WINDSOR

MATHEMATICS

STATISTICS

REPORT

Nonparametric Shrinkage estimation for Aalen’s additive hazards model

by

A. A. Hussein*, S. Nkurunziza* and K. Tomanelli*
*Department of Mathematics and Statistics
University of Windsor
Windsor, ON, N9B 3P4, Canada

WMSR # 12-02
July 17, 2012
Nonparametric Shrinkage estimation for Aalen’s additive hazards model

A. A. Hussein¹, S. Nkurunziza¹ and K. Tomanelli¹

¹Department of Mathematics and Statistics, University of Windsor, 401 Sunset Avenue, Windsor, ON, N9B 3P4, Canada

Abstract

Aalen’s nonparametric additive model in which the regression coefficients are assumed to be unspecified functions of time is a flexible alternative to the Cox’s proportional hazards model when the proportionality assumption is doubtful. In this manuscript, we incorporate a general linear hypothesis into the estimation of the time-varying regression coefficients. We combine unrestricted least squares estimators and estimators that are restricted by the linear hypothesis and produce James-Stein-type shrinkage estimators of the regression coefficients. We develop the asymptotic joint distribution of such restricted and unrestricted estimators and use it for studying the relative performance of the proposed estimators via their asymptotic integrated distributional risks. We conduct Monte Carlo simulations to examine relative performance of the estimators in terms of their integrated mean square errors. We also compare the performance of the proposed estimators to a re-

June 30, 2012
cently devised $l_1$-penalized estimators in the context of data on the survival of primary biliary cirrhosis patients.

**Keywords:**

1. Introduction

Survival analysis is a branch of statistics which deals with the analysis of time-to-event (or in general event history). Applications of event history analysis methodologies are numerous in the medical field, but are also found in economics, engineering and sociology. Usually, the investigator would collect data on the occurrence times of a certain event of interest along with a set independent predictor variables (covariates) such as gender, age, social status, biomarkers of diseases and similar variables. The investigator would then desire to know if and how such covariates influence the occurrence rates (intensity) of the event of interest. The Cox’s proportional hazards (PH) model is one of the earliest and perhaps the most used statistical model which attempts to address such questions. The Cox’s original model was later formulated in terms of counting process theory and named multiplicative intensity model (Andersen and Gill (1982)). The multiplicative intensity model (MI) has extended the PH model in the sense that it allowed multiple events and time varying covariate processes. The PH model and its vari-
ants assume that the intensity function of the counting process defined by the events of interest is made up of the product of a baseline nonparametric intensity function and a parametric part consisting of a function of linear combination of the independent variables. The effect of the covariates are measured through the unknown coefficients of the linear combination (regression parameters) which do not depend on time. This entails that the hazard ratio of two individuals differing just by the level of a given covariate is constant over time. This property, known as the proportional hazards assumption, gives an attractive interpretation in terms of risk ratio and is mathematically tractable. However, such assumption is sometimes not satisfied by the data at hand. A remedy to this is the use of time varying regression coefficients. As an alternative to the MI models, Aalen (1993) proposed an additive regression model whereby the intensity function is governed by the covariates as well as the past events through a linear regression with time-varying coefficients. Estimation of Aalen’s nonparametric time-varying regression coefficients are performed via weighted least squares and their asymptotic properties are studied by using the martingale theory for counting processes (Martinussen & Scheike (2006)).

