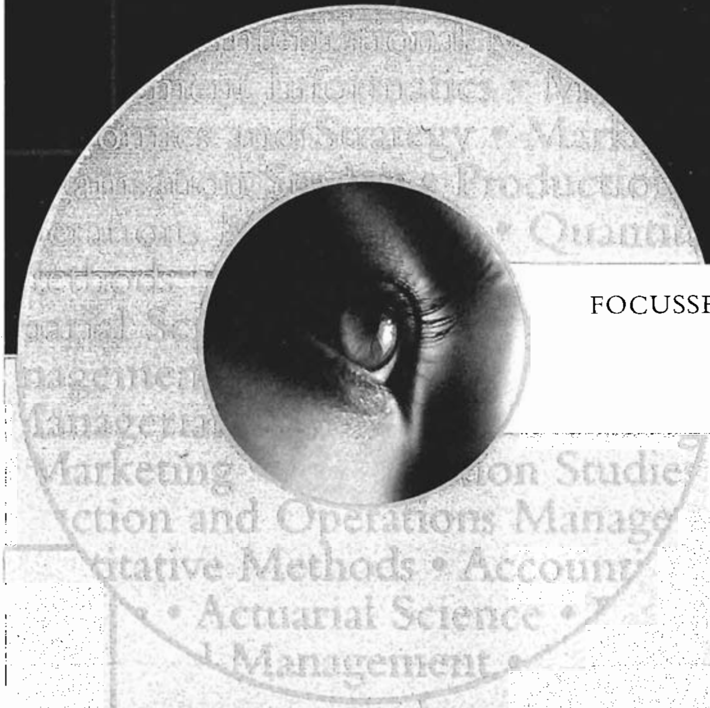
DEPARTMENT OF APPLIED ECONOMICS



RESEARCH REPORT

FOCUSSED INFORMATION CRITERIA AND MODEL AVERAGING FOR COX'S HAZARD REGRESSION MODEL

Nils Lid Hjort • Gerda Claeskens

OR 0458

Focussed information criteria and model averaging for Cox's hazard regression model

Nils Lid Hjort and Gerda Claeskens

University of Oslo and Katholieke Universiteit Leuven

November 2004

Abstract. This article is concerned with variable selection methods for the proportional hazards regression model. Including too many covariates causes extra variability and inflated confidence intervals for regression parameters, so regimes for discarding the less informative ones are needed. Our framework has p covariates designated as 'protected' while variables from a further set of q covariates are examined for possible in- or exclusion. In addition to deriving results for the AIC method, defined via the partial likelihood, we develop a focussed information criterion that for given interest parameter finds the best subset of covariates. Thus the FIC might find that the best model for predicting median survival time might be different from the best model for estimating survival probabilities, and the best overall model for analysing survival for men might not be the same as the best overall model for analysing survival for women. We also develop methodology for model averaging, where the final estimate of a quantity is a weighted average of estimates computed for a range of submodels. Our methods are illustrated in simulations and for a survival study of Danish skin cancer patients.

Key words: *Akaike's information criterion, covariate selection, Cox regression, focussed information criteria, median survival time, model averaging*

1. Introduction and summary

Suppose survival data of the form $(t_i, \delta_i, x_i, z_i)$ are recorded for n individuals, where t_i is life-time, possibly censored, δ_i is an indicator for non-censoring, x_i contains say p covariates that are deemed necessary in the regression model, while z_i has say q further covariates potentially worthy of inclusion. The most popular model for such data is the Cox model of proportional hazards, where the hazard rate for individual i is expressed as

$$h_i(u) = h_0(u)\exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma) \quad \text{for } i = 1, \ldots, n. \tag{1.1}$$

Here $h_0(u)$ is assumed to be continuous and positive over the range of life-times of interest, but is otherwise not specified. This makes the model partly parametric and partly nonparametric. Inference about (β, γ) typically proceeds using the well-known partial likelihood $L_n(\beta, \gamma)$, properly defined in Section 2.

This article is concerned with developing methods for selecting the in some sense best covariates $z_{i,j}$ among the q. The argument against simply including all of them is that this may cause too much estimation variability, leading to inflated confidence intervals and less powerful tests. On the other hand including too few covariates could mean serious modelling bias and missing important explanatory features in the analysis. Thus selecting 'the best' set is a statistical balancing act between bias and variance.

1.1. The Danish malignant melanoma study. For an illustration we study the skin cancer survival analysis data set that is described and analysed extensively in Andersen, Borgan, Gill and Keiding (1993) and elsewhere. In this Danish study, 205 patients with malignant melanoma had radical removal surgery and were followed after operation over the time period 1962–1977. Several covariate variables are of potential interest for studying survival chances, including

x_1, indicator for sex of the patient (woman = 1, man = 2);
z_1, thickness of the tumour, more precisely $z_1 = (z_1^0 - 292)/100$ where z_1^0 is the real thickness, in 1/100 mm, with the average value 292 subtracted out;
z_2, infection infiltration level, a measure of resistance against the tumour, from high resistance 1 down to low resistance 4);
z_3, presence indicator of so-called epithelioid cells (present = 1, non-present = 2);
z_4, ulceration presence (present = 1, non-present = 2);
z_5, invasion depth (at levels 1, 2, 3); and
z_6, age of the patient at the operation (in years).

Patients dead of other causes or still alive in 1977 are treated as censored observations. Among the findings in Andersen et al. (1993) were that men tend to have higher hazard than women. That is why we designate x_1 as 'protected' here, and look for 'the best' covariates to keep among $z_1, \ldots, z_6$. Data rows for the first five and the last five of the 205 are as follows (we have sorted the 205 rows by increasing life-times). Here ε_i is 1 if dead from the illness, 2 if censored, and 4 if dead from other reasons, so that $\delta_i = I\{\varepsilon_i = 1\}$.

	t_i	ε_i	x_1	z_1	z_2	z_3	z_4	z_5	z_6
1	10	4	2	3.84	3	2	1	2	76
2	30	4	2	-2.27	1	1	2	1	56
3	35	2	2	-1.58	3	1	2	2	41
4	99	4	1	-0.02	3	1	2	1	71
5	185	1	2	9.16	3	2	1	3	52
...									
201	4492	2	2	4.14	4	2	1	3	29
202	4668	2	1	3.20	3	2	2	3	40
203	4688	2	1	-2.44	2	2	2	1	42
204	4926	2	1	-0.66	2	1	2	1	50
205	5565	2	1	-0.02	3	1	2	2	41

TABLE 1.1. *The first five and the last five rows from the Danish malignant melanoma survival data set, with life-times t_i (in days), censoring indicator ε_i, and covariates $x_1, z_1, \ldots, z_6$ as described above.*

1.2. Some selection methods. There are rather few well-developed variable selection methods for the Cox model. Methods involving pre-testing of coefficients and variants of backward and forward regression can be put forward, in partial analogy with linear or generalised linear regression theory; we know of no serious study of the performance of such methods in the Cox model context, however. The general model averaging theory we develop in Section 8 below will actually accurately describe the performance of such methods. Fan and Li (2002) propose a penalised version of the log-partial likelihood, with a penalty called the smoothly clipped absolute deviation. This penalty depends on two unknown parameters where the first is fixed at a pre-determined value while the second is chosen via an approximation to generalised cross-validation. Tibshirani (1997) uses the lasso method for variable selection in the Cox model; this and similar L_1 based methods refined later in Efron, Hastie, Johnstone and Tibshirani (2004) are of particular value when the number q of non-protected covariates is large. Bunea and McKeague (2004) also introduce a penalised partial likelihood, where now the penalty depends on both the number of parameters in the parametric part of the model and on the number of components in the sieve construction to estimate the unknown baseline hazard function.

More traditional model selection methods such as AIC and BIC are not automatically defined for Cox models, since there is no workable full likelihood for data. One may however choose to use the partial likelihoods, say $L_{n,S}$ for the model that only uses covariates $z_{i,j}$ for $j \in S$, which leads to

$$\begin{aligned} \mathrm{AIC}_{n,S} &= 2\log L_{n,S}(\widehat\beta_S, \widehat\gamma_S) - 2(p + |S|), \\ \mathrm{BIC}_{n,S} &= 2\log L_{n,S}(\widehat\beta_S, \widehat\gamma_S) - (p + |S|)\log n, \end{aligned} \tag{1.2}$$

in terms of the Cox estimators $(\widehat\beta_S, \widehat\gamma_S)$ inside the S submodel. Here $|S|$ denotes the number of elements in S, and the model with the highest score is selected. These model selection schemes are easily implemented using software for handling the Cox regression model. Volinsky and Raftery (2000) investigate some aspects of the BIC scheme, including discussion of other penalty factors, along with versions of Bayesian model averaging strategies for the Cox model. Less theory has however been developed for the AIC and BIC methods valid for the Cox model than for fully parametric regression models. It should be noted that these criteria work only with the parametric part of the (1.1) model, thus ignoring the nonparametric part. It is therefore not clear whether model selectors using (1.2) are relevant when it comes to consequences for questions that relate to all of the (1.1) model, like survival probabilities and median survival time. For the melanoma data set, at any rate, the AIC method selects variables 2, 3, 4, 5, 6 among the $z_{i,j}$s while the BIC regime chooses variables 4, 5.

1.3. Focussed information criteria and model averaging. Different selection method find different 'best subsets', as dramatically witnessed for the Danish melanoma data set with the AIC and the BIC. What all the methods mentioned above have in common is however that they advocate one and only one final model, regardless of its intended use, whether this involves predicting median survival time or estimating survival chances for patients with unusual characteristics and so on. We shall develop a certain 'focussed information criterion', the FIC, that for each parameter of interest finds the best submodel for that purpose. Specifically, for a given focus parameter $\mu(\beta, \gamma, H_0)$, where H_0 is the cumulative baseline hazard rate, we are able to estimate the mean squared error for each of the many candidate estimators, say $\widehat{\mu}_S$ for the model indexed by subset S. The FIC strategy is to select the model with lowest possible mean squared error estimate.

We do not view this as a paradox, even if it means leaving behind the traditionally strong and conceptually sirening paradigm of finding one adequate model to explain all aspects and facets of the data. Thus for the Danish skin cancer data we shall see in Section 9 that when it comes to estimating a certain relative risk parameter, then the FIC selects the narrow model as the best one, with only x_1 and none of the z_js; while for estimating a certain survival probability, FIC chooses to include z_5.

When a model selection scheme like the AIC or FIC is followed to produce an estimator it is important to realise that the real variance involved is larger than if the selected model had been given in advance. Studying statistical properties of such post-selection estimators involves more work than simply understanding the limit distributions of the Cox estimators. In our article we reach precise large-sample results for a broad class of 'compromise estimators' that interpolate between all candidate models, with the post-selection estimators constituting special cases. The limit distributions involved are not normal, but rather non-linear mixtures of different normals.

The theory and results for our FIC and model average estimators parallel development and findings in our earlier articles Claeskens and Hjort (2003) and Hjort and Claeskens (2003a, 2003b), hereafter referred to as CH and HC (2003). These articles were concerned with likelihood methods for general parametric models, including various regression models, but the results reached there do not capture and cannot be applied directly inside the Cox regression model. This is partly because of the censoring and of the semiparametric nature of model (1.1), rendering analysis of estimators that combine both H_0 and (β, γ) estimators difficult. Thus a separate development for building a proper FIC along with proper model average methods for the Cox model has been necessary.

We learn in doing so that the CH and HC (2003) theory and methods carry over with reasonable ease to situations which involve only the regression parameters (β, γ); in other words, as long as questions are posed that can be answered in terms of (β, γ), one does not need significant extensions of the already available theory. This comment will be seen to apply also to results for the AIC strategy. Many questions of interest relate however to the full (1.1) model, including the hazard rate part, like the median survival time $H_0^{-1}(\log 2/\exp(x^{\rm t}\beta + z^{\rm t}\gamma))$, the survival probability $\exp\{-\exp(x^{\rm t}\beta + z^{\rm t}\gamma)H_0(t)\}$, likewise conditional survival probabilities given than one has survived up to a certain time point, etc. It is for such questions that the CH and HC (2003) theory needs harder work to be appropriately extended, as seen in the sections to follow.

1.4. The present article. Our article is organised as follows. Section 2 sets the basic framework, properly defining all subset estimators $\widehat{\mu}_S$ for a given focus parameter $\mu(H_0, \beta, \gamma)$, and provides a local neighbourhood formulation that turns out to give fruitful large-sample approximations to modelling bias, variance, and distributions of estimators. In Sections 3 and 4 we develop theory for describing the behaviour of submodel-based estimators $\widehat\beta_S$ and $\widehat\gamma_S$ for the regression coefficients and $\widehat H_{0,S}$ for the cumulative hazard. This is used in Section 5 to provide precise large-sample results for limit distributions and limiting risk for all submodel estimators. This is a harder task than proving limit theorems for the Cox estimators, in that these need to be studied also outside model conditions and since we need to care about simultaneous aspects of all the estimators involved.

In Section 6 we go through a list of particularly important parameters of interest, including survival probability curves for given strata of patients, median survival times, and relative risks. This is also where we describe how to estimate various quantities necessary for implementing the FIC methods. Section 7 gives the proper machinery for the FIC and its averaged versions. Then in Section 8 a 'master theorem' is provided that accurately describes the limit distribution for a large class of model average estimators. This in particular provides precise descriptions of the large-sample behaviour of all post-selection strategies, like the AIC and the FIC. Section 9 illustrates our methods in some settings with simulated data and goes on to analyse the Danish melanoma survival data set. Our article ends with a list of concluding remarks in Section 10, some of which might lead to further research work, and with Section 11, where we gather all proofs of lemmas from earlier sections.

