages 402–408, Cambridge, 1997. MIT Press.
14. D. Altman and P.K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771–783, 1989.
15. P. Verweij and H. van Houwelingen. Penalized likelihood in Cox regression. *Statistics in Medicine*, 13:2427–2436, 1994.
16. J. van Houwelingen. Predictability of the survival of patients with advanced ovarian cancer. *Journal of Clinical Oncology*, 7:769–773, 1989.
17. M. Goldstein and A. Smith. Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society*, 36:284–291, 1974.
18. D. MacKay and M. Jordan(ed.). *Learning in Graphical Models*. Kluwer Academic Publishers, The Netherlands, 1998.
19. D. MacKay. Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network*, 6:469–505, 1995.
20. H. Migon and D. Gamerman. *Statistical Inference: an Integrated Approach*. Arnold, London, 1999.
21. J. Bjornstad and R. Butler. The equivalence of backward elimination and multiple comparisons. *Journal of the American Statistical Association*, 83:136–144, 1988.
22. J. Berger and M. Delampady. Testing precise hypotheses. *Statistical Science*, 2:317–352, 1987.
23. H. Jeffreys. *Theory of Probability*. Oxford University Press, London, third edition, 1967.
24. P. van de Laar and T. Heskes. Pruning using parameter and neuronal metrics. *Neural Computation*, 11(4), 1999.
25. B. Hassibi and D. Stork. Second order derivatives for network pruning: optimal brain surgeon. In S. Hanson, J. Cowan, and L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 164–171, San Mateo, 1993. Morgan Kaufmann.
26. L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24:2350–2383, 1996.
27. P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

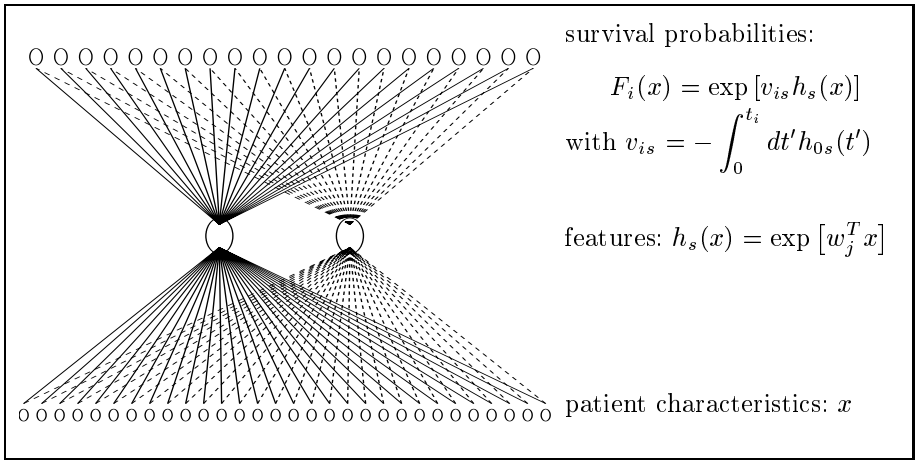

Figure 1. Neural interpretation of survival analysis.

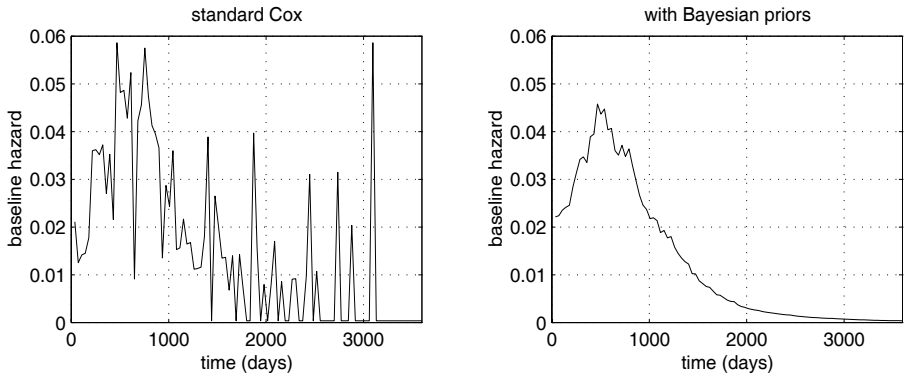

Figure 2. The baseline hazard $h_0(t)$ as obtained in classical Cox analysis (left), and after imposing a Bayesian prior on the parameters v (right). In both cases h_0 is optimised on a training set of 600 patterns, chosen randomly from our database. The effect of the prior is a considerable smoothing of the hazard function.