Often there is a large number of potential covariates to be included in a
regression model. However, it may turn out that only a handful of such covariates is relevant to explain the outcome of interest. Investigators usually employ either their own judgement or a model selection mechanism in order to filter out most of the unimportant covariates and obtain a parsimonious and slim final model. Despite their appeal, model selection methodologies have the disadvantage of introducing biases due to the fact that the discarded covariates may not be completely irrelevant. Therefore, the question of whether to settle for reduced (uncertain) model or a full (possibly inefficient) model remains open. A way out is to construct James-Stein-type shrinkage estimators which incorporate both models into the estimation process. In the classical linear and partially linear regression models and in censored data models, the shrinkage estimators are known to dominate the unrestricted estimators (based on the full model) over the whole parameter space and dominate the restricted estimators (based on the linear hypothesis) except in a small neighborhood of the linear restriction (Ahmed et. al (2007), Raheem et. al (2012), Nkurunziza and Ahmed (2011)). It is important to note that the regression coefficients in this context can be seen as infinite dimensional parameters which makes Aalen’s model a high-dimensional model by nature.
In this manuscript, we propose James-Stein-type shrinkage estimators for the nonparametric regression coefficients in the Aalen’s additive model under a general linear hypothesis about the coefficients. We study the asymptotic joint distribution of the restricted and unrestricted estimators of such coefficients via the martingale central limit theorems. Consequently, we define integrated distributional quadratic risks of the proposed shrinkage estimators and compare them analytically to those of the restricted and unrestricted estimators. We then take the task of comparing the performance of the estimators via Monte Carlo simulations. We also compare the performance of the proposed estimators to a recent $l_1$-penalized estimators in the context of data on the survival of primary billiary cirrhosis patients.

The manuscript is organized as follows. In Section 2, we introduce the Aalen’s additive model and define a general linear hypothesis about the regression coefficients. We provide restricted estimators of the cumulative coefficients, and study, the joint asymptotic normality of the restricted and unrestricted estimators. We then define James-Stein-type shrinkage estimators of the cumulative coefficients. We define and study the integrated distributional quadratic risks of the proposed shrinkage estimators and compare them to those of the restricted and unrestricted estimators. In Section 3, we
conuct Monte Carlo simulations examining the performance of the estimators. Furthermore, we compare the performance of the proposed estimators to \( l_1 \)-penalized estimators in the context of data on the survival of primary biliary cirrhosis patients. Section 4 gives concluding remarks and proofs of technical results are outlined in the Appendix.

2. The proposed methodology

2.1. Aalen's additive model and the unrestricted estimator

Event history data usually come in the form of a triplets \([N_i(t), Y_i(t), X_i(t)]\) for a sample of individuals \( i = 1, \ldots, n \), where \( N_i(t) \) is the number of events up to time \( t \), \( X_i(t) \) is a vector of \( k \) covariates. For mathematical technicality, let \( \{ \mathcal{F}_{it}, t \geq 0 \} \), be the family of \( \mathcal{G} \)-fields generated by the processes \( \{ N_i(t); t \geq 0 \} \) and define \( \mathcal{F}_t \) as the \( \mathcal{G} \)-field generated by \( \bigcup_{i=1}^n \mathcal{F}_{it} \). The natural filtration of any counting process is right-continuous, so we can state that \( \{ \mathcal{F}_t, t \geq 0 \} \) is a right-continuous filtration.

In this paper, we assume that \( \{(X_i(t), Y_i(t)); t \geq 0\} \) is a process adapted to the filtration \( \{ \mathcal{F}_t; t \geq 0 \} \). Further, for simplicity of notations, let us redefine the vector of covariate to incorporate the risk indicator functions, and organize them in a design matrix, \( X(t) = (Y_1(t)X_1(t), \ldots, Y_n(t)X_n(t))' \).
Further, the derivation of the main results given in this paper involve operations such as supremums and infimums of an uncountable collection of random variables. Thus, in order to guarantee that the resulting quantities are random variables, we assume that all processes under consideration are separable. It should be noted that this restriction is without loss of generality, and this is just for technical consideration. Indeed, it is well known that every stochastic process is equivalent to a separable process Billingsley (1995, p. 531)).

To define some other notations, we first recall that a metric space-valued function is called càdlàg if it is right-continuous with left limits. Thus, let $\mathcal{D}([0, \tau], \mathbb{R}^p)$ denote the space that consists of càdlàg functions on $[0, \tau]$ into $\mathbb{R}^p$ and endowed with the sup-norm Skorohod topology (see Billingsley, 1995). For the sake of simplicity, we let $\mathcal{D}([0, \tau])$ stand for $\mathcal{D}([0, \tau], \mathbb{R}^p)$.