2. A framework for covariate subset selection

Working inside the (1.1) model, with life-time data observed or partly observed over a time horizon $[0,\tau]$, the log-partial likelihood can be written

$$\log L_n(\beta,\gamma)=\sum_{i=1}^n\int_0^\tau\Big[x_i^{\rm t}\beta+z_i^{\rm t}\gamma-\log\Big\{\sum_{i=1}^n Y_i(u)\exp(x_i^{\rm t}\beta+z_i^{\rm t}\gamma)\Big\}\Big]\,{\rm d}N_i(u), \qquad (2.1)$$

where $Y_i(u)=I\{t_i\ge u\}$ and ${\rm d}N_i(u)=I\{t_i\in[u,u+{\rm d}u],\delta_i=1\}$. The maximum partial likelihood estimators, also referred to as the Cox estimators, are the $(\widehat\beta,\widehat\gamma)$ values maximising (2.1). The theory to be developed in our article will partly utilise the large-sample theory associated with counting processes and martingales, as exposited e.g. in Andersen et al. (1993), where results are most comfortably reached if the upper time limit τ is finite, so we shall assume it to be so; somewhat more technical assumptions are needed if one wishes to obtain results valid for $\tau=\infty$.

For each subset S of $\{1,\dots,q\}$ we may study the model indexed by (β,γ_S), where γ_S contains precisely those γ_j coefficients where $j\in S$, i.e. corresponding to including in the model only those $z_{i,j}$ where $j\in S$, excluding those with $j\notin S$. There is a total of 2^q such submodels to consider. Sometimes several of these might be ruled out on a priori grounds, e.g. when there is a natural ordering in complexity, in which case only the $q+1$ nested submodels of $\{1,\dots,q\}$ are considered. For each submodel S there are

Cox estimators $(\widehat\beta_S,\widehat\gamma_S)$, and also an accompanying Aalen–Breslow type estimator of the cumulative baseline hazard function, namely

$$\widehat H_{0,S}(t)=\int_0^t\frac{\sum_{i=1}^n \mathrm{d}N_i(u)}{\sum_{i=1}^n Y_i(u)\exp(x_i^{\mathrm t}\widehat\beta_S+z_{i,S}^{\mathrm t}\widehat\gamma_S)},\tag{2.2}$$

where $z_{i,S}$ means the components $z_{i,j}$ of z_i for which $j\in S$. For given estimand of interest, say $\mu=\mu(\beta,\gamma,H_0)$, there is accordingly a list of potential estimators

$$\widehat\mu_S=\mu(\widehat\beta_S,\widehat\gamma_S,0_{S^c},\widehat H_{0,S}),\tag{2.3}$$

one for each submodel. The notation indicates that one uses the null value $\gamma_j=0$ for $j\notin S$, i.e. for $j\in S^c$, the complement set. The μ in question may also depend on covariate positions x and z, as exemplified in Sections 6 and 9.

We shall study questions of covariate inclusion and exclusion inside a large-sample framework where γ is small or moderate, and where the largest of the models, the one containing all $p+q$ covariates, contains the truth. More specifically, the real hazard rate functions are taken to be

$$h_{i,\mathrm{true}}(u)=h_0(u)\exp(x_i^{\mathrm t}\beta+z_i^{\mathrm t}\eta/\sqrt n)\quad\text{for } i=1,\ldots,n,\tag{2.4}$$

for suitable $(\beta_1,\ldots,\beta_p)^{\mathrm t}$ and $(\eta_1,\ldots,\eta_q)^{\mathrm t}$. This turns out to be a fruitful framework for deriving accurate approximations to modelling bias and variances and hence mean squared errors for different estimators, essentially because variances and squared biases now become exchangeable currencies, both of order $O(1/n)$. See also the general discussion surrounding these issues in CH and HC (2003).

3. Submodel estimators for hazard regression coefficients

This section develops theory for the large-sample behaviour of all submodel estimators $(\widehat\beta_S,\widehat\gamma_S)$. This requires some modifications and extensions of standard theory for the Cox regression models, in that we need to analyse estimators for models that are perhaps approximately but not fully correct. We also need an apparatus for handling estimators from different submodels simultaneously. See Andersen and Gill (1982), Gill (1984) and Andersen et al. (1993) for such 'standard theory'.

To properly analyse the submodel estimators we need to introduce certain random quantities and their limit functions. Let

$$\begin{aligned}
G_n^{(0)}(u,\beta,\gamma)&=n^{-1}\sum_{i=1}^n Y_i(u)\exp(x_i^{\mathrm t}\beta+z_i^{\mathrm t}\gamma),\\
G_n^{(1)}(u,\beta,\gamma)&=n^{-1}\sum_{i=1}^n Y_i(u)\exp(x_i^{\mathrm t}\beta+z_i^{\mathrm t}\gamma)\begin{pmatrix}x_i\\ z_i\end{pmatrix},\\
G_n^{(2)}(u,\beta,\gamma)&=n^{-1}\sum_{i=1}^n Y_i(u)\exp(x_i^{\mathrm t}\beta+z_i^{\mathrm t}\gamma)\begin{pmatrix}x_i\\ z_i\end{pmatrix}\begin{pmatrix}x_i\\ z_i\end{pmatrix}^{\mathrm t},
\end{aligned}$$

along with

$$E_n(u,\beta,\gamma) = G_n^{(1)}(u,\beta,\gamma)/G_n^{(0)}(u,\beta,\gamma).$$

We shall also need the sub-functions associated with subset S being used instead of the full $\{1,\ldots,q\}$; thus $G_{n,S}^{(1)}$ is the $p+|S|$-vector where the first p components make up $G_{n,0}^{(1)}$ and the next $|S|$ components define $G_{n,1,S}^{(1)}$, and similarly with the ratio $E_{n,S}$ which has p components giving $E_{n,0}$ and then $|S|$ components defining $E_{n,1,S}$.

As is commonly assumed in treatises on the Cox regression model, we postulate that these functions have limits in probability $g^{(0)}(s,\beta,\gamma)$, $g^{(1)}(s,\beta,\gamma)$, $g^{(2)}(s,\beta,\gamma)$, and that these limit functions are continuous in s. Actually, since we work under the (2.4) assumption, we are more concerned with the related condition that $G_n^{(0)}(s,\beta,\eta/\sqrt{n}) \to_p g^{(0)}(s,\beta,0)$, and so on. We also write $e(s,\beta,0)$ for the limit function of $G_n^{(1)}(s,\beta,\eta/\sqrt{n})/G_n^{(0)}(s,\beta,\eta/\sqrt{n})$.

Let U_n and V_n be the derivatives with respect to β and γ of the log-likelihood normalised by n^{-1}, so that

$$\begin{pmatrix} U_n(\beta,\gamma) \\ V_n(\beta,\gamma) \end{pmatrix} = n^{-1}\sum_{i=1}^n \int_0^\tau \Big\{\begin{pmatrix} x_i \\ z_i \end{pmatrix} - E_n(u,\beta,\gamma)\Big\}\,\mathrm{d}N_i(u). \tag{3.1}$$

Let also $I_n(\beta,\gamma)$ be the $(p+q)\times(p+q)$ matrix of second order derivatives, leading to

$$-I_n(\beta,\gamma) = \int_0^\tau \Sigma_n(u,\beta,\gamma)G_n^{(0)}(u,\beta,\gamma)h_0(u)\,\mathrm{d}u,$$

in which $\Sigma_n = G_n^{(2)}/G_n^{(0)} - E_nE_n^{\mathrm{t}}$.

Under standard assumptions about the covariate sequences x_i and z_i, and in the framework defined by (2.4), it follows that $-I_n(\beta,\eta/\sqrt{n})$ as well as $-I_n(\beta,0)$ have as limit in probability a $(p+q)\times(p+q)$ matrix

$$J_{\mathrm{full}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}, \tag{3.2}$$

which we also take to be positive definite; see again Andersen et al. (1993) for more details. In fact, also $J_{n,\mathrm{full}} = -I_n(\widehat\beta_{\mathrm{full}},\widehat\gamma_{\mathrm{full}})$ tends in probability to J_{full}, under condition (2.4). The estimator we shall use for J_{full} is

$$\begin{aligned}\widehat J_{\mathrm{full}} &= \int_0^\tau \Sigma_n(u,\widehat\beta_{\mathrm{full}},\widehat\gamma_{\mathrm{full}})G_n^{(0)}(u,\widehat\beta_{\mathrm{full}},\widehat\gamma_{\mathrm{full}})\,\mathrm{d}\widehat H_{0,\mathrm{full}}(u) \\ &= \int_0^\tau \Sigma_n(u,\widehat\beta_{\mathrm{full}},\widehat\gamma_{\mathrm{full}})\,n^{-1}\sum_{i=1}^n \mathrm{d}N_i(u) = n^{-1}\sum_{i=1}^n \Sigma_n(t_i,\widehat\beta_{\mathrm{full}},\widehat\gamma_{\mathrm{full}})\delta_i.\end{aligned} \tag{3.3}$$

We shall also have occasion to need the $(p+|S|)\times(p+|S|)$ submatrix J_S, with blocks say $J_{00}, J_{01,S}, J_{10,S}, J_{11,S}$. It is convenient to phrase some of the results in terms of the projection function $\pi_S\colon \mathcal{R}^q \to \mathcal{R}^{|S|}$ which takes $v=(v_1,\ldots,v_q)^{\mathrm{t}}$ to its subvector v_S with those v_j for which $j\in S$. Thus π_S is an $|S|\times q$ matrix of 1s and 0s.

The following key lemma, with its proof in Section 11, gives the precise large-sample behaviour of the S-submodel Cox estimators, under the (2.4) assumption. We assume that 'ordinary regularity conditions', as spelled out and discussed in Andersen et al. (1993, Ch. VII) are in force; these may actually also be substantially weakened, as discussed in Hjort (1992) and Hjort and Pollard (1996).

Lemma 1. *Assume that the conditions just described are in force. Then, under the sequence of true hazard rate functions (2.4),*

$$\begin{pmatrix} \sqrt{n}(\widehat\beta_S-\beta) \\ \sqrt{n}\widehat\gamma_S \end{pmatrix} \to_d \begin{pmatrix} B_S \\ C_S \end{pmatrix} \sim \mathrm{N}_{p+|S|}(J_S^{-1}\begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix}\eta, J_S^{-1}).$$

Armed with this lemma we may derive useful expressions for the approximate mean squared error of estimators $\mu(\widehat\beta_S,\widehat\gamma_S)$ of estimands of the type $\mu(\beta,\gamma)$. Since we shall take an interest in more general estimands, which also may depend on H_0, further efforts are needed to determine the behaviour of $\widehat H_{0,S}$ estimators.

4. Submodel estimators for the cumulative baseline hazard

Inside a submodel S, which gives maximum partial likelihood estimators $(\widehat\beta_S,\widehat\gamma_S)$ for the (2.1) model, we now study the accompanying Aalen–Breslow type estimator given in (2.2) for the cumulative hazard function $H_0(t)=\int_0^t h_0(u)\,\mathrm{d}u$. To reach a precise result, consider first

$$W_n(t)=n^{-1/2}\int_0^t \frac{\sum_{i=1}^n \mathrm{d}M_i(u)}{G_n^{(0)}(u,\beta,0)}.$$

This is a martingale with variance function converging towards $\int_0^t g^{(0)}(u,\beta,0)^{-1}\,\mathrm{d}H_0(u)$, which implies that the $W_n(.)$ process tends in distribution to a Gaußian zero-mean martingale $W(.)$ with $\mathrm{Var}\,\mathrm{d}W(u)=\mathrm{d}H_0(u)/g^{(0)}(u,\beta,0)$. One also finds that the W_n process and the vector of $\sqrt{n}U_n(\beta,0)$ and $\sqrt{n}V_n(u,\beta,0)$ are independent in the limit, that is, the W process becomes independent of each B_S, C_S of Lemma 1. To see this, work first with $\mathrm{cov}\{\sqrt{n}U_n(\beta,0),\mathrm{d}W_n(u)\}$, which by martingale theory can be expressed as the mean of say S_n, where

$$S_n=n^{-1}\sum_{i=1}^n\{x_i-E_{n,0}(u,\beta,0)\}\frac{Y_i(u)\exp(x_i^{\mathrm{t}}\beta+z_i^{\mathrm{t}}\eta/\sqrt{n})}{G_n^{(0)}(u,\beta,0)},$$

which is seen to be composed of $G_{n,0}^{(1)}(u,\beta,0)-E_{n,0}(u,\beta,0)G_n^{(0)}(u,\beta,0)$, which vanishes, plus a term of order $O_p(n^{-1/2})$. A similar calculation confirms the claim for $\sqrt{n}V_n(u,\beta,0)$ and W_n.

For the next central result, its proof placed in Section 11, let us introduce the $(p+q)$-vector function

$$F(t)=\int_0^t e(u,\beta,0)\,\mathrm{d}H_0(u)=\begin{pmatrix} F_0(t) \\ F_1(t) \end{pmatrix},$$

where the first p components comprise $F_0(t)$ and the final q components make up $F_1(t)$. We also use $F_{1,S}(t)$ to denote the subset of $F_1(t)$ with components belonging to subset S, and finally $F_S(t)$ for the $p+|S|$-vector with $F_0(t)$ and $F_{1,S}(t)$.

