Smoothed Nonparametric Estimation in Window Censored Semi Markov Processes

Enrique E. Alvarez - Department of Statistics
University of Connecticut
215 Glenbrook Rd - Storrs CT 06269 - 4120
email: ealvarez@merlot.stat.uconn.edu

Abstract

Consider a process that jumps among a finite set of states, with random times spent in between. In semi-Markov processes transitions follow a Markov chain and the sojourn distributions depend only on the connecting states. Suppose that the process started far in the past, achieving stationarity. We consider nonparametric estimation by modelling the log-hazard of the sojourn times through linear splines; and we obtain maximum penalized likelihood estimators when data consist of several i.i.d. windows. We prove consistency using Grenander’s method of sieves.

Keywords: Semi-Markov process; stationarity; window censoring; maximum penalized likelihood, sieved MLE, linear splines.

Running headline: Nonparametric Estimation in SMPs

1 Introduction

Consider a continuous time stochastic process \{W(t), t ≥ 0\} that visits a finite set of states in \(\mathcal{K} = \{1, 2, \ldots, k\}\), with random times between transitions. In engineering applications the process \(W(t)\) usually represents a system, and the simplest example is a machine that is either on or off. In such case \(\mathcal{K} = \{0, 1\}\), and \(W(t)\) is an alternating renewal process (ARP). More general state spaces allow to partition \(\mathcal{K}\) into two sets, \(U\) and \(D\), where \(U\) are “up” states and \(D\) is the set of a “down” or repair states. Typically systems start in one particular state in \(U\) that represents the technical perfection, to then transit through suboptimal states and, finally undergo technical service when \(D\) is hit.

We consider a situation in which the duration of the system in each state depends solely on the two connecting states, implying that \(\{W(t), t ≥ 0\}\) is a finite state semi-Markov process. Our interest is to estimate the sojourn distributions
and the transition probabilities, when observations consist of window censored sample paths and the process has already achieved stationarity.

Inference in semi-Markov processes has received attention in the literature, especially for ARPs. Baxter and Li (1994) study the estimation of the “point availability, i.e., the probability that a system modeled by an ARP is on at a given time. They consider data arising from one (not necessarily stationary) ARP and the estimator that results from plugging empirical distributions into the so-called “availability functional”. An important generalization was obtained by Ouhbi and Limnios (2003), who extended the result to semi-Markov processes. Other classic references in the subject are Greenwood and Wefelmeyer (1996), the early paper by Lagakos et. al. (1978) and its following extension by Dinse (1986).

The most related antecedent of the present study, however, is the pioneering paper of Ouhbi and Limnios (2000). They study estimation of the Markov renewal matrix and the Semi-Markov transition matrix for the first time, with right-censored data. Their estimation method is based on a stepwise constant hazard function and for a censoring time $T \to \infty$ they show consistency and asymptotic normality, point-wise, as the mesh becomes finer at a suitable rate.

Our work differs from Ouhbi and Limnios (2000) in several ways. The first difference is that we consider window censored data (i.e. censored on the left as well as on the right) from a stationary semi-Markov process. Secondly, instead of taking piecewise constant hazard functions, we selected a linear spline model for the log-hazard. A usual model in survival analysis, the Weibull model, is a special case of our estimators. So are continuous time Markov chains, i.e., semi-Markov processes with exponential sojourn times.

Estimation of densities or hazards through maximum likelihood presents in the same difficulties in the semi-Markov context as with i.i.d. data (i.e., the unrestricted MLE fails to exist unless extra assumptions are incorporated, or some smoothing method is used). We opted to maximize the penalized likelihood, with penalty functionals chosen so that the estimates of the sojourn distributions are close to either Weibull or exponential densities.