Following Martinussen & Scheike (2006), Aalen’s nonparametric additive regression model is defined through the intensity function of the counting process, $N(t)$, as follows

$$\lambda(t) = X'(t)\beta(t).$$

(2.1)

In general, a major objective is to estimate the cumulative regression
coefficients defined by

$$B_j(t) = \int_0^t \beta_j(u) du \quad \text{for} \quad j = 1, \ldots, k.$$ 

As proposed in Aalen (1993), this is achieved by using the least squares estimators,

$$\hat{B}(t) = \int_0^t X^{-1}(s)dN(s), \quad (2.2)$$

where $X^{-1}(t) = (X'(t)W(t)X(t))^{-1}X'(t)W(t)$ is a generalized inverse and $X'(t)W(t)X(t)$ is assumed to be a full rank matrix, for some predictable diagonal $k \times k$ weight matrix $W(t)$. We call this estimators, the Unrestricted estimator. Under some regularity conditions (Martinussen & Scheike (2006)),

$$\sqrt{n}(\hat{B}(t) - B(t)) \overset{D}{\rightarrow} U \text{ on } D([0, \tau]),$$

where $U$ is a mean zero Gaussian martingale with covariance function

$$\Phi(t) = \int_0^t \Omega(s)E[W_1^2(s)X_1'(t)X_1(t)X_1'(t)\beta(s)]\Omega(s)ds,$$

and $\Omega^{-1}(t) = E[W_1(t)X_1'(t)X_1(t)]$.

2.2. The restricted estimator

Often, investigators have some idea about the importance of the covariates in the model in the sense of suspecting that a reduced model whereby some of the covariates are set to zero would be preferable. Here we define a
general linear hypothesis about the cumulative coefficients in the form

\[ H_1 : R\beta(t) = r_1(t) \iff H_0 : RB(t) = r_2(t) = \int_0^t r_1(s)ds \]

for \(0 \leq t \leq \tau\). By using Lagrange multiplier, it is easy to derive the restricted estimator of the cumulative coefficients under the above linear hypothesis as,

\[ \tilde{B}_R(t) = \int_0^t (I_k - A_n(s)R)\hat{b}_{LS}(s)ds + \int_0^t A_n(s)r_1(s)ds, \quad (2.3) \]

where \(\hat{b}_{LS}(t) = [X(t)W(t)X'(t)]^{-1}X(t)W(t)Y(t)\) and

\[ A_n(t) = [X(t)W(t)X'(t)]^{-1}R'[R[X(t)W(t)X'(t)]^{-1}R']^{-1}. \quad (2.4) \]

We show the joint normality of the restricted and unrestricted estimators under a sequence of local alternatives of the form,

\[ H_{2,n} : RB(t) = r_2(t) + \frac{\delta_2(t)}{\sqrt{n}} \quad (2.5) \]

where \(R\) is a known \(q \times k\) full-rank matrix, and \(r_1(t)\) is a known function of time. Under the same sequence of local alternatives, we also provide an asymptotic noncentral \(\chi^2\) quantity useful for constructing a test statistic for \(H_1\). Let \(\eta_n(t) = \sqrt{n}(\tilde{B}_R(t) - B(t)), \xi_n(t) = \sqrt{n}(\tilde{B}_R(t) - \hat{B}_{LS}(t))\), and

\[ J(t) = R' \left( \int_0^t R\Omega(s)R'ds \right)^{-1} R, \quad \delta^*(t) = \int_0^t \Omega(s)R'[R\Omega(s)R']^{-1}\delta(s)ds, \]

\[ \Sigma_{11}(t) = \int_0^t \Omega(s)R'[R\Omega(s)R']^{-1}R\Omega(s)ds \]

\[ \Delta(t) = \left( \int_0^t \delta^*(s)ds \right) \times R' \left( \int_0^t R\Omega(s)R'ds \right)^{-1} R \times \left( \int_0^t \delta^*(s)ds \right). \]
Proposition 2.1. Under condition $A$ in the appendix and for the sequence of local alternatives given above, we have
\[(\xi'_n(t), \eta'_n(t))' \overset{D}{\to} n \to \infty (\xi'(t), \eta'(t))',\]
on $\mathcal{D}([0, \tau])$, where \{$(\xi'(t), \eta'(t))', t \geq 0$\} is the Gaussian martingale with
\[(\xi'(t), \eta'(t))' \sim N_{2k} \left[ \int_0^t \begin{pmatrix} I_k \\ I_k \end{pmatrix} A(s) \delta_1(s) ds, \Phi^{***}(t) \right],
\]
\[\Phi^{***} = \int_0^t \begin{pmatrix} A(s) R \Omega(s) \\ 0 \\ 0 \end{pmatrix} ds,
\]
for $0 \leq t \leq \tau$, with $A(t) = \Omega(t) R' [R \Omega(t) R']^{-1}$. Furthermore,
\[\varphi_n(t) = \xi'_n(t) \tilde{J}(t) \xi_n(t) \sim \chi^2_q(\Delta(t)), \quad 0 \leq t \leq \tau,
\]
where $q = \text{rank}(\Sigma_{11}(t) J(t) \Sigma_{11}(t))$, $\tilde{J}(t)$ is a consistent estimator for $J(t)$, uniformly on $[0, \tau]$.

