Computational Statistics and Data Analysis 68 (2013) 155–169

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bandwidth selection in marker dependent kernel hazard estimation☆

María Luz Gámiz Pérez [a], Lena Janys [b], María Dolores Martínez Miranda [a,c], Jens Perch Nielsen [c,*]

[a] University of Granada, Spain
[b] University of Mannheim, Germany
[c] Cass Business School, City University London, UK

ARTICLE INFO

Article history:
Received 18 December 2012
Received in revised form 3 June 2013
Accepted 10 June 2013
Available online 4 July 2013

Keywords:
Local linear estimation
Bandwidth
Cross-validation
Indirect cross-validation
Aalen's multiplicative model
Survival

ABSTRACT

Practical estimation procedures for the local linear estimation of an unrestricted failure rate when more information is available than just time are developed. This extra information could be a covariate and this covariate could be a time series. Time dependent covariates are sometimes called markers, and failure rates are sometimes called hazards, intensities or mortalities. It is shown through simulations and a practical example that the fully local linear estimation procedure exhibits an excellent practical performance. Two different bandwidth selection procedures are developed. One is an adaptation of classical cross-validation, and the other one is indirect cross-validation. The simulation study concludes that classical cross-validation works well on continuous data while indirect cross-validation performs only marginally better. However, cross-validation breaks down in the practical data application to old-age mortality. Indirect cross-validation is thus shown to be superior when selecting a fully feasible estimation method for marker dependent hazard estimation.

1. Introduction

Marker dependent hazard estimation is omnipresent in the mathematical statistical literature. Applications exist in many fields, such as actuarial science, applied statistics, biostatistics, econometrics, engineering and finance. The semiparametric structure considered in Cox (1972) and Andersen and Gill (1982) is widely used in the literature and in practice. Additionally, an enormous amount of semi-parametric dynamic survival models can be found in the literature (see for example Andersen et al. (1993), Fleming and Harrington (1991), Martinussen and Scheike (2009), Devarajan and Ebrahimi (2011), Li et al. (2012) and Zhang et al. (2013)). We study the fully unspecified multivariate hazard estimation problem, which has received less attention in the literature than semiparametric hazard models. We work with general filtered survival data, allowing for repeated left truncations and right censoring, as well as fully general time dependent structures on our markers or covariates. Our starting point is the multivariate local linear estimator of Nielsen (1998). It arises from a local linear minimisation principle around the observed counting process, mimicking the delta function approach developed earlier in one-dimensional density estimation by Jones (1993). It is perhaps surprising that a fully feasible estimation procedure has not yet been published for the multivariate local linear estimator (see Nielsen and Tanggaard (2001) and Bagkavos (2011)

☆ Supplementary material can be found online.

* Correspondence to: Cass Business School, Faculty of Actuarial Science and Insurance, 106 Bunhill Row, London EC1Y 8TZ, UK. Tel.: +44 0 20 7040 0909; fax: +44 0 20 7040 8572.
E-mail address: Jens.Nielsen.1@city.ac.uk (J.P. Nielsen).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.010

for bandwidth selectors in the one-dimensional situation). In this paper we develop the classical cross-validation procedure for the marker dependent hazard estimator and we show that it works well in our finite sample studies. However, cross-validation breaks down in our application based on aggregated data. Indirect cross-validation is known to have a better theoretical and practical performance than cross-validation, and it is known to be more robust when applied to discrete data (see Martínez-Miranda et al. (2009), Savchuk et al. (2010), Mammen et al. (2011) and Gámiz et al. (2013) for the related density case). Consequently, in this paper we develop indirect cross-validation for the local linear estimator, which works well when applied to our aggregated data.

The remainder of the paper is organised as follows. In Section 2 we formulate the estimation problem and present the local linear principle following Nielsen (1998). The estimator is formulated in the general counting process formulation. Direct and indirect cross-validation methods are developed in Section 3. The asymptotic theory necessary to implement indirect cross-validation is provided in Appendix A. Simulation experiments are presented in Section 4 and a real data application to old-age mortality is presented in Section 5. These sections are supplemented by Appendix B, which contains discrete approximations of the estimation strategy in order to work with occurrences and exposures. The explicit algorithms used in the simulation experiments are also described there. Some concluding remarks are given in Section 6.