Lemma 2. *Under the (2.4) assumptions, along with other conditions stated in connection with Lemma 1, the $A_{n,S}(t) = n^{1/2}\{\widehat H_{0,S}(t) - H_0(t)\}$ process tends in distribution to the process*

$$A_S(t) = W(t) - \begin{pmatrix} F_0(t) \\ F_{1,S}(t) \end{pmatrix}^{\rm t} \begin{pmatrix} B_S \\ C_S \end{pmatrix} + F_1(t)^{\rm t}\eta.$$

Note that

$$\operatorname{Var} A_S(t) = \int_0^t \frac{{\rm d}H_0(u)}{g^{(0)}(u,\beta,0)} + F_S(t)^{\rm t} J_S^{-1} F_S(t),$$

getting larger when more covariates are included. This needs to be weighted against its bias level, which can be read off from Lemmas 1 and 2. An expression for the bias will also flow from the efforts of the next section.

5. Limiting risk of submodel estimators

Consider an estimand of the general type $\mu(\beta,\gamma,H_0(t))$, with t at the moment kept fixed, taken to be smooth in the sense of having continuous derivatives in a neighbourhood of $(\beta,0,H_0(t))$. There is one potential estimator $\widehat\mu_S = \mu(\widehat\beta_S,\widehat\gamma_S,0_{S^c},\widehat H_{0,S}(t))$ for each regressor subset $S \subset \{1,\ldots,q\}$. We shall reach a precise limit distribution result for $\widehat\mu_S$. Our limiting risk results will involve concise expressions for bias and variance in terms of the quantities

$$\Omega_S = \pi_S^{\rm t} K_S \pi_S K^{-1}, \quad \omega = J_{10}J_{00}^{-1}\tfrac{\partial\mu}{\partial\beta} - \tfrac{\partial\mu}{\partial\gamma}, \quad \kappa = \kappa(t) = \{J_{10}J_{00}^{-1}F_0(t) - F_1(t)\}\tfrac{\partial\mu}{\partial H_0}. \tag{5.1}$$

Here $K = J^{11}$ is a $q \times q$ matrix, computed from the inverse of $J_{\rm full}$, while $K_S = J^{11,S} = (\pi_S K^{-1} \pi_S^{\rm t})^{-1}$ similarly is the lower right hand corner $|S| \times |S|$ submatrix of the inverse of J_S. The partial derivatives are evaluated at the centre point $(\beta,0,H_0(t))$; thus both $\frac{\partial\mu}{\partial H_0}$ and $\kappa = \kappa(t)$ depend upon the t under consideration. Note that both ω and κ are of dimension q. Finally define

$$\tau_0^2 = (\tfrac{\partial\mu}{\partial H_0})^2 \int_0^t \frac{{\rm d}H_0(u)}{g^{(0)}(u,\beta,0)} + \{\tfrac{\partial\mu}{\partial\beta} - \tfrac{\partial\mu}{\partial H_0}F_0(t)\}^{\rm t} J_{00}^{-1} \{\tfrac{\partial\mu}{\partial\beta} - \tfrac{\partial\mu}{\partial H_0}F_0(t)\}, \tag{5.2}$$

which will be seen to be the minimal possible limiting variance of the $\widehat\mu_S$ estimators. The model underlying the data is again taken to be that of (2.4), under which $\mu_{\rm true} = \mu(\beta,\eta/\sqrt{n},H_0(t))$.

Lemma 3. *Under conditions laid out for Lemmas 1 and 2, and under circumstances (2.4), the variable $\Lambda_{n,S} = \sqrt{n}(\widehat\mu_S - \mu_{\rm true})$ tends in distribution to*

$$\Lambda_S = (\tfrac{\partial\mu}{\partial\beta})^{\rm t} B_S + (\tfrac{\partial\mu}{\partial\gamma_S})^{\rm t} C_S - (\tfrac{\partial\mu}{\partial\gamma})^{\rm t}\eta + \tfrac{\partial\mu}{\partial H_0} A_S(t). \tag{5.3}$$

This is a normal variable with mean and variance respectively equal to

$$(\omega-\kappa)^{\rm t}(I-\Omega_S)\eta \quad \textit{and} \quad \tau_0^2 + (\omega-\kappa)^{\rm t}\Omega_S K \Omega_S^{\rm t}(\omega-\kappa).$$

Following the details of the proof, which is placed in Section 11, we also establish a quite fruitful representation of the limit distribution, namely

$$\Lambda_S = (\tfrac{\partial\mu}{\partial H_0})W(t) + (\tfrac{\partial\mu}{\partial\beta} - \tfrac{\partial\mu}{\partial H_0}F_0(t))^{\mathrm{t}} J_{00}^{-1} U + (\omega - \kappa)^{\mathrm{t}}\{(I - \Omega_S)\eta - \Omega_S V'\}, \tag{5.5}$$

in which U and V' are independent and respectively $\mathrm{N}_p(0, J_{00})$ and $\mathrm{N}_q(0, K)$. It also follows from this that the mean squared error of the large-sample limit of n times $\widehat{\mu}_S$, i.e. the limiting risk associated with using the S subset, is

$$\mathrm{E}\Lambda_S^2 = \tau_0^2(t) + (\omega - \kappa(t))^{\mathrm{t}}\{(I - \Omega_S)\eta\eta^{\mathrm{t}}(I - \Omega_S^{\mathrm{t}}) + \Omega_S K \Omega_S^{\mathrm{t}}\}(\omega - \kappa(t)), \tag{5.6}$$

allowing our notation here to reflect that both τ_0 and κ depend on the t engaged in the estimand $\mu = \mu(\beta, \gamma, H_0(t))$.

Remark. There is a result corresponding to that of Lemma 3 in CH and HC (2003), valid for general parametric families, but involving only a quantity similar to the ω. It is the semiparametric nature of the Cox regression model that here leads to the more general $\omega - \kappa$ quantity. ■

6. Risk calculation and estimation for important estimands

Note that the limit distributions and limiting risks derived in the previous section depend crucially on both $\tau_0(t)$ and the coefficients of $\omega - \kappa(t)$, which vary from one parameter to the next. This is illustrated now for a brief list of examples, before we turn to the task of estimating these and other quantities involved in the limiting risk expressions. The important case of the median survival time, or more generally the task of estimating the quantile distribution of the survival time, needs some technical development of separate interest, and is treated in Section 6.2.

6.1. A list of foci.

(i) One may naturally compare hazard level for individuals with covariate x with hazard level for those with covariate x_0 using the hazard ratio, say $h(s\,|\,x,z)/h(s\,|\,x_0,z) = \exp\{(x-x_0)^{\mathrm{t}}\beta\}$. This focus parameter has $\omega = \exp\{(x-x_0)^{\mathrm{t}}\beta\}J_{10}J_{00}^{-1}(x-x_0)$ and $\kappa = 0$.

(ii) A natural parameter of interest is the relative risk $\mu = \exp(x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma)$ at position (x,z) in the covariate space; here $\omega = \exp(x^{\mathrm{t}}\beta)(J_{10}J_{00}^{-1}x - z)$ while $\kappa = 0$.

The quantity just discussed can be seen as the relative risk in comparison with an individual with covariates $(x,z) = (0,0)$. This is a natural quantity in situations where the covariates have been centred to have mean zero; in this case, the 'relative' in 'relative risk' would mean in comparison with 'the average individual'. Similarly, if x and z represent risk factors, scaled such that zero level corresponds to normal healthy conditions and positive values correspond to increased risk, then the $\mu = \mu(x,z)$ above is relative risk increase at level (x,z) in comparison with normal health level.

In yet other situations it would be more natural to compare individuals with an existing or hypothesised individual with suitable null-covariates (x_0, z_0), say. This corresponds to focussing on the relative risk $\mu = \exp\{(x - x_0)^{\rm t}\beta + (z - z_0)^{\rm t}\gamma\}$, and leads to $\kappa = 0$ and

$$\omega = \exp\{(x - x_0)^{\rm t}\beta\}\{J_{10}J_{00}^{-1}(x - x_0) - (z - z_0)\}. \tag{6.1}$$

Note in particular that different covariate levels give different ω vectors, which in view of Lemma 3 and risk expression (5.6) means that there might well be different optimal S submodels for different covariate regions. This is accounted for in our focussed information criterion for model selection, as discussed in Section 7.

(iii) For the problem of estimating $H_0(t)$ separately, the ω vector is zero while $\kappa = J_{10}J_{00}^{-1}F_0(t) - F_1(t)$.

(iv) Estimating a survival probability for a given individual translates to

$$\mathrm{Su}(t \mid x, z) = \exp\{-\exp(x^{\rm t}\beta + z^{\rm t}\gamma)H_0(t)\},$$

for which one finds

$$\begin{aligned}\omega &= -\mathrm{Su}(t \mid x, z)H_0(t)(J_{10}J_{00}^{-1}x - z),\\ \kappa &= -\mathrm{Su}(t \mid x, z)\exp(x^{\rm t}\beta)\{J_{10}J_{00}^{-1}F_0(t) - F_1(t)\}.\end{aligned}$$

(v) Consider now a patient's chance of surviving t, given that he has managed to survive up to time t_0. This probability is $\exp[-\{H_0(t) - H_0(s)\}\exp(x^{\rm t}\beta + z^{\rm t}\gamma)]$. Handling this estimand calls for some modifications of Lemmas 2 and 3, in that $\int_{t_0}^t \mathrm{d}H_0(u)$ is at work rather than the full $H_0(t)$. Lemma 2 may be extended to reach parallel results involving $A_S(t) - A_S(t_0)$ rather than simply $A_S(t)$, without serious difficulties. This includes revised definitions of κ and τ_0^2, replacing $F_0(t)$ and $F_1(t)$ with $F_0(t) - F_0(t_0)$ and $F_1(t) - F_1(t_0)$.

6.2. Estimating median survival time. A patient's median survival time, in terms of his covariates, can be expressed as $\xi = H_0^{-1}(\log 2/\exp(x^{\rm t}\beta + z^{\rm t}\gamma))$. That this is a quantity of serious interest, and sometimes more important than say the mean survival time, is made clear in e.g. Gould (1995). Earlier work on conditional median survival time includes Dabrowska and Doksum (1987) and Burr and Doss (1993). Handling the case of such conditional quantiles here, in general

$$\xi(r) = H_0^{-1}(-\log(1 - r)/\exp(x^{\rm t}\beta + z^{\rm t}\gamma)) \quad \text{for } 0 < r < 1,$$

requires some separate development, and it is not a priori clear that the limiting distribution of say

$$\sqrt{n}(\widehat\xi_S - \xi_{\rm true}) = \sqrt{n}\Big\{\widehat H_{0,S}^{-1}\Big(\frac{\log 2}{\exp(x^{\rm t}\widehat\beta_S + z_S^{\rm t}\widehat\gamma_S)}\Big) - H_0^{-1}\Big(\frac{\log 2}{\exp(x^{\rm t}\beta + z^{\rm t}\eta/\sqrt{n})}\Big)\Big\},$$

has the same appealing structure as in Lemma 3 of Section 5, since the $\xi(\beta,\gamma,H_0)$ under consideration now does not only depend on H_0 at a single value.

Consider an estimand of the general form $\xi = H_0^{-1}(f(\beta,\gamma))$, where $f(\beta,\gamma)$ is some smooth function of the regression coefficients, and for which we contemplate using any of the estimates

$$\widehat{\xi}_S = \widehat{H}_{0,S}^{-1}(f(\widehat{\beta}_S,\widehat{\gamma}_S,0_{S^c})) = \sup\{t\colon \widehat{H}_{0,S}(t) \le f(\widehat{\beta}_S,\widehat{\gamma}_S,0_{S^c})\}.$$

The following is proved in Section 11.

Lemma 4. *Assume conditions laid out for Lemmas 1 and 2 are in force. Then, under circumstances (2.4), the variable $\Lambda_{n,S} = \sqrt{n}(\widehat{\xi}_S - \xi_{\rm true})$ tends in distribution to*

$$\Lambda_S = h_0(\xi_0)^{-1}\{-W(\xi_0) + R_0^{\rm t}J_{00}^{-1}U + \zeta^{\rm t}\Omega_S V' - \zeta^{\rm t}(I-\Omega_S)\eta\},$$

with U and V' as in (5.5), where $\xi_0 = H_0^{-1}(f(\beta,0))$, $R_0 = F_0(\xi_0) + \frac{\partial f}{\partial\beta}$, $R_1 = F_1(\xi_0) + \frac{\partial f}{\partial\gamma}$, and $\zeta = R_1 - J_{10}J_{00}^{-1}R_0$, and where the partial derivatives of f are evaluated at $(\beta,0)$. The limit distribution is normal, with mean and variance

$$-h_0(\xi_0)^{-1}\zeta^{\rm t}(I-\Omega_S)\eta \quad \textit{and} \quad h_0(\xi_0)^{-2}\{\mathrm{Var}\,W(\xi_0) + R_0^{\rm t}J_{00}^{-1}R_0 + \zeta^{\rm t}\Omega_S K\Omega_S^{\rm t}\zeta\},$$

respectively.

The limit distribution involves the baseline hazard rate $h_0(u)$, which for its estimation would require a suitable smoothing operation, using e.g. a kernel smoother on $\widehat{H}_0(\cdot)$. This is however not really required here, since our aim is to compare mean squared errors, and the very same multiplicative factor $h_0(\xi_0)^{-1}$ enters each of the Λ_S. We may therefore compare and estimate mean squared errors of the variables $\bar\Lambda_S = h_0(\xi_0)\Lambda_S$, which has

$$\mathrm{E}\bar\Lambda_S^2 = \mathrm{Var}\,W(\xi_0) + R_0^{\rm t}J_{00}^{-1}R_0 + \zeta^{\rm t}\Omega_S K\Omega_S^{\rm t}\zeta + \zeta^{\rm t}(I-\Omega_S)\eta\eta^{\rm t}(I-\Omega_S)^{\rm t}\zeta.$$

Here the first three terms combine to give the variance part while the fourth term stems from the model bias. Also, the two first terms are not affected by the model choice S, and represent the minimal possible variance, achieved by using the narrow model, where $S=\emptyset$. We see that the structure of these limiting risk expressions is precisely of the same form as in (5.6), as found there via Lemma 3; the only essential difference is that

$$\zeta = R_1 - J_{10}J_{00}^{-1}R_0 = \frac{\partial f}{\partial\gamma} - J_{10}J_{00}^{-1}\frac{\partial f}{\partial\beta} - \{J_{10}J_{00}^{-1}F_0(\xi_0) - F_1(\xi_0)\}$$

of Lemma 4 replaces $\omega-\kappa$ of Lemma 3. Clearly ζ is very similar to $\omega-\kappa$, but as explained above Lemma 3 does not cover the type of estimands handled by Lemma 4, which needed a separate treatment.