The format of this paper is as follows. In Section 2, we define semi-Markov processes, obtain the stationary distribution, and express the likelihood ratio for
window censored paths. Section 3 proposes the maximum penalized likelihood method, including a discussion of the choice of the penalty functional. Next in section 4, a linear spline model is proposed for the log-hazard functions of the sojourn times. We develop a numerical method to perform the maximization and we finish the section with a numerical example. Finally in section 5, we prove consistency by embedding the maximum penalized likelihood estimation problem into the framework of Grenander’s method of sieves (e.g., 1981).

2 Stationary semi-Markov processes

2.1 Construction and immediate properties

Suppose \( \{W(t), t \geq 0\} \) visits states in \( \mathcal{K} = \{1, 2, \ldots, k\} \) according to \( Q_{ij}(t) \), which denotes the probability that after making a transition into state \( i \), the process makes a transition into state \( j \) in a time no longer than \( t \). We call a transition \( i \xrightarrow{} j \) feasible when \( \theta_{ij} := \lim_{t \to \infty} Q_{ij}(t) > 0 \). For those, the sojourn times have conditional distribution \( F_{ij}(t) := Q_{ij}(t)/\theta_{ij} \). A typical sample path is shown in Figure 1.

For identifiability, we assume that the transition matrix \( \theta \) has a null diagonal, and that it is irreducible. An alternative way to study the process \( \{W(t), t \geq 0\} \), not pursued in this paper, is to let \( N_i(t) \) count the number of transitions into state \( i \) which occur in \( (0, t] \) and look at the process \( \mathbf{N}(t) := (N_1(t), N_2(t), \ldots, N_k(t)) \). Such \( \{\mathbf{N}(t), t \geq 0\} \) is called a (finite state) Markov Renewal Process. When transitions into the same state are disallowed (\( \theta_{ii} = 0 \)), the two processes \( W(t) \) and

\[ K \]

\[ d \times s_0 \]

\[ x_0 \times s_1 \]

\[ x_1 \times s_2 \]

\[ x_2 \times s_3 \]

\[ x_3 \times s_4 \]

\[ x_4 \times s_\tau \]

\[ \cdots \]

\[ x_\tau \]

\[ T \]

Figure 1: A Sample Path from a 4 state Semi-Markov Process.
\( N(t) \) are in one to one correspondence. One immediate consequence is that on any finite time interval, there is a finite expected number of jumps, i.e., \( E\{N_i(t)\} < \infty \). for all \( i \).

It is also important to notice that when \( \mathcal{K} = \{0,1\} \), \( W(t) \) is an alternating renewal process, and if it further holds that the sojourn densities are of exponential form \( f_{01}(x) = \lambda_1 \exp(-\lambda_1 x) \) and \( f_{10}(x) = \lambda_2 \exp(-\lambda_2 x) \), then \( \{W(t), t \geq 0\} \) is a 2-states continuous time Markov chain (CTMC), for which some results in likelihood based estimation are known. Alvarez (2003b) shows that the MLE fails to exist for some data sets, and gives a characterization of the “regular” data configuration (in which the MLE exists and is interior to the parameter space). It further shows that ignoring stationarity in the estimation represents often a substantial loss of efficiency for most parameters of interest.

2.2 Stationarity

When we arrive at the observation site we immediately see that \( W(\cdot) \) lies in an initial state \( j_0 := W(0) \) (though we ignore for how long) and that it remains there for a time \( x_0 \) before jumping to \( j_1 := W(x_0) \). Under stationarity, \( P(J_0 = j_0) = P(W(t) = j_0) =: P_j \) needs to be constant. In order to guess what it should be, we need to introduce the following types of expected times:

1. \( \mu_{ij} := \int_0^\infty [1 - F_{ij}(t)] dt \) is the expected time in \( i \) before transiting to \( j \),

2. \( \mu_i := \sum_{j=0}^{\mathcal{K}} \theta_{ij} \mu_{ij} \) is the expected time in \( i \) before a transition anywhere,