2. The local linear principle for multivariate kernel hazard estimation

In this section we define the local linear marker dependent hazard estimator. We assume that the data follow Aalen's multiplicative intensity model (see Aalen (1978) and Andersen et al. (1993)), which is defined as follows: Let $Z(t)$ be a d-dimensional time dependent covariate or marker dependent process, and let $\lambda(t)$ be the stochastic hazard for an individual with history $\{Z(s); s \leq t\}$. We examine the following marker dependent hazard model:

$$\lambda(t) = \alpha\{t, Z(t)\}Y(t),$$

where $Y(t)$ is an indicator of survival at time t. Suppose we are observing n individuals and let N_i count observed failures for the ith individual in the time interval, which for simplicity is assumed to be $(0, 1)$, for $i = 1, \ldots, n$. Let $\mathbf{N}^{(n)} = (N_1, \ldots, N_n)$ be a n-dimensional counting process with respect to an increasing, right continuous, complete filtration $\mathcal{F}_t$, $t \in (0, 1)$, i.e. one that obeys *les conditions habituelles* (see Andersen et al. (1993, p. 60)). The random intensity process $\lambda^{(n)} = (\lambda_1, \ldots, \lambda_n)$ of $\mathbf{N}^{(n)}$ is then modelled as depending on the d-dimensional marker dependent processes $Z_1(t), \ldots, Z_n(t)$ by

$$\lambda_i^{(n)}(t) = \alpha\{t, Z_i(t)\}Y_i(t), \tag{1}$$

with no restriction on the functional form of $\alpha(\cdot)$. Here Y_i is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is under risk, for $i = 1, \ldots, n$. The marker process $Z_i = (Z_{i1}, \ldots, Z_{id})$ is a d-dimensional, predictable, *CADLAG* covariate. Let $F_s(z) = \Pr(Z_i(s) \leq z \,|Y_i(s) = 1)$ be the conditional distribution function of the covariate process at time s. Furthermore, let $f_s(z)$ be the corresponding density with respect to the d-dimensional Lebesgue measure. We assume that the marker process is supported on the unit cube and that $E\{Y_i(s)\} = y(s)$, where $y(\cdot)$ is continuous. The marker $Z_i(s)$ is only observed for those s where $Y_i(s) = 1$. Let

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

and assume that the stochastic processes $(N_1, Z_1^*, Y_1), \ldots, (N_n, Z_n^*, Y_n)$ are i.i.d. for n individuals and $\mathcal{F}_t = \sigma(\mathbf{N}^{(n)}(s), \mathbf{Z}(s), \mathbf{Y}(s); s \leq t)$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)$. Hereafter we simplify the notation by writing $x = (t, z)$ and $W_i(s) = \{s, Z_i(s)\}$, both being vectors with dimension $d + 1$ and elements enumerated from 0 to d. Let $\mathcal{K}$ be a $d + 1$-dimensional kernel and $\underline{b} = (b_0, \ldots, b_d)$ a $d + 1$-dimensional bandwidth vector. Let $\mathcal{K}_{\underline{b}}(x - y) = |\underline{b}|^{-1}\mathcal{K}\{(x_0 - y_0)/b_0, \ldots, (x_d - y_d)/b_d\}$, where $x = (x_0, \ldots, x_d)$ and $y = (y_0, \ldots, y_d)$ are $(d + 1)$-dimensional vectors and $|b| = \prod_{j=0}^{d} b_j$. We restrict ourselves to the case of multiplicative kernels, that is, $\mathcal{K}(u) = \prod_{j=0}^{d} K_j(u_j)$, where K_j is a univariate kernel.

The local linear estimator of the hazard rate α is then defined as the solution of the following minimisation problem:

$$\begin{pmatrix} \widehat{\Theta}_0 \\ \widehat{\Theta}_1 \end{pmatrix} = \arg\min_{\Theta_0, \Theta_1} \sum_{i=1}^{n} \int \left[\Delta N_i(s) - \Theta_0 - \sum_{j=0}^{d} \Theta_{1j}(x_j - W_{ij}(s)) \right]^2 K_{\underline{b}}(x - W_i(s)) Y_i(s) ds. \tag{2}$$

Here we have used the notation $\int g(s)\Delta N_i(s)ds \equiv \int g(s)dN_i(s)$ for any function g. By solving the above problem in Θ_0, the estimator can be written as an intuitive ratio of the smoothed occurrences and smoothed exposures given by:

$$\widehat{\alpha}_{\mathcal{K},\underline{b}}(x) = \frac{\sum_{i=1}^{n} \int_0^1 \{1 - u^t D(x)^{-1} c_1(x)\} \mathcal{K}_{\underline{b}}(x - W_i(s)) dN_i(s)}{\sum_{i=1}^{n} \int_0^1 \{1 - u^t D(x)^{-1} c_1(x)\} \mathcal{K}_{\underline{b}}(x - W_i(s)) Y_i(s) ds} := \frac{O_{11}(t, z)}{E_{11}(t, z)}, \tag{3}$$