SUPER-EFFICIENT PREDICTION BASED ON HIGH-QUALITY MARKER INFORMATION

By

Jens Perch Nielsen

Codan, Gammel Kongevej 60, 1799 Copenhagen, Denmark

Abstract

Nielsen (1999) showed the surprising fact that a nonparametric one-dimensional hazard as a function of time can be estimated $\sqrt{n}$-consistently if a high quality marker is observed. In this paper we show that the hazard relevant for predicting remaining duration time, given the current status of a high quality marker, can be estimated $\sqrt{n}$-consistently if a Markov type property holds for the high quality marker.

Keywords and Phrases

Counting Process, Hazard, Kernel, Marker, Nonparametric estimation, Prediction, Survival analysis, Asset-Liability management, Datamining, High frequency data.

1. Introduction

Prediction of future events is important in many fields (e.g. biostatistics, actuarial science, finance and economics). For example in prevalent data studies, one often wishes to be able to make predictions based on the available information. In this paper we consider a general way of modelling a marker process and a hazard such that good prediction power remains. We do not have any parametric assumptions of the marker process or the hazard, but we do assume that the marker process obeys a Markov property and that the marker process is of high quality as defined in Nielsen (1999) that estimates the traditional deterministic hazard rate as a function of time. This estimation is improved through the knowledge of a high quality marker and the resulting procedure is more efficient than traditional approaches to nonparametric hazard estimation that do not use this extra information, see Nielsen (1998) for an overview of some of these more traditional kernel hazard estimators. The present paper is concerned about the estimation of the future hazard given the

ASTIN BULLETIN, Vol. 30, No. 2, 2000, pp. 295-303

current state of a continuous marker. The marker therefore enters as an integrated part of the model and the stochastic future hazard that we aim at estimating. Therefore the concept of high quality markers seems to even more important in the current study than in Nielsen (1999).

A lot of actuarial work is about modelling, in which the important thing is to make the right approximation to the problem at hand. Statistical testing is not always appropriate since the actuary knows very well that his model is not true. The actuary has another consideration – how can I make a very general model that is easy to understand and that gives me appropriate predictions for the future? The modelling in this paper can be considered as a procedure following this actuarial tradition. Though testing procedures of the validation of the model indeed need to be developed, the most important fact is that here is a rather general model that gives good prediction power. Our postulate of good predicting power is based on the fact that our non-parametrically based estimator of the future hazard given current marker conditions is estimated $\sqrt{n}$-consistently. This is in general not possible without parametric assumptions, the surprising fact that it holds in our situation is a result of the complicated interaction between the model assumptions of the marker process and the hazard function. We illustrate the applicability of the model in § 5 by considering possible applications in so diverse fields as asset-liability management, datamining, biostatistics and the analysis of high frequency financial data.

To get started, consider n individuals with survival times $T_1, ..., T_n$ and marker processes $\{Z_1(s), s \leq T_1\}, ..., \{Z_n(s), s \leq T_n\}$ and suppose we wish to estimate the future hazard based on current marker status. That is, we wish to estimate the marker conditional future hazard (MCFH)

$$h_{y,t_0}(t) = \Pr\{T_i \in (t + t_0, t + t_0 + dt) \mid Z_i(t_0) = y, T_i > t + t_0\}.$$

We make the stability assumption that $h_{y,t_0}(t)$ does not depend on t_0 and write $h_y(t) \equiv h_{y,t_0}(t)$. This assumption is realistic if the marker process is a Markov process and of high quality. A marker is of high quality if the hazard at any given time only depends on the marker, see Nielsen (1999). Under the stability assumption and the assumption that the marker is of high quality, we show that $h_y(t)$ can be estimated $\sqrt{n}$-consistently. In the more precise model formulation below, we allow for filtered data including prevalent data. In § 2 we formulate the two models, the simple time model and the marker-only model. In § 3 the estimators are defined and in § 4 we state the pointwise asymptotic theory of the $\sqrt{n}$-consistent estimator of the hazard as a function of time. We use standard counting process theory to formulate our model and consider n individuals $i = 1, ..., n$ with $N_i^{(n)}$ counting observed failures for the i'th individual in the time interval $[0, T]$. We assume that $\mathbf{N}^{(n)} = (N_1^{(n)}, ..., N_n^{(n)})$ is an n-dimensional counting process with respect to the increasing, right continuous filtration $\mathcal{F}_t = \sigma(\mathbf{N}(s), \mathbf{Z}(s), \mathbf{Y}(s); s \leq t), t \in [0, T]$ where $\mathbf{Y}^{(n)} = (Y_1^{(n)}, ..., Y_n^{(n)})$ is the n-dimensional exposure process, where $Y_i^{(n)}$ is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the i'th individual is under risk. We assume that we get some further marker information $\mathbf{Z}^{(n)} = (Z_1^{(n)}, ..., Z_n^{(n)})$,