Going back to the quantiles of the conditional survival time, for an individual with covariate (x,z), we have $\xi(r) = H_0^{-1}(f(\beta,\gamma))$ with $f(\beta,\gamma) = c\exp(-x^{\rm t}\beta - z^{\rm t}\gamma)$ and $c =$

$-\log(1-r)$ for the rth quantile. Thus $\zeta = R_1 - J_{10}J_{00}^{-1}R_0$ with $R_0 = F_0(\xi_0) - c\exp(-x^{\rm t}\beta)x$ and $R_1 = F_1(\xi_0) - c\exp(-x^{\rm t}\beta)z$, and $c = \log 2$ for the case of the median.

6.3. Estimation of risk quantities. Theory and calculations developed in the previous sections led to the definition of various model-based quantities that need to be estimated in practice. This is in particular required in light of the model selection criteria of the next section. Here we describe the required estimators. In this subsection we use $(\widehat\beta, \widehat\gamma)$ to signal the full-model based partial likelihood estimators.

This is a convenient place to discuss estimation of τ_0, $J = J_{\rm full}$, K, Ω_S, Ω_S, K_S, ω, κ, ζ. As the theory is being used in the following sections it demands that the estimators $\widehat\tau_0$, $\widehat J$ etc. that are used are consistent, i.e. that they should converge in probability to the relevant quantities as n grows, under the local neighbourhood circumstances (2.4). There are in fact several possibilities for say $\widehat J$ here, typically ranging from say $-I_n(\widehat\beta_{\rm narr}, 0)$ that uses estimators from the narrow model to $-I_n(\widehat\beta, \widehat\gamma)$ that employs estimators in the fullest $p+q$-parameter model. The first-order large-sample theory that we develop does not distinguish between these estimators, as long as they are consistent. We will in practice typically prefer the full-model based versions, partly for reasons of model-robustness. See CH and HC (2003) for parallel discussion.

We start with the information matrix J, for which we use the full-model based estimator (3.3). From this matrix we extract further estimates $\widehat J_{00}$, $\widehat J_{01}$, $\widehat J_{11}$, $\widehat J_S$, and furthermore $\widehat K = \widehat J^{11}$ with consequent $\widehat\Omega_S = \pi_S^{\rm t}\widehat K_S\pi_S\widehat K^{-1}$ matrices. We use full-model based parameter estimates also for estimating the $(p+q)$-vector function F of Section 4, giving

$$\begin{aligned}\widehat F(t) &= \int_0^t E_n(u, \widehat\beta, \widehat\gamma)\,{\rm d}\widehat H_0(u)\\ &= \sum_{i=1}^n I\{t_i \le t\} E_n(t_i, \widehat\beta, \widehat\gamma)\,\Delta\widehat H_0(t_i) = n^{-1}\sum_{i=1}^n I\{t_i \le t\}\frac{G_n^{(1)}(t_i, \widehat\beta, \widehat\gamma)}{G_n^{(0)}(t_i, \widehat\beta, \widehat\gamma)^2}\delta_i.\end{aligned}$$

Finally there are a couple of options when estimating the partial derivatives of $\mu(\beta, \gamma, H_0)$, which are required for arriving at $\widehat\omega$ and $\widehat\kappa$. In most of our examples we are able to find explicit expressions for these, as for the earlier examples in this section, after which we again insert parameter estimates from the fullest model. General numerical recipes might also be used in situations where explicit expressions are harder to come by.

For $\tau_0 = \tau_0(t)$ of (5.2), we insert estimates already described for the partial derivatives of μ with respect to β, γ, and $H_0(t)$, and likewise for $F_0(t)$ and $F_1(t)$. The remaining integral $\int_0^t g^{(0)}(u, \beta, 0)^{-1}\,{\rm d}H_0(u)$ is estimated as

$$\int_0^t \frac{{\rm d}\widehat H_0(u)}{G_n^{(0)}(u, \widehat\beta, \widehat\gamma)} = \sum_{i=1}^n I\{t_i \le t\}\frac{\Delta\widehat H_0(t_i)}{G_n^{(0)}(t_i, \widehat\beta, \widehat\gamma)} = n^{-1}\sum_{i=1}^n I\{t_i \le t\}\frac{\delta_i}{G_n^{(0)}(t_i, \widehat\beta, \widehat\gamma)^2}.$$

For handling the model selection problems associated with conditional median or quantile survival distributions, or more generally situations where Lemma 4 is applicable, one needs estimates of ξ_0, R_0, R_1 and ζ. We use $\widehat{\xi}_0 = \widehat{H}_0^{-1}(f(\widehat{\beta},\widehat{\gamma}))$, so for the median case we employ $\widehat{H}_0^{-1}(c\exp(-x^{\rm t}\widehat{\beta} - z^{\rm t}\widehat{\gamma}))$ with $c = \log 2$. Similarly we use

$$\widehat{R}_0 = \widehat{F}_0(\widehat{\xi}_0) - c\exp(-x^{\rm t}\widehat{\beta} - z^{\rm t}\widehat{\gamma})x, \quad \widehat{R}_1 = \widehat{F}_1(\widehat{\xi}_0) - c\exp(-x^{\rm t}\widehat{\beta} - z^{\rm t}\widehat{\gamma})z,$$

leading also to the crucial quantity $\widehat{\zeta} = \widehat{R}_1 - \widehat{J}_{10}\widehat{J}_{00}^{-1}\widehat{R}_0$.

7. The AIC and the FIC for Cox regression

This section uses theory developed in earlier sections to properly analyse the natural partial-likelihood based version of the AIC, and then goes on to derive a focussed information criterion, the FIC announced in Section 1.3, for general use in Cox proportional hazards regression models.

7.1. The AIC for the Cox model. Let $\widehat{\eta}$ be the estimator of $\eta = \sqrt{n}\gamma$ in the full Cox model with all $p + q$ covariates included. From Lemma 1,

$$\widehat{\eta} = \sqrt{n}\widehat{\gamma}_{\rm full} \to_d D \sim {\rm N}_q(\eta, K). \tag{7.1}$$

The natural statistic monitoring for absence or presence of η is

$$Z_n = \widehat{K}^{-1/2}\widehat{\eta} = \widehat{K}^{-1/2}\sqrt{n}\widehat{\gamma}_{\rm full} \to_d Z \sim {\rm N}_q(K^{-1/2}\eta, I),$$

with $\widehat{K}$ defined in Section 6.3.

The Akaike information criterion AIC is generally applicable for comparing competing parametric models; see e.g. Burnham and Anderson (2002) for a broad introduction with applications to many kinds of models. The arguments behind its construction do not necessarily apply to the Cox regression model, however, due to its semiparametric nature. We are not aware of any other attempts in the literature to define or discuss aspects or performance of the AIC for the Cox model. We are however free to define and analyse

$${\rm AIC}_{n,S} = 2\log L_{n,S}(\widehat{\beta}_S, \widehat{\gamma}_S) - 2(p + |S|), \tag{7.2}$$

in the style of parametric models, where $L_{n,S}$ is the partial likelihood function engaging (β, γ_S), see (2.1). The submodel S with largest AIC score (7.2) is selected.

A useful representation of ${\rm AIC}_{n,S}$ can be derived, in terms of $\widehat{\eta}$ and hence Z_n;

$$\begin{aligned}{\rm AIC}_{n,S} - {\rm AIC}_{n,\emptyset} &= Z_n^{\rm t} K^{-1/2}\pi_S^{\rm t} K_S \pi_S K^{-1/2} Z_n - 2|S| + o_p(1)\\ &= \widehat{\eta}^{\rm t} K^{-1}\pi_S^{\rm t} K_S \pi_S K^{-1}\widehat{\eta} - 2|S| + o_p(1).\end{aligned}$$

This may be shown following arguments in HC (2003a). Note in particular that all ${\rm AIC}_{n,S}$ numbers, across submodels S, depend essentially only on the Z_n vector. One may also show from this that

$${\rm AIC}_{n,S} - {\rm AIC}_{n,\emptyset} \to_d Z^{\rm t} K^{-1/2}\pi_S^{\rm t} K_S \pi_S K^{-1/2} Z - 2|S| \sim \chi^2_{|S|}(\eta^{\rm t} K_S^{-1}\eta) - 2|S|.$$

These results imply in particular that there are well-defined precise limit probabilities for the different submodels being chosen by the AIC; specifically,

$$\begin{aligned}\mathrm{ch}_n(S,\eta) &= \Pr\{\text{model } S \text{ is chosen}\}\\ &\to \mathrm{ch}(S,\eta) = \Pr\{\alpha(S) \text{ is bigger than all other } \alpha(S')\},\end{aligned}$$

where $\alpha(S) = D^{\rm t}K^{-1}\pi_S^{\rm t}K_S\pi_S K^{-1}D$ and D are as in (7.1). As an illustration, assume we wish to select either the narrow model with only β or the fullest model with all of (β,γ). Then

$$\mathrm{ch}_n(\mathrm{full},\eta) \to \mathrm{ch}(\mathrm{full},\eta) = \Pr\{\chi^2_q(\eta^{\rm t}K^{-1}\eta) \geq 2q\},$$

a probability that increases with the distance between η and zero.

It is worth pointing out that the AIC as developed here, using the partial likelihood, in a sense does not care about H_0 or about how well the indirectly selected $\widehat{H}_{0,S}$ estimator performs. Our FIC methods, to be developed now, may be geared towards good estimation performance for H_0, for example.

7.2. The focussed information criterion. We have demonstrated in Sections 5 and 6 that for focus estimands $\mu(\beta,\gamma,H_0)$, with ensuing estimators $\widehat{\mu}_S = \mu(\widehat{\beta}_S,\widehat{\gamma}_S,\widehat{H}_{0,S})$, the limiting risk can be expressed as

$$\mathrm{risk}(S) = \tau_0^2 + (\omega-\kappa)^{\rm t}\{(I-\Omega_S)\eta\eta^{\rm t}(I-\Omega_S)^{\rm t} + \Omega_S K\Omega_S^{\rm t}\}(\omega-\kappa),$$

for the relevant τ_0, ω and κ. When the full $(p+q)$-parameter model is used, for example, $\Omega_S = I_q$ and the risk function is $\tau_0^2 + (\omega-\kappa)^{\rm t}K(\omega-\kappa)$, constant in η. The other extreme is to select the narrow p-parameter model, for which $\Omega_S = 0$, leading to risk function $\tau_0^2 + \{(\omega-\kappa)^{\rm t}\eta\}^2$.

In these risk expressions, quantities τ_0, ω, κ, Ω_S, K may all be estimated consistently, with ordinary $\sqrt{n}$ precision, see the recipes of Section 6.3. The only quantity that can not be estimated consistently is $\eta\eta^{\rm t}$. For this quantity, about the best we can do is $\widehat{\eta\eta^{\rm t}} - \widehat{K} = n\widehat{\gamma\gamma^{\rm t}} - \widehat{K}$, in that $\widehat{\eta\eta^{\rm t}} \to_d DD^{\rm t}$, a variable with mean $\eta\eta^{\rm t} + K$. Thus we have an asymptotically unbiased risk estimator

$$\widehat{\mathrm{risk}}(S) = \widehat{\tau}_0^2 + (\widehat{\omega}-\widehat{\kappa})^{\rm t}\{(I-\widehat{\Omega}_S)(\widehat{\eta\eta^{\rm t}}-\widehat{K})(I-\widehat{\Omega}_S)^{\rm t} + \widehat{\Omega}_S\widehat{K}\widehat{\Omega}_S^{\rm t}\}(\widehat{\omega}-\widehat{\kappa}), \qquad (7.3)$$

for each candidate model S. The *focussed information criterion*, or FIC, consists in selecting the model with smallest estimated risk.

It is useful in practice to compute each of these risk numbers, since they have direct interpretation as estimates of sample size times mean squared error. One may also usefully display $\mathrm{FIC}^*(S) = \{\widehat{\mathrm{risk}}(S)/n\}^{1/2}$, since these are estimates of root mean squared error. Statisticians are used to interpreting standard errors, i.e. estimated standard deviations, the ubiquitous companions to estimates of model parameters. Here we suggest supplying also the FIC* numbers, along with $\widehat{\mathrm{FIC}}$ scores defined in the next paragraph.

As long as emphasis is on model selection we may simplify the above algebra somewhat, and give a crisper, but equivalent, version of the FIC. For this we subtract the constant $\widehat{\tau}_0^2$ which does not affect model comparison, and further subtract out the quantity $(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{K}(\widehat{\omega}-\widehat{\kappa})$, which is also common to each risk estimate. These rearrangements lead to

$$\widehat{\rm FIC}(S)=(\widehat{\psi}-\widehat{\psi}_S)^2+2(\widehat{\omega}-\widehat{\kappa})^{\rm t}\pi_S^{\rm t}\widehat{K}_S\pi_S(\widehat{\omega}-\widehat{\kappa}), \qquad (7.4)$$

in terms of estimates $\widehat{\psi}=(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{\eta}$ and $\widehat{\psi}_S=(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{\Omega}_S\widehat{\eta}$; see CH (2003, Section 4). This conveys better the statistical balancing game between modelling bias (the first term) and estimation variability (the second term). The FIC sees to it that an optimal model is selected for the particular task at hand; different estimands $\mu(\beta,\gamma,H_0)$ correspond to different $\omega-\kappa$ and different ψ.