3. Finally let \( \xi_{ij} \) denote the mean time it takes to reach state \( j \) from state \( i \) with any number of visits to other states in between. This is for, \( i \neq j \),

\[
\xi_{ij} := E \left[ \inf \left\{ s > 0 : W(t+s) = j \mid W(t) = i, W(t-) \neq i \right\} \right],
\]

while the expected revisit time a state \( i \) is

\[
\xi_{ii} = E \left[ \inf_{s>0} \left\{ W(t+s) = i, W(t+u) \neq i \mid \exists 0 < u < s \mid W(t) = i, W(t-) \neq i \right\} \right].
\]
In addition, since the transition matrix $\theta$ is irreducible, a unique stationary distribution exists for the embedded Markov chain with state probabilities $\pi_i$ that are strictly positive. (e.g., Resnick, 1992).

In consideration of $P_j$, we could reason heuristically that since $\pi_i$ equals the (long-run) proportion of transitions that are into state $j$, since $\mu_j$ is the mean time spent in $j$ per transition, and since $\xi_{jj}$ is the average time it takes the process to revisit $j$, the state probabilities, if stationary, should be

$$P_j = \frac{\mu_j}{\xi_{jj}} = \frac{\pi_j \mu_j}{\sum_{K} \pi_j \mu_j} \quad \text{for all } j \in K. \quad (1)$$

This is both true (e.g. Ross, 1992), and important. Equation (1) reduces the problem of obtaining the stationary state probabilities of the Semi-Markov process $P_j$'s and the mean intervisit times $\xi_{jj}$'s into just calculating the stationary probabilities of the embedded Markov chain and the expected inter-arrival times for each pair of transitions. This is a valuable simplification, since the mean recurrence times $\xi_{ii}$ are functions of all transition probabilities $\theta_{ij}$ and all distribution functions $F_{ij}$.

Of course, stationarity of $W(\cdot)$ requires much more than constant $P_j$'s, as we now define:

**Definition 2.1.** The stochastic process $\{W(t), t \geq 0\}$ is stationary when for all $h \geq 0$, $\{W(t), t \geq 0\}$ and $\{W(t+h), t \geq 0\}$ are equal in distribution.

We ask now: which distribution for the first triplet $(J_0, J_1, X_0)$, if any, makes the SMP stationary? Such a distribution does exist and it is given next.

**Theorem 2.2.** The process $\{W(t), t \geq 0\}$ is stationary iff

$$P \{W(0) = j, J_1 = k, X_0 \leq x\} = \frac{\theta_{jk}}{\xi_{jj}} \int_0^x [1 - F_{jk}(t)] dt. \quad (2)$$

**Remark 2.3.** If there is only one state, equation (2) gives the usual length-biased distribution, and if $W(t)$ is an ARP with “on” times $Z \sim F_Z$ and “off” times $Y \sim F_Y$ it gives,

$$P \{W(0) = 1, Z_0 \leq z\} = \frac{\mu_Z}{\mu_Z + \mu_Y} \int_0^z [1 - F_Z(t)] dt,$$

$$P \{W(0) = 0, Y_0 \leq y\} = \frac{\mu_Z}{\mu_Z + \mu_Y} \int_0^y [1 - F_Y(t)] dt,$$

a known result for ARPs (e.g. Alvarez, 2003b).
3 Maximum penalized likelihood estimation

Consider a single typical sample path of a window censored SMP, as shown in Figure 1. In order to express the likelihood we need to introduce some notations.

**Jump times** $S_0 := 0$ and $S_i := \inf\{t > S_{i-1} : W(t) \neq W(S_{i-1}), t > 0 \}$.

**Soujourn times** $X_i := S_i + 1 - S_i$.

**Chain of states visited** $J_i := W(S_i)$.

**Number of jumps** $\tau := \sup\{i \geq 0 : S_i < T\}$.

With these, we express the likelihood over a single window by

$$L[w(t), 0 \leq t \leq T] := \frac{\theta_{j_0j_1}}{\xi_{j_0j_0}} F_{j_0j_1}(x_0) \prod_{i=1}^{\tau-1} \theta_{j_i,j_{i+1}} F_{j_i,j_{i+1}}(x_i)$$