https://doi.org/10.2143/AST.30.2.504636 Published online by Cambridge University Press

where $Z_i^{(n)}$ is a d-dimensional, predictable *CADLAG* marker process and that we are in the *high quality marker case*, which Nielsen (1999), defined in the following way: the stochastic intensity process $\lambda^{(n)} = (\lambda_1^{(n)}, ..., \lambda_n^{(n)})$ based on the increased filtration $\mathcal{F}_t = \sigma(\mathbf{N}(s), \mathbf{Z}(s), \mathbf{Y}(s); s \leq t)$ depends only on the value of the marker in the following sense

$$\lambda_i^{(n)}(t) = \alpha\left\{Z_i^{(n)}(t)\right\} Y_i^{(n)}(t).$$

Apart from smoothness assumptions the functional form of the marker-only hazard, α, is assumed to be unknown. We furthermore assume that the marker process have the *Markov property* that

$$F_{y,t,t_0}(z) = \Pr\{Z_i(t + t_0) \leq z \mid Z_i(t_0) = y, T_i > t + t_0\}$$

is independent of t_0 and use the notation

$$F_{y,t}(z) = F_{y,t,t_0}(z).$$

Let $f_{y,t}(z)$ be the density of $F_{y,t}(z)$ with respect to the d-dimensional Lebesgue measure. We also assume the following condition of *unbiased filtering*:

$$F_{y,t}(z) = \Pr\{Z_i(t + t_0) \leq z \mid Z_i(t_0) = y, Y_i(t_0) = Y_i(t + t_0) = 1\}$$

Let also

$$F_t(z) = \Pr\{Z_i(t) \leq z \mid Y_i(0) = 1\}$$

and let $f_t(z)$ be the density of $F_t(z)$ and

$$H_y(t) = \Pr\{Y_i(t + t_0) = 1 \mid Z_i(t_0) = y, Y_i(t_0) = 1\}.$$

We assume that the marker $Z_i(s)$ has support on some compact set $\mathcal{N}$ and that the densities $f_{y,t}$ and the hazard α are uniformly bounded away from zero and infinity.

We assume that $E\{Y_i(s)\} = H(s)$, where $H(\bullet)$ is continuous. The marker $Z_i(s)$ is only observed for those s such that $Y_i(s) = 1$. Let

$$Z_i^*(s) = \begin{cases} Z_i(s) & \text{when } Y_i(s) = 1 \\ -\infty & \text{when } Y_i(s) = 0. \end{cases}$$

We call Z_i^* the observed marker process. We assume that the stochastic processes $(N_1, Z_1^*, Y_n), ..., (N_n, Z_n^*, Y_n)$ are *iid* for the n individuals. Under the above model assumptions a more precise definition of the MCFH is

$$h_y(t) = E[\alpha\{Z_i(t + t_0)\} \mid Z_i(t_0) = y, Y_i(t_0) = Y_i(t + t_0) = 1] = \int \alpha(z) f_{y,t}(z) dz.$$

In the following we will show that under the above assumption h_y can be estimated $\sqrt{n}$-consistently.

https://doi.org/10.2143/AST.30.2.504636 Published online by Cambridge University Press

2. Definition of the estimator

The following estimator of α was introduced and analyzed in Nielsen & Linton (1995) and it was employed by Fusaro et al. (1993) to estimate the risk of Aids given current marker status based on the San Fransisco Mens' Health Study. Let $\widehat{\mathcal{F}}_i(z) = n^{-1}\sum_{k\neq i}\int_0^T K_b\{z - Z_k(s)\}\,Y_k(s)ds$ and

$$\widehat{\alpha}_i(z) = \frac{n^{-1}\sum_{k\neq i}\int_0^T K_b\{z - Z_k(s)\}dN_k(s)}{\widehat{\mathcal{F}}_i(z)},$$

where K is a second order kernel of d dimensions. Both $\widehat{\mathcal{F}}_i$ and $\widehat{\alpha}_i$ are leave-one-out according to the definition given in Nielsen, Linton & Bickel (1998).