\[\square\]

2.3. The Shrinkage estimators and their asymptotic performance

Now we are in a position to define the proposed James-Stein-type estimators as follows,
\[\tilde{B}^{S}(t) = \tilde{B}_R(t) + (1 - c \varphi^{-1}_n(t))(\tilde{B}_{LS}(t) - \tilde{B}_R(t)) \quad (2.7)
\]
where $c$, which is known as the shrinkage constant, is chosen in an interval such that $\tilde{B}^{S}(t)$ dominates $\tilde{B}_{LS}(t)$, and $\varphi_n(t)$ is defined in (2.6). In the sequel, we will consider $c = q - 2$. 

10
The shrinkage estimator tends to overshrink the estimator, especially when \( \phi_n(t) \) is very small in comparison with \( c \). To remedy this issue, the *Positive-part Shrinkage (PS) Estimator* was developed by truncating the shrinkage estimator in the following way:

\[
\hat{B}^+(t) = \tilde{B}_R(t) + \max(0, 1 - c\phi_n^{-1}(t)) (\hat{B}(t) - \tilde{B}_R(t))
\]

\[
= \tilde{B}_R(t) + \max \left( \left[ 1 - \frac{q - 2}{\phi_n(t)} \right], 0 \right) (\hat{B}(t) - \tilde{B}_R(t)). \tag{2.8}
\]

It is common to compare shrinkage estimators in terms of their asymptotic distributional risk functions (ADR). Here we introduce a new risk based on quadratic integrated loss function, which we call integrated asymptotic distributional risk (IADR). We define the IADR of an estimator \( \hat{B}^* \) as

\[
IADR(\hat{B}^*, B; W^*) = E[\Psi], \quad \text{where } \Psi \text{ is the distributional limit of the loss function}
\]

\[
L(\hat{B}^*, B; W^*) = n \int_0^\tau (\hat{B}^*(s) - B(s))^\prime W^*(s)(\hat{B}^*(s) - B(s)) ds, \tag{2.9}
\]

in which \( W^*(s) \) is a known weight matrix. In particular, if this weight matrix is identity, the risk reduces to the usual integrated mean squared error. In the following proposition we state the risk dominance of the proposed shrinkage estimators with respect to the restricted and unrestricted estimators. The proof of the proposition is lengthy and hence omitted, but with the help of
the joint asymptotic normality of the restricted and unrestricted estimators given in the last proposition, one can easily work out the proof by using the fact that the integral of any nonnegative integrable function is nonnegative, and completes it by following the same lines as for example in Saleh (2006), Ahmed et al (2007), Nkurunziza (2010), and Nkurunziza and Ahmed (2011).

**Proposition 2.2.** Suppose that Conditions $A$, and the sequence of local alternatives in (2.5) hold. Also, suppose $W^*(t)$, is an element of the set

$$
\{W^*(t) : 0 \leq (q + 2)\text{Ch}_{\max}(W^*(t)\Sigma_{11}(t)) \leq 2\text{trace}(W^*(t)\Sigma_{11}(t)), t \geq 0\}.
$$

Then,

$$
\text{IADR}(\hat{B}^S, B, W^*) \leq \text{IADR}(\tilde{B}^S, B, W^*) \leq \text{IADR}(\tilde{B}_{LS}, B, W^*),
$$

where $\text{Ch}_{\max}(A)$ denotes the largest eigenvalues of the matrix $A$.

\[\square\]