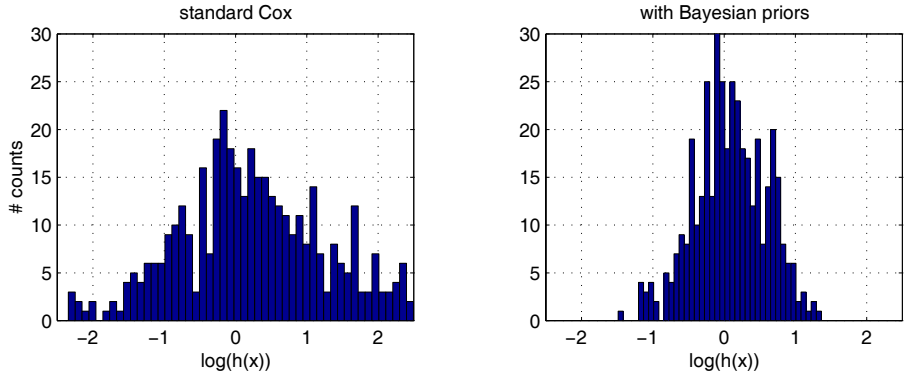

Figure 3. Histograms of the (log of the) proportional hazard $h(x)$ as obtained in classical Cox analysis (left), and after imposing a Bayesian prior on the parameters w (right). In both cases the parameters w of the proportional hazard are optimised on a training set of 600 patterns, chosen randomly from our database. The effect of the prior is a considerable "shrinking" of the proportional hazard.

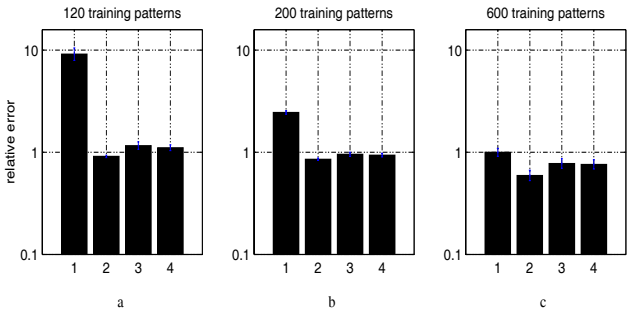

Figure 4. Relative error after training on three partitions of the database (120 patients in the left panel, 200 patients in the middle panel and 600 patients in the right panel). In each panel, from left to right the bars represent the error in: the ML Cox method(1), the HMCMC sampling approach(2), the variational approach(3), the Laplace approximation(4), all with one hidden unit, and the ML Cox method(5) and HMCMC sampling(6) with two hidden units. All differences are significant, except for the one between the variational approach and the Laplace approximation.

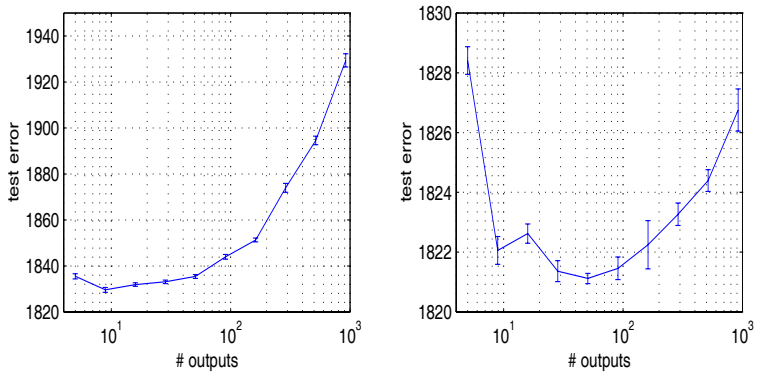

Figure 5. Test error as a function of the number of discrete outputs for both the ML Cox method (left panel) and the variational approach (right panel). The errors are the average of the test error obtained from 25 random draws of 600 patients from the database. For each draw we first selected the relevant inputs to the model, as will be described in Sections 8 and 9, and then varied the number of outputs.