We estimate $h_y(t)$ by the empirical version of $\int \alpha(z)f_{y,t}(z)dz$:

$$\widehat{h}_y(t) = \frac{\sum_{i=1}^n \int_0^T \widehat{\alpha}_i\{Z_i(t+s)\}\,Y_i(t+s)\,Y_i(s)K_b\{y - Z_i(s)\}ds}{\sum_{i=1}^n \int_0^T Y_i(t+s)\,Y_i(s)K_b\{y - Z_i(s)\}ds}.$$

Remark. Full knowledge of the marker process is only possible if the marker changes deterministically between observed time points. In many examples, including using CD4 cell counts or other surrogate markers to predict onset of Aids of HIV+ infected individuals, the marker is only observed at discrete timepoints. Therefore the methodology of this paper requires an interpolation and extrapolation technique. Interpolation is done between two observed points in time and extrapolation is performed from a point in time, where no succeeding observation exists within a suitable time, see Fusaro et al. (1993) and Nielsen (1999) for more comments on this issue.

3. Properties of the estimator

Let

$$\rho_{y,t}(z) = \int_0^T f_{y,t}(z)H_y(t)f_s(y)H(s)/\left\{\int_0^T f_u(z)H(u)du\right\}ds,$$

$$g_y(t) = \int_0^T H_y(t)f_s(y)H(s)ds$$

and

$$\widehat{g}_y(t) = n^{-1}\sum_{i=1}^n \int_0^T Y_i(t+s)\,Y_i(s)K_b\{y - Z_i(s)\}ds.$$

Let also

$$X_i = \int_0^T \alpha_i\{Z_i(t+s)\}\,Y_i(t+s)\,Y_i(s)K_b\{y - Z_i(s)\}ds - h_y(t)\widehat{g}_y(t).$$

https://doi.org/10.2143/AST.30.2.504636 Published online by Cambridge University Press

The error of estimation can be written as

$$\begin{aligned}\widehat{h}_y(t) - h_y(t) = &\{\widehat{g}_y(t)\}^{-1} n^{-1} \sum_{i=1}^{n} X_i \\ &+ \{\widehat{g}_y(t)\}^{-1} n^{-1} \sum_{i=1}^{n} \int_0^T \rho_{y,t}\{Z_i(s)\} dM_i(s) \\ &+ \{\widehat{g}_y(t)\}^{-1} \{R_1(t) + R_2(t)\},\end{aligned}$$

where

$$\widehat{\rho}_{i,y,t}(z) = n^{-1} \sum_{k \neq i} \int_0^T K_b\{Z_k(t+s) - z\} K_b\{y - Z_k(s)\} Y_k(s) / \widehat{\mathcal{F}}_k\{Z_k(t+s)\} ds,$$

$$R_1(t) = n^{-1} \sum_{i=1}^{n} \int_0^T \left[\widehat{\rho}_{i,y,t}\{Z_i(s)\} - \rho_{y,t}\{Z_i(s)\}\right] dM_i(s),$$

$$R_2(t) = n^{-1} \sum_{i=1}^{n} \int_0^T \left[(\alpha_i^* - \alpha)\{Z_i(t+s)\}\right] Y_i(t+s) Y_i(s) K_b\{y - Z_i(s)\} ds,$$

where α_i^* corresponds to $\widehat{\alpha}_i$, but with N_i replaced by its compensator Λ_i:

$$\alpha_i^*(z) = \frac{n^{-1} \sum_{k \neq i} \int_0^T K_b\{z - Z_k(s)\} \alpha\{Z_k^{(n)}(s)\} Y_i^{(n)}(s) ds}{\widehat{\mathcal{F}}_i(z)}.$$

The remainder terms $R_1(t)$ and $R_2(t)$ are of lower order of magnitude in probability. The proof of this is omitted since it follows the same lines as the proof of Theorem 1 in Nielsen (1999). The main technical difficulty of this proof is to solve to the so-called predictability issue, see Nielsen, Linton & Bickel (1998), Nielsen (1998) and Nielsen (1999). To get the asymptotic variance of $\widehat{h}_y$, we need to get the asymptotic distribution of the two first terms. Therefore the asymptotic distribution of $\{\widehat{h}_y(t) - h_y(t)\}\{\widehat{g}_y(t)\}^{-1}$ can be calculated as the asymptotic distribution of the sum of independent identically distributed stochastic variables $\sum_{i=1}^{n}(X_i + W_i)$, where $W_i = \int_0^T \rho_{y,t}\{Z_i(s)\} dM_i(s)$. The following theorem is a consequence of the law of large numbers. Notice that it is necessary to undersmooth the preliminary estimator, $\widehat{\alpha}_i$, used to calculate the final estimator. This type of undersmoothing is well known from semiparametric analyses, see Bickel, Klaassen, Ritov, Wellner (1993) and Nielsen et al. (1998). It also enters in marginal integration (one kind of additive regression), see Linton and Nielsen (1995), and it appeared in a similar way in the study of Nielsen (1999). The variance expression corresponding to W_i is a nice expression resulting from a standard counting process martingale calculation. The variance and covariance term corresponding to X_i is, however, not a nice expression. We have chosen a short hand notation for these two latter expressions. The variance can be estimated in a straightforward manner by first estimating the stochastic terms, X_i and W_i, by