We have reached formulae (7.3)–(7.4) in the framework of Section 5, using in particular Lemma 3, covering a broad variety of situations. For the case of quantile survival time estimators we need to employ Lemma 4 rather than Lemma 3, with a more complicated limit distribution. However, as argued after Lemma 4, the same structure emerges when we work with $\bar{\Lambda}_S=h_0(\xi_0)\Lambda_S$, which means that the FIC formulae above still work, with $\widehat{\zeta}$ replacing $\widehat{\omega}-\widehat{\kappa}$.

Remark 7.1. The FIC as developed here resembles the FIC model selector developed in CH (2003) for general parametric models. That theory could however not be applied directly to the proportional hazards model, partly because of its semiparametric nature and partly because the partial likelihood does not involve the baseline hazard. ■

Remark 7.2. The risk estimate (7.3) is the sum of a variance and a squared bias estimate. These terms can with a little algebra in combination with (7.1) be expressed as

$$V_n(S)=\widehat{\tau}_0^2+(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{\Omega}_S\widehat{K}\widehat{\Omega}_S^{\rm t}(\widehat{\omega}-\widehat{\kappa})=\widehat{\tau}_0^2+(\widehat{\omega}-\widehat{\kappa})^{\rm t}\pi_S^{\rm t}\widehat{K}_S\pi_S(\widehat{\omega}-\widehat{\kappa})$$

and

$$\begin{aligned}(B^2)_n(S)&=(\widehat{\omega}-\widehat{\kappa})^{\rm t}(I-\widehat{\Omega}_S)(\widehat{\eta}\widehat{\eta}^{\rm t}-\widehat{K})(I-\widehat{\Omega}_S)^{\rm t}(\widehat{\omega}-\widehat{\kappa})\\&=n\{(\widehat{\omega}-\widehat{\kappa})^{\rm t}(I-\widehat{\Omega}_S)\widehat{\gamma}\}^2-(\widehat{\omega}-\widehat{\kappa})^{\rm t}(\widehat{K}-\pi_S^{\rm t}\widehat{K}_S\pi_S)(\widehat{\omega}-\widehat{\kappa}),\end{aligned}$$

demonstrating also that $(B^2)_n(S)$ typically will increase with n, with a size essentially determined by $(\omega-\kappa)^{\rm t}(I-\Omega_S)\eta$. It can nevertheless happen that the event

$$\text{negligible bias}:\quad n\{(\widehat{\omega}-\widehat{\kappa})^{\rm t}(I-\widehat{\Omega}_S)\widehat{\gamma}\}^2<(\widehat{\omega}-\widehat{\kappa})^{\rm t}(\widehat{K}-\pi_S^{\rm t}\widehat{K}_S\pi_S)(\widehat{\omega}-\widehat{\kappa}) \qquad (7.5)$$

takes place, in which case we choose to redefine $(B^2)_n(S)$ as zero, to avoid estimating the squared bias with a negative number. Thus we redefine $\widehat{\rm risk}(S)=V_n(S)$ and

$$\widehat{\rm FIC}(S)=V_n(S)+(B^2)_n(S)-\widehat{\tau}_0^2-(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{K}(\widehat{\omega}-\widehat{\kappa})=(\widehat{\omega}-\widehat{\kappa})^{\rm t}(\pi_S^{\rm t}\widehat{K}_S\pi_S-\widehat{K})(\widehat{\omega}-\widehat{\kappa})$$

for those instances where the negligible bias event (7.5) takes place. ■

7.3. Securing good average performance. The FIC apparatus introduced above is at the outset tailor-made for optimal model selection when considering a single parameter of interest. One would not infrequently encounter focus parameters that depend on a covariate value, or a time point, that one next would wish to study across portions of the covariate space or time scale. We shall see now that the FIC machinery also yield methods for dealing with such problems.

For concreteness of illustration, consider the estimand $\mu = \mu(z) = \exp\{(x-x_0)^{\rm t}\beta + (z-z_0)^{\rm t}\}$ discussed in Section 6.1, but now viewed as a function of z with levels x_0 and x kept fixed. We wish to select a submodel S that provides optimal precision for estimates $\widehat\mu_S(z)$, across many values of z. From Lemma 3, $\sqrt{n}\{\widehat\mu_S(z)-\mu(z)\} \to_d \Lambda_S(z)$, say, of the form given there, with $\omega = \omega(z)$ calculated in (6.1). We infer from this that if Q_n is some distribution of covariates z, tending to some limit Q, then under mild conditions

$$\rho_n(S) = n\int \{\widehat\mu_S(z)-\mu(z)\}^2\,{\rm d}Q_n(z) \to_d \rho(S) = \int \Lambda_S(z)^2\,{\rm d}Q(z),$$

say. Hence the average risk of using $\widehat\mu_S(z)$, say ${\rm risk}_n(S) = {\rm E}\rho_n(S)$, converges to

$$\begin{aligned}{\rm risk}(S) = {\rm E}\rho(S) &= \int \big[\tau_0^2 + \omega(z)^{\rm t}\{(I-\Omega_S)\eta\eta^{\rm t}(I-\Omega_S)^{\rm t} + \Omega_S K\Omega_S^{\rm t}\}\omega(z)\big]\,{\rm d}Q(z)\\ &= \tau_0^2 + {\rm Tr}\Big[\{(I-\Omega_S)\eta\eta^{\rm t}(I-\Omega_S)^{\rm t} + \Omega_S K\Omega_S^{\rm t}\}\int \omega(z)\omega(z)^{\rm t}\,Q({\rm d}z)\Big].\end{aligned}$$

This can be estimated as in the previous subsection, plugging in consistent estimators of τ_0, Ω_S, K, along with the distribution Q_n for Q and $\widehat\eta\widehat\eta^{\rm t} - \widehat K$ for $\eta\eta^{\rm t}$. This leads to a list of $\widehat{{\rm risk}}(S)$ numbers to be minimised over candidate models S.

In the present case, with (6.1) for $\omega(z)$, the crucial $\int \omega(z)\omega(z)^{\rm t}\,Q({\rm d}z)$ matrix becomes

$$\begin{aligned}&J_{10}J_{00}^{-1}(x-x_0)(x-x_0)^{\rm t}J_{00}^{-1}J_{01} + \int (z-z_0)(z-z_0)^{\rm t}\,Q({\rm d}z)\\ &\qquad - J_{10}J_{00}^{-1}(x-x_0)(\bar z - z_0)^{\rm t} - (\bar z - z_0)(x-x_0)^{\rm t}J_{00}^{-1}J_{01},\end{aligned}$$

where $\bar z = \int z\,Q({\rm d}z)$. We could for example take Q_n to be the empirical distribution of $z_1,\ldots,z_n$, and z_0 to be the average of these, in which case the estimated $\int \omega\omega^{\rm t}\,{\rm d}Q$ matrix becomes $\widehat J_{10}\widehat J_{00}^{-1}(x-x_0)(x-x_0)^{\rm t}\widehat J_{00}^{-1}\widehat J_{01} + S_n$, with S_n being the empirical covariance matrix of the z_is.

The above FIC-averaging scheme is illustrated in Section 9.2 for the Danish skin cancer survival data. Note that the reasoning above is general in nature, and can be applied with appropriate variations to the task of finding the best subset S for best average estimation of the nine decile survival times ${\rm Su}^{-1}(j/10\,|\,x,z)$, for example.

8. Model average estimators

The previous section developed the FIC, to be used for model selection purposes in connection with any focus parameter μ of interest. It is also of interest to understand the statistical behaviour of the resulting estimator-post-selection strategy. Such estimators take the form $\widehat{\mu} = \widehat{\mu}(\widehat{S})$, say, where $\widehat{S}$ is the randomly selected submodel. This is a special case of a more general class termed *compromise estimators* in HC (2003a). This section develops theory for such model average strategies for the Cox model.

8.1. Model average estimators. When several candidate models are being considered, as above, a natural idea is to form compromise estimators that weight across models in a suitable fashion. Specifically, consider now

$$\widehat{\mu} = \sum_S w_n(S \,|\, \widehat{\eta})\widehat{\mu}_S, \tag{8.1}$$

where the weights depend on $\widehat{\eta} = \sqrt{n}\widehat{\gamma}_{\rm full}$ and sum to one. The AIC and FIC strategies are of this form, with weight 1 for the chosen submodel and 0 for the others. We may now state the following; the proof is in Section 11.

Lemma 5. *Suppose regularity conditions used in Lemmas 1, 2, 3 continue to hold, and assume that the random weights $w_n(S \,|\, \widehat{\eta})$ used in the compromise estimator (8.1) are such that the vector of $w_n(S \,|\, \widehat{\eta})$ tends in distribution to the vector of $w(S \,|\, D)$, in terms of the limit D of $\widehat{\eta}$ as in (7.1), where each $w(S \,|\, D)$ has at most a finite number of discontinuities in D. Then*

$$\sqrt{n}(\widehat{\mu} - \mu_{\rm true}) \to \Lambda = \Lambda_0 + (\omega - \kappa)^{\rm t}\{\eta - \widehat{\eta}(D)\}.$$

Here $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ is independent of $D \sim {\rm N}_q(\eta, K)$, and $\widehat{\eta}(D) = \sum_S w(S \,|\, D)\Omega_S D$.

As a consequence of this result, we may for any model average estimator compute its limit risk function under squared error loss as ${\rm risk}(\eta) = {\rm E}\Lambda^2 = \tau_0^2 + R(\eta)$, say, where

$$R(\eta) = {\rm E}[(\omega - \kappa)^{\rm t}\{\widehat{\eta}(D) - \eta\}]^2 = (\omega - \kappa)^{\rm t}\,{\rm E}\{\widehat{\eta}(D) - \eta\}\{\widehat{\eta}(D) - \eta\}^{\rm t}\,(\omega - \kappa).$$

One should note the broad generality here; the performance of almost every model average strategy can be precisely assessed, for large n, by evaluating the precision of the estimator $(\omega - \kappa)^{\rm t}\widehat{\eta}(D)$ for the estimand $(\omega - \kappa)^{\rm t}\eta$, in the limit experiment where $D \sim {\rm N}_q(\eta, K)$, and where all quantities are known apart from η. In particular performance of the AIC versus that of the FIC and so on can be studied, for different situations determined by K and $\omega - \kappa$.

8.2. Smoothed AIC and smoothed FIC. Lemma 5 is of course very general and allows a broad class of model average estimators. Among these we may single out two procedures, namely smoothed versions of the AIC and the FIC. Further options are discussed in HC (2003a).

For the smoothed AIC, use (8.1) with data-determined weights

$$w_{\rm aic}(S) = \frac{\exp(\frac{1}{2}\lambda {\rm AIC}_{n,S})}{\sum_{{\rm all}\ S'} \exp(\frac{1}{2}\lambda {\rm AIC}_{n,S'})}, \tag{8.2}$$

in terms of the AIC scores (7.2). The sum in the denominator extends over all submodels under consideration; the list of these does not have to be extensive, as often some submodels might be ruled out on a priori grounds. The λ of (8.2) is like a smoothing parameter, dictating the amount of smoothing between candidate models. If λ is large, then the method is essentially equivalent to the AIC selection scheme; if $\lambda = 0$ then all methods are weighted equally. Certain arguments discussed in Buckland, Burnham and Augustin (1997) and Burnham and Anderson (2002) on an ad hoc basis and more fully in HC (2003a) advocate taking $\lambda = 1$ in (8.2). This is the value we take in our simulations and illustrations in Section 9. We use the same strategy to form a smoothed BIC, simply replacing the AIC scores in (8.2) with those of the BIC.

In a similar manner, the smoothed FIC uses (8.1) with weights

$$w_{\rm fic}(S) = \exp\Big\{-\tfrac{1}{2}\lambda \frac{\widehat{\rm FIC}(S)}{(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{K}(\widehat{\omega}-\widehat{\kappa})}\Big\}\Big/ \sum_{{\rm all}\ S'} \exp\Big\{-\tfrac{1}{2}\lambda \frac{\widehat{\rm FIC}(S')}{(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{K}(\widehat{\omega}-\widehat{\kappa})}\Big\}, \tag{8.3}$$

again with a parameter λ determining the degree to which lower FIC scores should be compared to higher ones. The point of the scaling here, via the $(\widehat{\omega}-\widehat{\kappa})^{\rm t}\widehat{K}(\widehat{\omega}-\widehat{\kappa})$ factor, is to make different situations similar with respect to the scale of the smoothing parameter λ; $(\omega-\kappa)^{\rm t}K(\omega-\kappa)$ is the constant risk of the minimax method in the limit experiment alluded to after Lemma 5. In our illustrations we have taken $\lambda = 1$. Larger values would push the model average method closer to the FIC method, and values closer to zero would correspond to equal weights across the submodels considered.

Variations exist, like using (8.2) and (8.3) involving say only the ten top marked models. Lemma 5 still applies and describes accurately the large-sample performance also of such model average schemes. Limit distributions are non-linear mixtures of normals, and as such non-normal; see the illustration of Section 10.6.

9. Illustrations and applications

In this section we illustrate our FIC and model average methods in simulations, where we find that the post-selection FIC as well as the smoothed FIC methods may perform well in comparison with for example the AIC and BIC regimes. Then we analyse the Danish skin cancer survival study that was described in Section 1.1.