3. Empirical Studies

3.1. Simulation study

In this section we study the performance of the proposed shrinkage estimators by using Monte Carlo simulations. To this end, we consider a simple survival model whose intensity function for the $i$th individual is given by

$$
\lambda(t|X_i(t)) = Y_i(t)X_i\beta(t)
$$

where we set $\beta(t) = (\beta_0 t, \beta_2 t, ..., \beta_4 t)$ where
\( \beta_q \) for \( q = 0, \ldots, 4 \) are unknown but constants and the covariate process is time independent. Thus, we assumed the time-dependent regression coefficients to be linear. This leads to a cumulative intensity functions given by \( \Lambda(T) = T^2X_i\beta \) where \( \beta = (\beta_0, \ldots, \beta_4)/2 \). This enables us to generate random times, \( T \), via uniform(0, 1) numbers, \( U \), by inverting the relationship, \( 1 - F(T) = \exp(-\Lambda(T)) \). We generated the covariates \( X_{i1}, \ldots, X_{i4} \) from uniform(0, 20) and we set \( \beta = c(2, 0, 0, 0, 0) \), under the null hypothesis and \( \beta = c(2, 0, 0, 0, 0 + \delta) \) under the alternative hypothesis where \( \delta \) varied from zero to one with steps of .05. Thus generated random survival times were then censored by using independent random variates generated from Uniform(0, 3). This setting led to censoring rates varying from 5-15%. Each scenario was simulated 1000 times for sample sizes of \( n = 250, 500, 750, 1000 \). In each scenario, we computed the empirical mean squared errors (MSE) of the all the estimators (shrinkage, positive shrinkage, restricted and unrestricted) and by taking the unrestricted estimator as benchmark, we reported the ratios of these MSEs relative to the benchmark. The results are summarised in Figures 1.

It is clearly seen that the proposed shrinkage estimators outperform the usual restricted and unrestricted estimators on almost all of the parameter
space. When the null hypothesis is true (in other words, $\delta = 0$), we see that the best estimator is the restricted estimator as foreseen from the analytic results, while its performance deteriorates substantially when we get away from the null space. On the other hand, the positive shrinkage estimator dominates the unrestricted estimator throughout the null and alternative space and converges to it in terms of MSE for all the sample sizes considered. However, the shrinkage estimator seems to be worse than the unrestricted at the null hypothesis for sample sizes that are smaller than $n = 1000$. Since the problem of estimating the nonparametric regression coefficients is high-dimensional by nature, the joint asymptotic normality requires quite large samples to kick in.

3.2. Application and comparison with LASSO

In this section, we apply the proposed shrinkage estimation strategies to the famous PBC data from the trial in primary biliary cirrhosis (Fleming and Harrington, 1991) for the purpose of comparing its performance to a recently proposed $l_1$ penalized estimator for the additive model with constant regression coefficients (Gorst-Rasmussen and Scheike (2012)). The PBC clinical trial was a randomized placebo-controlled trial of the drug D-penicillamine, and there were 424 patients in total at the clinic who were eligible to par-
Figure 1: Ratios of MSE for all estimators relative to the unrestricted estimator with delta varying over the parameter space.
participate. The data is mostly complete for the first 312 patients, however the last 112 did not participate in the clinical trial, consenting only to have measurements recorded and followed for survival. In addition to the treatment indicator, there were a total of 16 covariates collected from the patients who participated in the clinical trial, including age, sex, and various biomarkers. The outcome of interest was the patient’s survival time. Applying the gradient descent algorithm of Gorst-Rasmussen and Scheike, we produced LASSO estimators for the 17 covariates by using based a 10-fold cross-validation for choosing the optimal tuning parameter. The overall mean squared error in estimating the 17 time-invariant coefficients was computed by using 1000 bootstrap samples and found to be 15.52 relative to the overall MSE of the vector of the unrestricted estimators. On the other hand, to make our comparison on equal footing, we produced our shrinkage estimators based on the linear hypothesis $H_1 : \beta = 0$, which assumes no specific knowledge about the coefficients. The relative (to the unrestricted) mean squared errors of the shrinkage and positive shrinkage estimators were found to be 1.29, 1.53 respectively. This indicates, at least in the context of the example, that albeit its simplicity, the proposed positive shrinkage estimator is as good as estimators based on the LASSO.
4. Conclusion