https://doi.org/10.2143/AST.30.2.504636 Published online by Cambridge University Press

inserting consistent versions of the unknown quantities and then use these estimators to estimate the empirical variance of the sum of the identically distributed stochastic variables.

Theorem 1. [Pointwise Convergence] *Assume that* α *is twice continuously differentiable and that* $nb^{d+1} \to \infty$ *and* $nb^4 \to 0$. *Assume finally that the kernel* K *satisfies the Lipschitz condition given in Theorem 2 in Nielsen & Linton (1995) in each coordinate. Then*

$$n^{\frac{1}{2}}\{\widehat{h}_y(t) - h_y(t)\} \Rightarrow N(0, \sigma^2_{y,t}),$$

where

$$\sigma^2_{y,t} = \{g_y(t)\}^{-2}(\sigma^2_{1,y,t} + \sigma^2_{2,y,t} + 2\sigma^2_{1,2,y,t})$$

and

$$\sigma^2_{1,y,t} = V(X_i), \sigma^2_{2,y,t} = \int\int_0^T \rho^2_{y,t}(z)\alpha(z)f_s(z)H(s)dsdz,$$
$$\sigma^2_{1,2,y,t} = E(X_i W_i).$$

4. Extensions

The $\sqrt{n}$-consistency of the predicting hazard estimator will remain even if the hazard model is extended to some semiparametric model

$$\lambda_i^{(n)}(t) = \alpha\{Z_i^{(n)}(t)\}m_\theta\{L_i^{(n)}(t)\}Y_i^{(n)}(t),$$

where m_θ is some parametric model specification and $L_i^{(n)}(t)$ is some marker process that can contain continuous as well as descrete markers. A traditional actuarial finite state space Markov model can therefore be included in the model. The estimation of the parameter θ can be performed by a traditional semiparametric analysis, see Nielsen et al. (1998) for an example of a semiparametric hazard estimation technique beyond the traditional cox regression. Also the Markov assumption of the marker process can be relaxed. One can allow for a parametric trend, for example a linear trend and estimate and include the parameters without violating the $\sqrt{n}$-consistency of the predicting hazard estimator.

5. Examples

In this section we specify four examples where the above methodology seems to provide a helpful methodology for prediction. The examples are taken from asset-liability management, datamining, biostatistics and an area of recent interest, namely high frequency transaction data in financial studies.

https://doi.org/10.2143/AST.30.2.504636 Published online by Cambridge University Press

5.1. Asset-Liability management: Predicting the hazard of the remaining time to prepayment of mortgage bonds

Danish prepaid bonds can be a difficulty in asset-liability management of Danish insurance companies, where up to 60% of the invested capital is in this type of bonds. In particular it can be hard to say something precise regarding the expected duration of the bonds. Here we think of duration of bonds in the original sense of Macauley (1938) and of asset-liability management in the sense indicated by Redington (1952). The duration of prepaid bonds is difficult because the payment times are stochastic variables. People might choose to prepay their loan and they might not. One good marker of whether people actually choose to prepay their mortgage is if the market interest rate is considerable below the nominal interest rate. Therefore the difference between the market interest rate and the nominal interest rate is a good candidate for a high quality marker. It is reasonable to assume that this marker obeys the needed Markov property and that the hazard of prepayment can be described from the marker information alone. The procedure of this paper therefore seems valuable when calculating the predicting hazard of prepayment bonds and hereby their expected remaing life time or their duration.

5.2. Datamining: Predicting the hazard of the remaining time a given customer will remain in an insurance company or predicting the remaing time before the next claim

Let us assume that we through datamining have come up with some relevant marker for expected customer loyalty. If this marker is assumed to obey the Markov property of this paper, then the method of this paper can be used to give a precise estimate of the expected customer loyalty in the future. Another example from datamining could be to predict the expected time to the next claim. Continuous measures such as the historical claim intensity for the last three years can be combined with other markers to predict the expected time to the next claim using the methodology of this paper.