9.1. Results of a simulation study. Data t_i are generated following a Cox proportional hazard regression model with constant baseline hazard $h_0(t) = 1$. Covariates are generated from independent standard normal distributions. In each setting we use $p = 2$ protected variables and decide on $\beta = (1,1)^{\rm t}$. Censoring times are generated from an exponential distribution with mean 10/9. In our settings this corresponds to an average proportion of uncensored observations about 52%. In setting (i) $q = 4$ and data are generated under the narrow model assumption, i.e. $\eta = (0,0,0,0)^{\mathrm{t}}$. Situation (ii) is as the first one but corresponds to the full model with $\eta = (3,-3,3,-3)^{\mathrm{t}}$. In setting (iii) a situation in between the narrow and the full model is taken with $q = 6$ but $\eta = (0,0,3,-3,3,-3)^{\mathrm{t}}$. In these situations we study four focus parameters. Focus parameter (a) is the relative risk of a subject with covariates at value 0.5, relative to the mean covariate values, i.e. $\mu(x,z) = \exp(x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma)$ with each x and each z fixed at 0.5. Focus point (b) is $H_0(t)$, the cumulative baseline hazard rate function at time $t = 0.5$. Our third focus (c) is the survival probability $\mathrm{Su}(t\,|\,x,z)$ for a subject with the same covariates and time values as in (a) and (b). The last focus point (d) is the median survival time $\xi(0.5\,|\,x,z)$, with again the same covariate values as in (a).

For each of the sim = 1000 simulation runs we compute for all subsets the estimators of the focus parameters μ in each of the 2^q models, together with the values of AIC, BIC, as per formulae (1.2), and for each focus parameter the corresponding FIC. The final estimators are the post-model-selection estimators pAIC, pBIC, pFIC as well as model averaged, weighted, estimators wAIC, wBIC, wFIC, with model weights based on the values of the information criteria for that model, see Section 8.2. We compute the root mean squared errors $\{\mathrm{sim}^{-1}\sum_{j=1}^{\mathrm{sim}}(\widehat{\mu}_j - \mu_{\mathrm{true}})^2\}^{1/2}$, across the simulations. Two sample sizes are used, $n = 150$ and $n = 300$. The results are summarised in the Table 9.1 below. The winning criterion is in each situation identified with its score given in boldface. It should be noted that with sim = 1000 runs there is some simulation uncertainty that leaves some of the comparisons still open and the identified 'winners' not quite clear.

		wFIC	pFIC	wAIC	pAIC	wBIC	pBIC
				Setting (i)			
$n = 150$	(a)	0.444	0.441	0.451	0.474	**0.391**	0.402
	(b)	0.089	0.089	0.089	0.090	**0.088**	0.089
	(c)	0.062	0.062	0.063	0.065	**0.058**	0.060
	(d)	0.046	**0.042**	0.047	0.049	0.043	0.044
$n = 300$	(a)	0.276	0.276	0.281	0.302	**0.247**	0.248
	(b)	0.063	0.063	0.063	0.063	0.063	**0.062**
	(c)	0.043	0.043	0.043	0.045	**0.040**	0.040
	(d)	0.034	**0.031**	0.034	0.036	0.032	0.032
				Setting (ii)			
$n = 150$	(a)	**0.503**	0.530	0.600	0.645	0.596	0.681
	(b)	**0.090**	0.092	0.092	0.093	0.091	0.092
	(c)	**0.069**	0.076	0.077	0.080	0.078	0.086
	(d)	0.062	**0.056**	0.060	0.063	0.064	0.072
$n = 300$	(a)	**0.304**	0.315	0.364	0.383	0.368	0.430
	(b)	**0.063**	0.064	0.064	0.064	0.064	0.064
	(c)	**0.047**	0.051	0.052	0.054	0.053	0.059
	(d)	0.041	**0.037**	0.041	0.043	0.043	0.048

Setting (iii)							
$n = 150$	(a)	**0.572**	0.582	0.674	0.739	0.625	0.725
	(b)	**0.089**	0.090	0.091	0.093	0.090	0.092
	(c)	**0.071**	0.075	0.080	0.086	0.079	0.088
	(d)	0.066	0.063	**0.062**	0.066	0.064	0.070
$n = 300$	(a)	0.328	**0.325**	0.398	0.426	0.392	0.458
	(b)	0.063	0.064	0.064	0.064	**0.063**	0.064
	(c)	**0.048**	0.051	0.053	0.056	0.053	0.059
	(d)	0.042	**0.038**	0.041	0.044	0.042	0.047

TABLE 9.1. *Root mean squared errors over 1000 simulation runs of the post model selection and model averaged estimators based on FIC, AIC and BIC for focus parameters (a) relative risk, (b) cumulative hazard, (c) survival probability and (d) median. Setting (i) corresponds to* $\eta = (0,0,0,0)^{\rm t}$, *(ii) to* $\eta = (3,-3,3,-3)^{\rm t}$ *and (iii) to* $\eta = (0,0,3,-3,3,-3)^{\rm t}$.

In setting (i) the narrow model is the true model. It is known that the BIC is a consistent model selector and that it often works well for models with a small number of parameters. Hence it is expected to do well for this situation, as is indeed seen from the simulation results. It should be noticed that especially for focus parameters (b) and (d) the differences with the FIC values are only minor. For settings (ii) and (iii) where there are four more non-zero parameters in the true model than in the narrow model, BIC is no longer preferred. The smoothed FIC is clearly the best choice for setting (ii), where the wide model is true, while for the median as a focus point, the post-FIC selector gives the best results. The picture is more undecided for setting (iii) in between narrow and full model. For the smaller sample size for focus (c) the post-AIC gives the smallest simulated mse. Overall, model averaging tends to yield smaller simulated mse than post-model selection. Considering only the post-model selection estimators, we see that the pBIC performs well for the simplest setting where all extra parameters are zero. For setting (ii) the pFIC is the best, with the same conclusion for setting (iii), where all three criteria perform about equal for focus (b).

9.2. Survival analysis for malignant melanoma. Here we examine the data set described in Section 1. As already motivated, we include x_1 in every model, and select amongst the other variables $z_1, \ldots, z_6$ using an all subsets search. The seven hazard regression coefficient estimates were 0.535 (0.277) for β_1, 0.036 (0.052) for γ_1, 0.321 (0.192) for γ_2, -0.707 (0.314) for γ_3, -0.995 (0.324) for γ_4, 0.334 (0.241) for γ_5, 0.017 (0.008) for γ_6, with the estimated standard deviation (standard error) in parentheses, computed using the full model. In particular variables x_1, z_2, z_3, z_4, z_6 might be considered to have a reasonably clear influence on life-times, as measured by the ratios estimate divided by standard error. Coefficients γ_1 and γ_5 would however not be seen as significantly different from zero in most analyses. We shall nevertheless see that variable z_5 often will be selected, by different criteria, for different purposes.

The model selection methods applied are AIC, BIC and four versions of FIC, each corresponding to a particular focus parameter. FIC1 corresponds to $\mu_1 = \exp\{(x-x_0)^{\rm t}\beta + (z-z_0)^{\rm t}\gamma\}$, relative risk of a man, with average tumour thickness amongst all men participating in the study, infection infiltration level $z_2 = 3$, epithelioid cells not present $(z_3 = 2)$, ulceration present $(z_4 = 1)$, invasion depth $z_5 = 1$, and average men's age in the study, as compared to that of women with averages of thickness of tumour and age computed over the subgroup of women in the study, and the other covariates remaining the same. The first set of covariates for men defines the variables level (x, z), while the second set corresponds to (x_0, z_0). FIC2 computes the FIC values for $\mu_2 = H_0(t)$ at time $t = 1584$ days which corresponds to the time where the estimated Kaplan–Meier survival probability reaches 0.85. The third focus, which defines FIC3, is the survival probability at time $t = 1584$ for the same set of covariates (x, z) as for the first focus parameter, i.e. $\mu_3 = {\rm Su}(t \,|\, x, z)$. The final focus parameter is $\mu_4 = \xi(0.10) = {\rm Su}^{-1}(0.90 \,|\, x, z)$, the time at which at least 90% of the patients with covariate level (x, z) are still alive.

Table 9.2 shows the 20 highest ranked score values for each of the criteria in question, after having searched through all $2^6 = 64$ candidate models, along with the selected z_j variables; '245' means that variables z_2, z_4, z_5 are included, etc. Note that the values are sorted in importance per criterion.

vars	AIC	vars	BIC	vars	FIC1	vars	FIC2	vars	FIC3	vars	FIC4
23456	−527.15	45	−542.72	∅	2.84	5	4.70	5	0.12	∅	0.029
3456	−528.29	14	−542.98	2	4.11	25	5.78	45	0.15	5	0.038
12346	−528.59	345	−544.04	26	4.17	235	5.81	6	0.15	56	0.038
123456	−528.69	24	−544.71	25	4.30	56	5.86	46	0.15	15	0.039
2346	−529.57	4	−544.89	256	4.35	356	5.87	16	0.15	25	0.039
13456	−529.65	3456	−544.91	6	4.63	2	6.05	146	0.15	156	0.040
2345	−529.94	124	−545.05	456	4.85	35	6.06	56	0.15	256	0.040
1234	−530.45	134	−545.42	14	5.22	125	6.23	456	0.15	3	0.042
1346	−530.51	456	−545.47	145	5.22	156	6.31	156	0.16	23	0.042
345	−530.74	245	−545.49	146	5.26	1356	6.32	1456	0.16	13	0.042
12345	−531.08	234	−545.61	1456	5.26	3456	6.37	15	0.17	123	0.042
2456	−531.26	146	−546.07	34	5.59	1235	6.46	2	0.17	1256	0.042
1246	−531.34	34	−546.11	346	5.78	26	6.65	25	0.18	36	0.043
124	−531.76	2346	−546.18	345	5.82	236	6.65	24	0.18	236	0.043
1345	−531.79	145	−546.52	23	5.94	126	6.73	245	0.18	136	0.043
12456	−532.10	2345	−546.55	5	5.94	1236	6.74	12	0.18	1236	0.043
134	−532.13	346	−546.83	12	6.01	2346	6.79	124	0.18	125	0.049
456	−532.18	1234	−547.06	134	6.10	456	7.01	125	0.18	1346	0.053
245	−532.20	246	−547.08	1345	6.11	2456	7.23	1245	0.18	345	0.055
234	−532.32	23456	−547.09	45	6.13	23456	7.25	145	0.21	2345	0.056

TABLE 9.2. *Values of the information criteria AIC, BIC and FIC for four focus parameters: (1) relative risk, (2) cumulative hazard, (3) survival probability, and (4) 10% quantile* $\xi(0.10) = {\rm Su}^{-1}(0.90)$. *The table shows the 20 largest AIC and BIC values and the 20 smallest FIC values for each of the focus parameters.*

The values are sorted for each criterion and for each focus parameter, and 'vars' indicate the selected variables among $z_1, \ldots, z_6$.

The AIC model choice method yields a model with the five variables z_2, z_3, z_4, z_5, z_6; only tumour thickness z_1 is not selected. The BIC on the other hand selects only variables z_4 (ulceration) and z_5 (invasion depth). Note that variable z_4 is present in all the 20 best BIC models. With the FIC there is not one single answer for the 'best model', as explained in Sections 1 and 7; the model chosen by FIC depends on the focus. The relative risk as a focus parameter lets FIC point to the narrow model, followed by models with variables z_2 and then $\{z_2, z_6\}$, and so on. In the second example, to estimate the cumulative hazard $H_0(t)$ at time $t = 1584$, only variable z_5 (invasion depth) is selected, with second best model being $\{z_2, z_5\}$. The picture is somewhat different when studying the survival probability, focus μ_3. Here variable z_5 is again the most important, followed by the model $\{z_4, z_5\}$, and so on. Note that variable z_6 (age) shows up quite frequently in the list of the best FIC3 models, whereas variable z_5 (invasion depth) appears to be the most important one for FIC2. For the 10% quantile, that is $\mathrm{Su}^{-1}(0.90 \,|\, x, z)$, we find the narrow model, including none of the extra variables, as the best FIC4 choice, with models $\{z_5\}$ and $\{z_5, z_6\}$ as second best.

The fact that different models are selected for different purposes should not lead to confusion; it should rather be seen as a way of strengthening the biostatistician's ability to produce more precise estimates or predictions for a specific patient or patient group.

In the next table we study the model selection problem for focus parameter (3), the survival probability for a patient with covariates (x, z), in more depth. We observe in particular that for the ten best models, according to FIC3, the estimates of the focus parameter $\widehat{\mu}_S = \widehat{\mathrm{Su}}(t \,|\, x, z)$ are quite variable, ranging from 0.209 (variables 5, 6) to 0.590 (variables 4, 5). For other situations the estimates for the best say ten models may be more homogeneous than for this particular case. Note that one version of the smoothed FIC method described in Section 8.2 is to take a weighted average of the ten best candidate estimates $\widehat{\mu}_S$ of Table 9.3, with weights as given there.

While the arguably most important output from such a FIC analysis would be this list of the most important models, along with FIC scores and the list of corresponding $\widehat{\mu}$ estimates, it is also often fruitful to examine the bias and standard deviation components that combine to give the risk estimate (7.3). As we saw in Section 7, this risk estimate is directly related to the FIC formula (7.4), and sorting models by the FIC score is equivalent to do the sorting by risk estimate. The bias can be estimated in a couple of different ways. What we may term the 'direct bias estimate' is to plug in estimates in the bias formula in Lemma 3 of Section 5, i.e. using bias_S equal to $(\widehat{\omega} - \widehat{\kappa})^{\mathrm{t}}(I - \widehat{\Omega}_S)\widehat{\eta}/\sqrt{n}$. We divide by $\sqrt{n}$ here to give an estimate for the genuine bias of $\widehat{\mu}_S$ at sample size n. The alternative way is as spelled out in Remark 7.2, with bias_S equal to the signed square root of $(B^2)_n(S)/\sqrt{n}$. It is with this second version, via the correct estimate of the squared bias, that we have the Pythagorean combination of bias_S and se_S giving $\mathrm{mse}_S/\sqrt{n}$.