The problem of estimating the nonparametric regression coefficients in Aalen-additive model is high-dimensional by nature. In this manuscript, we have proposed estimators which combine reduced and full model estimators of the Aalen’s additive hazards regression coefficients. The proposed estimators are computationally inexpensive and provide a performance that is better than estimators based on either of the reduced and full models and comparable to that of the recently proposed LASSO estimators.
References


Appendix A. Proofs and regularity conditions

Condition $A$

i) $\{(X_i(t), N_i(t)) , \ 0 \leq t \leq \tau \}$ for $i = 1, ..., n$ are i.i.d

ii) $E \left[ \sup_{0 \leq t \leq \tau} |W_1^2(t)X_1j(t)X_1s(t)X_1l(t)| \right] < \infty$ for $j, s, l = 1, ..., k$

iii) $E \left[ \sup_{0 \leq t \leq \tau} |W_1(t)X_1j(t)X_1s(t)| \right] < \infty$ for $j, s = 1, ..., k$

iv) $X'(t)W(t)X(t)$ is a positive-definite matrix, $\forall \ t \in [0, \tau]$

v) $E[W_1(t)X_1'(t)X_1(t)]$ is non-singular $\forall \ t \in [0, \tau]$

vi) $\int_0^t |\beta(s)| ds < \infty$, $\forall \ t \in [0, \tau]$.

vii) $W_i(t) = \frac{1}{X_i(t)\beta(t)}$.

Lemma Appendix A.1. Under condition $A$ and for the sequence of local alternatives in (2.5),

\[
\eta_n(t) \xrightarrow{D, n \to \infty} U^*
\]

on $\mathcal{D}([0, \tau])$, where $U^*$ is a Gaussian martingale such that

\[
U^*(t) \sim \mathcal{N}_k \left( -\int_0^t A(s)\delta_1(s) ds, \Phi^*(t) \right)
\]

for all $0 \leq t \leq \tau$, with

\[
\Phi^*(t) = \int_0^t [I - A(s)R]\Omega(s)\kappa_1(s)\Omega(s) [I - R' A'(s)] ds.
\]

□
Proof. We have \( \eta_n(t) = P_{1,n}(t) + P_{2,n}(t) - P_{3,n}(t) \), where

\[
\begin{align*}
P_{1,n}(t) &= n^{-\frac{1}{2}} \int_0^t [I_k - A_n(s)R] \Omega(s)X(s)W(s)dM(s), \\
P_{2,n}(t) &= n^{-\frac{1}{2}} \int_0^t [I_k - A_n(s)R] \left\{ (n^{-1}\Gamma(s))^{-1} - \Omega(s) \right\} X(s)W(s)dM(s), \\
P_{3,n}(t) &= \int_0^t A_n(s)\delta_1(s)ds.
\end{align*}
\]

Also, by using the Robolledo's Martingale Central Theorem, one can prove that \( \{P_{1,n}(t), t \geq 0\} \) converges in distribution to a Gaussian martingale on \( D([0,\tau]) \), with covariance function \( \Phi^*(t) \). Further, one can verify that \( P_{2,n}(t) \) converges in probability to 0, uniformly on \([0,\tau]\).

Finally, by some algebraic computations, one can verify that

\[
P_{3,n}(t) \xrightarrow{P_{n \to \infty}} \int_0^t A(s)\delta_1(s)ds
\]

uniformly over \([0,\tau]\), and this completes the proof.

\[
\square
\]

Proof of Proposition 2.1. Notice that one can rewrite \( (\xi_n^t, \eta_n^t)' \) as

\[
(\xi_n^t, \eta_n^t)' = \sqrt{n} \int_0^t (-R'A_n'(s), I_k - R'A_n'(s))' \Omega(s)X(s)W(s)dM(s)
\]

\[
+ \sqrt{n} \int_0^t (-R'A_n'(s), I_k - R'A_n'(s))' \left\{ (n^{-1}\Gamma(s))^{-1} - \Omega(s) \right\} X(s)W(s)dM(s)
\]

\[
+ \int_0^t (I_k, I_k)' A_n(s)\delta_1(s)ds.
\]
Then, the proof is completed by following the same steps as used in proof of the above lemma.