5.3. Biostatistics: Predicting the hazard of the remaining time to onset of AIDS in an HIV+ infected individual

In prevalent cohort studies the relevant time is often unknown and surrogate markers can be useful to alleviate this. A well known example is the early datasets on onset of AIDS, see Fusaro et al. (1993) for a description of the San Fransisco Mens' Health Study. In studies of onset of AIDS the predictive power of CD4 cell counts and other surrogate markers for HIV+ infected individuals are well known, see also Choi, Lagakos, Schooley & Volberding (1993) and Nielsen (1999). Fusaro et al. (1993) considered the effect of CD4, CD8 and Beta-2 Microglobulin on onset of AIDS. It does not seem unreasonable to assume that CD4 cell counts, perhaps combined with other surrogate markers, can be used as a marker of high quality. The Markov property assumption of

the markers also seems reasonable. Therefore the methodology of this paper can be used to give quite accurate estimates of the predicting hazard given current marker status. This is an unexpected and quite welcome result for this type of otherwise hard to handle prevalent cohort data sets.

5.4. High frequency financial analysis: Predicting the hazard of time between transactions in financial data

Due to the increasing storage capacity of data and due to the increasing speed of computations, all kinds of financial data is being collected at high frequencies. The methodology described in this paper can be useful for modelling the distribution of time between transactions of financial data. Engle and Russell (1998) describe a completely different approach using time series techniques. In their section 6.1 they analyze the time between transactions for IBM stocks. However, their problem is a problem of duration and traditional duration methods from biostatistics and renewal theory seem to be useful for this type of data. The methodology of this paper can be used to predict remaining time to the next transaction of a stock given the two most important markers, namely the bid-ask spread and the volume, see Engle and Russell (1998) for a definition of these markers. The Markov property of the markers seems to be realistic and it is also realistic to assume that the two markers are of high quality, see Shen and Starr (1998) for some indication of the predictive power of the bid-ask spread on bond returns. In our terminology they argue, that the bid-ask spread can be characterized as a marker for liquidity in the considered financial market. Liquidity in turn is a financial term closely connected to duration of time between transactions. Obviously the traded volume is also important for the duration of time between transactions. The marker consisting of the bid-ask spread and the traded volume is therefore likely to be of high quality in many cases. The predicting hazard of duration time between transactions given the current bid-ask spread and volume can therefore be expected to be estimated $\sqrt{n}$-consistently when using the method of this paper.

REFERENCES

BICKEL, P.J., KLAASSEN, C.A.J., RITOV, Y. and WELLNER, J.A. (1993) *Efficient and adaptive estimation for semiparametric models*. The John Hopkins University Press, Baltimore and London.

CHOI, S., LAGAKOS, S.W., SCHOOLEY, R.T. and VOLBERDING, P.A. (1993) CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. *Ann. Internal Med.* **118**, 674-680.

ENGLE, R.F. and RUSSELL, J.R. (1998) Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrika* **5**, 1127-1162.

FUSARO, R., NIELSEN, J.P. and SCHEIKE, T. (1993) Marker dependent hazard estimation. An application to Aids. *Statistics in Medicine* **12**, 843-865.

LINTON, O.B. and NIELSEN, J.P. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-101.

MACAULEY, F.R. (1938) Some theoretical problems suggested by the movements of interest rates, bond yields and stock prices in the U.S. since 1856. *New York: NBER.*

NIELSEN, J.P. (1998) Multiplicative bias correction in kernel hazard estimation. *Scand. J. Statist* **25**, 541-553.

NIELSEN, J.P. (1999) Super efficient hazard estimation based on high quality marker information. *Biometrika* **86**, 227-232.

NIELSEN, J.P. and LINTON O.B. (1995) Kernel estimation in a marker dependent hazard model. *Ann. Statist.* **23**, 1735-1748.

NIELSEN, J.P., LINTON O.B. and BICKEL, P. (1998) On a semiparametric survival model with flexible covariate effect. *Ann. Statist.* **26**, 215-241.

REDINGTON, F.M. (1952) Review of the principle of life office valuations. *Journal of the institute of actuaries* **18**, 286-340.

SHEN, P. and STARR, R.M. (1998) Liquidity of the treasury bill market and the term structure of interest rates. *Journal of Economics and Business* **50**, 401-417.

JENS PERCH NIELSEN
Codan
Gammel Kongevej 60
1799 Copenhagen
Denmark