The FIC reflects the bias–variance trade-off phenomenon, as seen here by having larger biases for simpler models and larger standard errors for more complex models, even though the standard errors and resulting mean squared error estimates are fairly close on this particular situation.

vars	$\widehat{\mu}_S$	se_S	bias_S	$(\mathrm{mse}_S/n)^{1/2}$	FIC_S
5	0.430	0.047	0.020	0.047	0.119
45	0.590	0.047	0.026	0.048	0.149
6	0.374	0.048	0.011	0.048	0.151
46	0.534	0.048	0.015	0.048	0.151
16	0.464	0.048	0.017	0.048	0.152
146	0.553	0.048	0.018	0.048	0.152
56	0.209	0.049	−0.003	0.049	0.154
456	0.363	0.049	0.004	0.049	0.155
156	0.325	0.049	0.006	0.049	0.159
1456	0.431	0.049	0.009	0.049	0.159
23456	0.225	0.052	−0.003	0.052	0.228
45	0.590	0.047	0.026	0.048	0.149

TABLE 9.3. *For focus parameter (3), the survival probability* $\mathrm{Su}(t\,|\,x,z)$, *we give for each of the ten best models according to FIC, as well as for AIC (one but last line) and BIC (last line) the variables in the model, the estimate* $\widehat{\mu}$, *classical standard error pretending this is the true model, estimated bias, root mean squared error, and the value of FIC.*

We take the FIC analysis of this dataset one step further by examining the average FIC and risk criteria developed in Section 7.3 to select models with good performance across user- and context-defined portions of the covariate space. Specifically, we study focussed model selection for the relative risk of a man, with average tumour thickness amongst all persons participating in the study, infection infiltration level $z_2 = 3$, epithelioid cells not present ($z_3 = 2$), ulceration present ($z_4 = 1$), invasion depth $z_5 = 1$, compared to the risk of women with covariate information (x_0, z_0) as given earlier in this section; this corresponds to a certain $\mu(x, z, x_0, z_0)$ parameter, with $x = 2$ for man vs. $x = 1$ for women. The present task is to find the best submodel for best average performance, weighted across all ages z_6. Following the method of Section 7.3 we compute, for all 205 subjects in the study, with ages ranging from 4 to 95, and for each of the $2^6 = 64$ models S, the values of FIC and the corresponding $(\mathrm{mse}_S/n)^{1/2}$. The resulting 205 × 64 FIC values can be analysed by patient. Summarising this we have the following situation:

Variables:	1	24	26	2	35	3	46	4	5	$\emptyset$
Times selected:	10	37	3	7	7	12	12	35	10	72

This corresponds to individual selection per patient, or per age. As there clearly is a difference in individual model choice, for an overall model we follow the approach of Section 7.3 and compute both the average FIC_S and the average $\widehat{\mathrm{risk}}(S)$ values, taken over all 205 patients, and next order the resulting 64 averaged values. Results for the five best models are given in Table 9.4. We see that the overall model with variables z_2 and z_6 is deemed the best choice. This application points yet again to the importance of first thinking about the use of the selected model before blindly applying a model selection criterion.

vars	avg FIC1	$(\mathrm{mse}_S/n)^{1/2}$
26	24.92	30.079
456	25.44	30.088
346	26.23	30.102
126	26.37	30.104
1346	26.39	30.105

TABLE 9.4. *For focus parameter (1), the relative risk as specified in the text, we give for each of the five best models according to averaged FIC the variables in the model, the value of the averaged FIC, and of the averaged risk.*

The picture does not always have to be so diverse. If we carry out the same exercise for focus parameters μ_2, the FIC2 best model is chosen in all of the 205 cases, and the model for best average performance coincides with the individual best models. This gives a good overall model for μ_2, independent of age.

10. Concluding remarks

We end our article with some comments and remarks, some of which might point to further research work.

10.1. 'Protected' versus 'open'. Our framework uses p 'protected' covariates, designated to always be inside the chosen models, along with q 'open' covariates that may or not may not be included in the final models. Deciding which is which is context-related and up to the statistician. For the analysis of Section 9.2 we could have chosen to use both sex x_1 and age z_6 as protected, for example, leaving the FIC and model average machinery to work with z_1, z_2, z_3, z_4, z_5 as open. The R software programmes we have developed make it easy to work through each desired combination. One may also choose not to pinpoint any protected covariates at all, i.e. using $p = 0$ and leaving it all to data to decide which covariates should be included for what purposes. Our Lemmas 1–5 can be extended to cover the required $p = 0$ case.

For the Danish melanoma data set we chose to use infection infiltration level z_2 as well as invasion depth z_5 as unperturbed covariates, on their original scale, viz. 1,2,3,4 for z_2 and 1,2,3 for z_5. These could also be broken down into sub-covariates, via indicator variables, and they could be taken as 'ordered' or not. This would mean more modelling robustness but also more parameters, in fact $1 + 9$ instead of $1 + 6$ regression parameters. We could still run our programmes searching for ex- or inclusion of $z_1, \ldots, z_6$, i.e. over the appropriate 2^6 subsets of the 2^9.

10.2. Pretesting, backward and forward selection. A simple pretesting approach is sometimes followed for covariate inclusion. For the skin cancer data one might include those

z_j for which $|W_{n,j}| = |\widehat\gamma_j/\mathrm{se}(\widehat\gamma_j)|$ exceeds say 1.645, corresponding to test significance level of 0.10 per coefficient; this yields the $\{z_2, z_3, z_4, z_6\}$ model, with z_1 and z_5 excluded. One might likewise have attempted versions of 'forward' and 'backward' selection strategies. It is important to realise that each of these methods is covered by Lemma 5 of Section 8, with appropriate non-normal limit distributions. This also makes it possible to compare performances via the risk function $R(\eta)$ given there. A systematic study of this sort would be useful. A tentative and partial conclusion from some evidence presented in CH and HC (2003) is that the AIC and the FIC will tend to outperform the pre-test regime, and that the smoothed versions wAIC and wFIC often are even better.

10.3. Two cultures. There are at least two uses of statistical modelling of the Cox regression variety (and more general models); they may be used primarily for interpretation, perhaps of a biomedical nature, like finding that high blood pressure is associated with higher risk, and they may be used primarily as a vehicle for producing precise estimation of quantities like median survival time, relative risk, survival probabilities, etc. Though these uses are not fully orthogonal, see e.g. the discussion in and to Breiman (2001), the viewpoint of the present article has (again, primarily) been the second one, aiming for as precise estimation and prediction as possible.

10.4. Even more submodels. We have studied estimators of the form $\mu(\widehat H_{0,S}, \widehat\beta_S, \widehat\gamma_S)$, but may in principle use different covariate subsets for the H_0 part and the (β, γ) part. The techniques of our paper allows extensions to be made to cover the required simultaneous distribution of all 2^{2q} combined estimators, and would lead to a somewhat more general FIC, and so on, but we abstain from pursuing this here.

10.5. Bayesian model averaging. Section 8 dealt with the large class of compromise or model average estimators, with Lemma 5 being a 'master theorem' giving the limit distribution of all such estimators. It can be further generalised in the direction of 'generalised ridging' estimators, as in HC (2003a, Section 8 and 9). These take the form

$$(\widetilde\beta_S, \widetilde\gamma_S) = (\widehat\beta_S, \varepsilon(S \,|\, \widehat\eta)\widehat\gamma_S) \quad \text{with shrinking factor} \quad \varepsilon(S \,|\, \widehat\eta) \in [0, 1],$$

with consequent $\widetilde\mu_S = \mu(\widetilde\beta_S, \widetilde\gamma_S, \widetilde H_{0,S})$, and may be particularly useful when the number q of extra covariates is becoming large. It turns out that a class of Bayesian model averaging methods, as worked with from several perspectives in Volinsky, Madigan, Raftery and Kronmal (1997), Clyde (1999), Hoeting, Madigan, Raftery and Volinsky (1999), Volinsky and Raftery (2000), becomes first-order equivalent to a special subclass of such generalised ridge estimators. Their performance can thus be studied along with those of the frequentist model average schemes of Section 8.2.

10.6. Confidence intervals and tests. Our article has developed methods for selecting 'the right' variables in applications with the Cox regression model, with different optimal subsets for different contexts and uses. The methodology has also extended to averages over candidate models and has reached precise descriptions of the large-sample distributions involved. These descriptions can e.g. be used to illustrate the 'overoptimism' involved when one applies the standard output from Cox regression analysis too simplistically, as when one sets a 95% confidence interval for a parameter based on the AIC selected model, without taking into account the extra uncertainty associated with the model selection step. See Section 4 of HC (2003a) for concrete illustrations of this. The same phenomenon affects the significance levels of tests, where a tentative 5% test might in reality have a size much bigger.

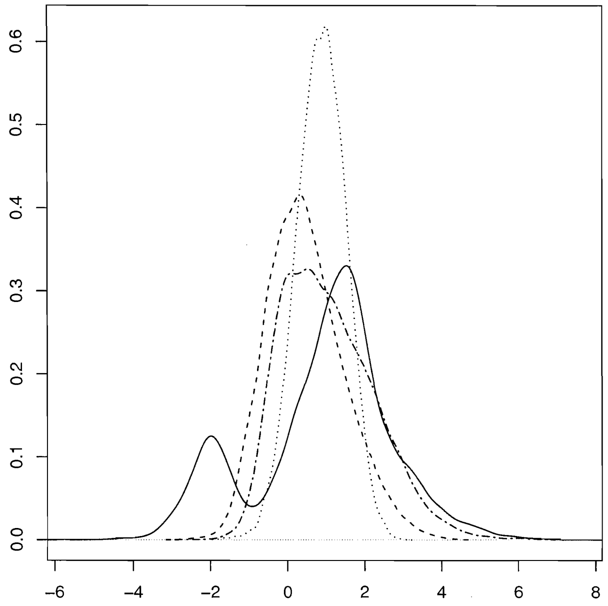


FIGURE. *For the second focus estimand μ_2 of Section 9.2, which is the cumulative hazard rate function for the Danish skin cancer survival data, the figure displays densities for the limit distributions Λ for $\sqrt{n}(\widehat{\mu}_2 - \mu_2)$, for four different estimators. These are the FIC method (solid line), the AIC method (dashed line), the smoothed FIC method (dot-dashed line), and the smoothed AIC method (dotted line). The densities are computed as kernel estimates based on 10,000 simulations from the four appropriate versions of Λ as per Lemma 5 of Section 8, at the position $\widehat{\eta} = \sqrt{n}\widehat{\gamma}_{\rm full}$ in the η parameter space.*

What we have not touched directly, so far, is the application of the large-sample results to supply 'real' confidence intervals and 'real' significance tests. This is not easy, partly because the required limit distributions of $\sqrt{n}(\widehat{\mu} - \mu)$ are non-linear mixtures of many normals. This is illustrated in the Figure, for four estimation strategies, for the case of the second focus estimand μ_2 in the Danish skin cancer survival study discussed in Section 9.2. Confidence intervals and tests with correct levels of confidence and significance for large samples can be constructed based on assessment of these limit distributions, in several ways; see the 'better confidence' recipe and the general discussion in HC (2003a, Section 4).

11. Proofs of lemmas

Proof of Lemma 1. We start from Taylor expansion

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} U_n(\widehat\beta_S, \widehat\gamma_S) \\ V_{n,S}(\widehat\beta_S, \widehat\gamma_S) \end{pmatrix} \doteq \begin{pmatrix} U_n(\beta, 0) \\ V_{n,S}(\beta, 0) \end{pmatrix} + I_{n,S}(\beta, 0) \begin{pmatrix} \widehat\beta_S - \beta \\ \widehat\gamma \end{pmatrix},$$

which with appropriate analysis of the error term involved leads to

$$\begin{pmatrix} \sqrt{n}(\widehat\beta_S - \beta) \\ \sqrt{n}\widehat\gamma_S \end{pmatrix} = \{-I_{n,S}(\beta, 0)\}^{-1} \begin{pmatrix} \sqrt{n}U_n(\beta, 0) \\ \sqrt{n}V_{n,S}(\beta, 0) \end{pmatrix} + o_p(1).$$

Introduce the martingales

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\exp(x_i^{\rm t}\beta + z_i^{\rm t}\eta/\sqrt{n})\,{\rm d}H_0(u) \quad \text{for } i = 1, \ldots, n.$$

These are orthogonal with variance functions $\langle M_i, M_i\rangle(t) = \int_0^t Y_i(u)\exp(x_i^{\rm t}\beta + z_i^{\rm t}\eta/\sqrt{n})$ ${\rm d}H_0(u)$, see e.g. Andersen et al. (1993). We may then write

$$\begin{pmatrix} \sqrt{n}U_n(\beta, 0) \\ \sqrt{n}V_{n,S}(\beta, 0) \end{pmatrix} = n^{-1/2}\sum_{i=1}^n \int_0^\tau \Big\{\begin{pmatrix} x_i \\ z_{i,S} \end{pmatrix} - E_{n,S}(u, \beta, 0)\Big\}\,{\rm d}M_i(u)$$
$$+\, n^{-1/2}\sum_{i=1}^n \int_0^\tau \Big\{\begin{pmatrix} x_i \\ z_{i,S} \end{pmatrix} - E_{n,S}(u, \beta, 0)\Big\} Y_i(u)\exp(x_i^{\rm t}\beta + z_i^{\rm t}\eta/\sqrt{n})\,{\rm d}H_0(u).$$