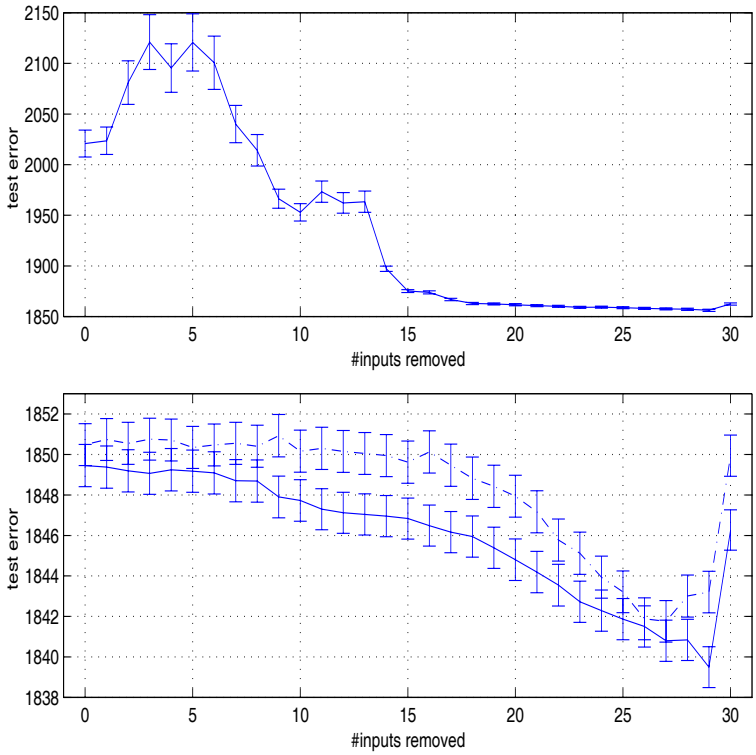

Figure 6. The test error (minus the log-likelihood of the data in the test set under the current model) as a function of the number of removed input parameters for the maximum likelihood Cox model (upper), the variational approach (lower, solid line) and the Laplace approximation (lower, dotted line). At zero, thirty one inputs are left in the model, at thirty, just one. The test error is the average error over 25 sessions, conducted on parameters obtained from random choices of training sets, each containing 600 patterns.

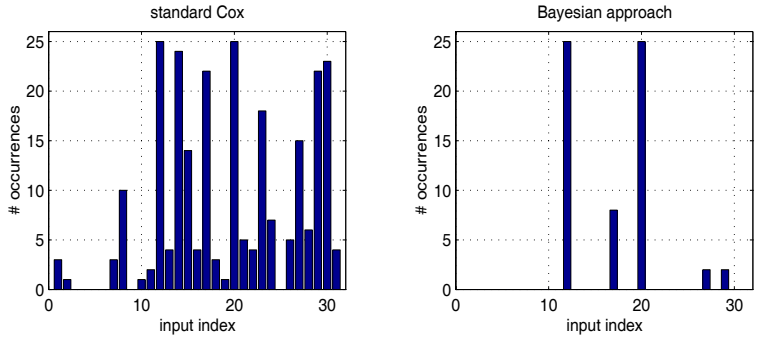

Figure 7. For the Bayesian approach (right panel), in each of the 25 model reduction sessions 2-4 inputs were left in the model. Two of the inputs (12 and 20) appear in almost all of the results. Inputs 17, 27 and 29 are less strong, but still clearly present in the results. For the ML Cox method (left panel), the reduction process is considerably less stable, resulting in a more noisy selection of remaining parameters.