The first term is an integral of a previsible function with respect to a martingale, is therefore itself a martingale evaluated at infinity, and with total variance matrix

$$n^{-1}\sum_{i=1}^n \int_0^\tau \Big\{\begin{pmatrix} x_i \\ z_{i,S} \end{pmatrix} - E_{n,S}(u, \beta, 0)\Big\}\Big\{\begin{pmatrix} x_i \\ z_{i,S} \end{pmatrix} - E_{n,S}(u, \beta, 0)\Big\}^{\rm t}$$
$$Y_i(u)\exp(x_i^{\rm t}\beta + z_i^{\rm t}\eta/\sqrt{n})\,{\rm d}H_0(u),$$

which is seen to converge in probability towards J_S. After some algebra and analysis the second term may be written

$$n^{-1}\sum_{i=1}^n \int_0^\tau \Big\{\begin{pmatrix} x_i \\ z_{i,S} \end{pmatrix} - E_{n,S}(u, \beta, 0)\Big\} z_i^{\rm t} Y_i(u)\exp(x_i^{\rm t}\beta)\,{\rm d}H_0(u)\,\eta + O_p(n^{-1/2}),$$

and this is seen to converge to the matrix with first p rows equal to $J_{01}\eta$ and the next $|S|$ rows equal to $\pi_S J_{11}\eta$. This proves the lemma. ■

Proof of Lemma 2. With some Taylor approximation analysis one finds that ${\rm d}\widehat H_{0,S}(u)$ may be expressed as

$$\frac{\sum_{i=1}^n {\rm d}N_i(u)}{nG_n^{(0)}(u, \widehat\beta_S, \widehat\gamma_S)} = \frac{\sum_{i=1}^n \{{\rm d}M_i(u) + Y_i(u)\exp(x_i^{\rm t}\beta + z_i^{\rm t}\eta/\sqrt{n})\}}{nG_n^{(0)}(u, \beta, 0)}$$
$$\times \Big\{1 - \frac{G_{n,S}^{(1)}(u, \beta, 0)^{\rm t}}{G_n^{(0)}(u, \beta, 0)}\begin{pmatrix} \widehat\beta_S - \beta \\ \widehat\gamma_S \end{pmatrix}\Big\}$$

plus terms of order $O_p(n^{-1})$, and which with some further efforts becomes

$$\Big[\mathrm{d}H_0(u) + \frac{\sum_{i=1}^n \mathrm{d}M_i(u)}{nG_n^{(0)}(u,\beta,0)} + n^{-1/2}\frac{\sum_{i=1}^n Y_i(u)\exp(x_i^{\mathrm{t}}\beta)z_i^{\mathrm{t}}\,\mathrm{d}H_0(u)}{nG_n^{(0)}(u,\beta,0)}\eta + o_p(n^{-1})\Big]$$
$$\times\Big\{1 - E_{n,S}(u,\beta,0)^{\mathrm{t}}\begin{pmatrix}\widehat\beta_S - \beta\\ \widehat\gamma_S\end{pmatrix}\Big\}.$$

This leads to

$$\begin{aligned} n^{1/2}\{\mathrm{d}\widehat H_{0,S}(u) - \mathrm{d}H_0(u)\} &= n^{-1/2}\frac{\sum_{i=1}^n \mathrm{d}M_i(u)}{G_n^{(0)}(u,\beta,0)} - E_n(u,\beta,0)^{\mathrm{t}}\begin{pmatrix}\sqrt{n}(\widehat\beta_S - \beta)\\ \sqrt{n}\widehat\gamma_S\end{pmatrix}\mathrm{d}H_0(u)\\ &\quad + G_n^{(0)}(u,\beta,0)^{-1}G_{n,2}^{(1)}(u,\beta,0)^{\mathrm{t}}\,\mathrm{d}H_0(u)\eta + o_p(1)\\ &\to_d \mathrm{d}W(u) - e(u,\beta,0)^{\mathrm{t}}\begin{pmatrix}B_S\\ C_S\end{pmatrix}\mathrm{d}H_0(u) + e_2(u,\beta,0)^{\mathrm{t}}\eta\,\mathrm{d}H_0(u), \end{aligned}$$

which proves the lemma. ■

Proof of Lemma 3. It is helpful to translate B_S and C_S to other representations that better reveal the underlying bias and variance balance. Going back to the proof of Lemma 1, let us write

$$\begin{pmatrix}\sqrt{n}U_n(\beta,0)\\ \sqrt{n}V_n(\beta,0)\end{pmatrix} \to_d \begin{pmatrix}J_{01}\eta\\ J_{11}\eta\end{pmatrix} + \begin{pmatrix}U\\ V\end{pmatrix}, \quad \text{where } \begin{pmatrix}U\\ V\end{pmatrix} \sim \mathrm{N}_{p+q}(0, J_{\mathrm{full}}).$$

Let next $V' = J^{10}U + J^{11}V = K(V - J_{10}J_{00}^{-1}U)$. One may show that

$$V' \sim \mathrm{N}_q(0,K) \quad \text{independently of} \quad U \sim \mathrm{N}_p(0, J_{00}).$$

Further algebra leads to

$$\begin{aligned} B_S &= (J^{00,S}J_{01} + J^{01,S}\pi_S J_{11})\eta + J^{00,S}U + J^{01,S}V_S\\ &= J_{00}^{-1}J_{01}(I - \Omega_S)\eta + J_{00}^{-1}U - J_{00}^{-1}J_{01}\pi_S^{\mathrm{t}}K_S\pi_S(V - J_{10}J_{00}^{-1}U)\\ &= J_{00}^{-1}J_{01}(I - \Omega_S)\eta + J_{00}^{-1}U - J_{00}^{-1}J_{01}\Omega_S V', \end{aligned}$$

while similarly

$$C_S = (J^{10,S}J_{01} + J^{11,S}\pi_S J_{11})\eta + J^{10,S}U + J^{11,S}V_S = K_S\pi_S K^{-1}(\eta + V').$$

We may also use this in conjunction with Lemma 2 to derive an alternative expression for $A_S(t)$, which better identifies the bias and variance parts.

We are now in a better position to work with $\widehat\mu_S$. It is not difficult to derive representation (5.3) from Lemmas 1 and 2, via the delta method. Using the expressions found for B_S and C_S, we may isolate the bias and the random parts of Λ_S. The random part may after some algebra be expressed as

$$(\tfrac{\partial\mu}{\partial H_0})W(t) + \{\tfrac{\partial\mu}{\partial\beta} - \tfrac{\partial\mu}{\partial H_0}F_0(t)\}^{\mathrm{t}}J_{00}^{-1}U + (\kappa - \omega)^{\mathrm{t}}\Omega_S V',$$

with variance as indicated above. Similarly the non-random parts may be collected together and expressed as $b_S^{\mathrm{t}}\eta$, where one indeed finds $b_S = (I - \Omega_S^{\mathrm{t}})(\omega - \kappa)$. ■

Proof of Lemma 4. An essential ingredient in our proof is that the earlier process convergence result $A_{n,S}(t) = \sqrt{n}\{\widehat{H}_{0,S}(t) - H_0(t)\} \to_d A_S(t)$, from Lemma 2, implies

$$\Gamma_{n,S}(u) = \sqrt{n}\{\widehat{H}_{0,S}^{-1}(u) - H_0^{-1}(u)\} \to_d \Gamma_S(u) = -\frac{A_S(H_0^{-1}(u))}{h_0(H_0^{-1}(u))}$$

as a process in u, in the Skorokhod topology over each compact interval. That this is true follows from general inversion results and techniques presented and discussed in Doss and Gill (1992) and Burr and Doss (1993).

We now work with

$$\begin{aligned}\widehat{\xi}_S - \xi_{\mathrm{true}} &= \widehat{H}_{0,S}^{-1}(f(\widehat{\beta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})) - H_0^{-1}(f(\beta, \eta/\sqrt{n})) \\ &= \widehat{H}_{0,S}^{-1}(f(\widehat{\beta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})) - H_0^{-1}(f(\widehat{\beta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})) \\ &\quad + H_0^{-1}(f(\widehat{\beta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})) - H_0^{-1}(f(\beta, 0)) + H_0^{-1}(f(\beta, 0)) - H_0^{-1}(f(\beta, \eta/\sqrt{n})),\end{aligned}$$

which decomposes our $\Lambda_{n,S}$ into three different sources of variation. With proper Taylor expansion arguments one finds

$$\begin{aligned}\Lambda_{n,S} &= \Gamma_{n,S}(f(\widehat{\beta}_S, \widehat{\gamma}_S)) + (H_0^{-1})'(f(\beta, 0))\{f(\widehat{\beta}_S, \widehat{\gamma}_S) - f(\beta, 0)\} \\ &\quad - (H_0^{-1})'(f(\beta, 0))\{f(\beta, \eta/\sqrt{n}) - f(\beta, 0)\} + o_p(1),\end{aligned}$$

which by Lemmas 1 and 2 must have a limit distribution with representation

$$\begin{aligned}\Lambda_S &= \Gamma_S(f(\beta, 0)) + (H_0^{-1})'(f(\beta, 0))\{(\tfrac{\partial f}{\partial \beta})^{\mathrm{t}} B_S + (\tfrac{\partial f}{\partial \gamma_S})^{\mathrm{t}} C_S\} - (H_0^{-1})'(f(\beta, 0))(\tfrac{\partial f}{\partial \gamma})^{\mathrm{t}}\eta \\ &= h_0(\xi_0)^{-1}[-A_S(\xi_0) + \{(\tfrac{\partial f}{\partial \beta})^{\mathrm{t}} B_S + (\tfrac{\partial f}{\partial \gamma_S})^{\mathrm{t}} C_S\} - (\tfrac{\partial f}{\partial \gamma})^{\mathrm{t}}\eta].\end{aligned}$$

In particular, the factor $h_0(H_0^{-1}(f(\beta, 0)))$ enters each of the contributions here, and will be of no consequence when the task is to compare limiting risk for different subset models S.

The statement of Lemma 4 follows from these results, upon using representations for B_S and C_S arrived at in the course of proving Lemma 3 above. ■

Proof of Lemma 5. There is joint convergence in distribution of all $(\Lambda_{n,S}, \widehat{\eta}, w_n(S \mid \widehat{\eta})$ to that of $(\Lambda_S, D, w(S \mid D))$, as can be seen from previous proofs. Also, from representation (5.5), $\Lambda_S = \Lambda_0 + (\omega - \kappa)^{\mathrm{t}}(\eta - \Omega_S D)$. The statement of the lemma follows from this. See also corresponding discussion in HC (2003a, 2003b). ■

Acknowledgements. The authors gratefully acknowledge hospitality and partial funding from the Department of Mathematics at the University of Oslo and the Institut de Statistique at Université catholique de Louvain at reciprocal visits. The research of Claeskens has been partly supported by the Belgian Federal Science Policy Office.

References

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1994). *Statistical Models Based on Counting Processes*. Springer-Verlag, Heidelberg.

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large-sample study. *annals* **10**, 1100–1120.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science* **16**, 199–231.

Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model selection: an integral part of inference. *Biometrics* **53**, 603–618.

Bunea, F. and McKeague, I.W. (2004). Covariate Selection for Semiparametric Hazard Function Regression Models. *Journal of Multivariate Analysis* (in press).

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer, New York.

Burr, D. and Doss, H. (1993). Confidence bands for the median survival time as a function of the covariates in the Cox model. *Journal of the American Statistical Association* **88**, 1330–1340.

Claeskens, G. and Hjort, N.L. (2003). The focused information criterion [with discussion]. *Journal of the American Statistical Association* **98**, 900–916.

Clyde, M. (1999). Bayesian model averaging and model search strategies [with discussion]. In *Bayesian Statistics VI* (J.M. Bernardo, A.P. Dawid, J.O. Berger and A.F. Smith, eds.), Oxford University Press, 157–185.

Dabrowska, D.M. and Doksum, K. (1987). Estimates and confidence intervals for median and mean life in the proportional hazard model. *Biometrika* **74**, 799–807.

Doss, H. and Gill, R.D. (1992). An elementary approach to weak convergence for quantile processes, with applications to censored survival data. *Journal of the American Statistical Association* **87**, 869–877.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). *Annals of Statistics* **32**, 407–499.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* **30**, 74–99.

Gill, R.D. (1984). Understanding Cox's regression model: a martingale approach. *Journal of the American Statistical Association* **79**, 441–447.

Gould, S.J. (1995). The median isn't the message. In *Adam's Navel and other essays*, Penguin Books, 1995.

Hjort, N.L. (1992). On inference in parametric survival data models. *International Statistical Review* **40**, 355–387.

Hjort, N.L. and Claeskens, G. (2003a). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association* **98**, 879–899.

Hjort, N.L. and Claeskens, G. (2003b). Rejoinder to the discussion of the FIC and FMA articles. *Journal of the American Statistical Association* **98**, 938–945.

Hjort, N.L. and Pollard, D.B. (1996). Asymptotics for minimisers of convex processes. Statistical Research Report, University of Oslo.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial [with discussion]. *Statistical Science* **19**, 382–417. [A version where the number of misprints has been significantly reduced is available at `http://www.stat.washington.edu/raftery/`.]

Raftery, A.E., Madigan, D. and Volinsky, C.T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance [with discussion]. In *Bayesian Statistics 5* (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), Oxford University Press, 323–350.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Volinsky, C., Madigan, D., Raftery, A.E. and Kronmal, R.A. (1997). Bayesian model averaging in proportional hazard models: Predicting the risk of a stroke. *Applied Statistics* **46**, 443–448.

Volinsky, C. and Raftery, A.E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256–262.