$$\left[\sum_{k=1}^{K} \theta_{j,k} \bar{F}_{j,k}(x_{\tau})\right] 1(\tau > 0) + \left[\sum_{k=1}^{K} \frac{\theta_{j_0k}}{\xi_{j_0j_0}} \bar{F}_{j_0k}(x_0)\right] 1(\tau = 0), \quad (3)$$

where $\bar{F}$ is the survival function. Equation (3) gives a “recipe” to compute the window-censored likelihood by multiplying three types of terms. In a typical sample path, where at least one transition in observed, we multiply: (i) the initial density $(\theta_{j_0j_1}/\xi_{j_0j_0}) F_{j_0j_1}(x_0)$ in accordance to equation (2), (ii) the densities of all non-censored sojourn times $\prod \theta_{j_i,j_{i+1}} F_{j_i,j_{i+1}}(x_i)$, and finally (iii) the survival function of the last state in the window $\sum \theta_{j,k} \bar{F}_{j,k}(x_{\tau})$. If, however, the window $[0, T]$ contains no jumps the likelihood equals $\sum_k (\theta_{j_0k}/\xi_{j_0j_0}) \bar{F}_{j_0k}(x_0)$.

Maximum likelihood estimation over $n$ i.i.d. windows involves maximizing the sample likelihood $L[w_1, w_2, \ldots, w_n] = \prod_{h=1}^{n} L[w_h(t), 0 \leq t \leq T]$ in the arguments $\hat{\theta}_{ij} = \hat{\theta}_{ij}(w_1, w_2, \ldots, w_n)$ and $\hat{F}_{ij} = \hat{F}_{ij}(w_1, w_2, \ldots, w_n)$ with the requirements that $\hat{\theta} \in \Theta := \{\theta \in [0, 1]^{k \times k} : \forall i \theta_{ii} = 0, \text{ irreducible}\}$ and the $\hat{F}_{ij}$’s belongs to some family of distributions $\mathcal{F}_{ij}$. Even in parametric contexts there is no guarantee that the MLE exists (e.g., Alvarez, 2003b). But more importantly, with an absolutely continuous non-parametric choice of $\mathcal{F}_{ij}$, maximization of equation (3) presents the same difficulties as ML estimation of densities (i.e., non-existence). For successful
likelihood based estimation it is necessary to either restrict the class $\mathcal{F}$ or alter the likelihood functional in a suitable way. This could be achieved by (i) assuming a parametric family, (ii) putting shape restrictions, or (iii) adding a roughness penalty to the likelihood (see Silverman, 1986, for an exposition of these various methods in the i.i.d. context).

The choice of Ouhbi and Limnios (2000) was to assume a piecewise constant form for the hazard functions. We, instead, seek a maximizer of the penalized likelihood $\ln L - \lambda R$, where $R$ is a penalty that measures the “roughness” of the estimate and $\lambda$ is a constant chosen to control the trade-off between fidelity to the data and the avoidance of roughness. There are different natural choices for $R$. We have selected $R = \sum_{ij} R_{ij}$ with

$$R(h_{ij}) = \int_0^T \left( \frac{d\log h_{ij}(s)}{d\ln s} \right)^2 ds,$$

which gives no penalty to exponential densities. With this, elements in the null family $\mathcal{R}_0 := \{ h \in \mathcal{H} : R(h) = 0 \}$ are continuous time Markov chains, which we consider the “smoothest” type of SMPs (as the Markov property holds). With the MPL method we can force the estimates to approximate $\mathcal{R}_0$ by choosing a large enough penalty constant. Another natural choice is

$$R^*(h_{ij}) = \int_0^T \left\{ \frac{d\log h_{ij}(s)}{d\ln s} - \int_0^T \left( \frac{d\log h_{ij}(u)}{d\ln u} \right) du \right\}^2 ds.$$

This forces the hazard to be close to linear in $\ln(x)$, producing a Weibull density, a widespread choice for modelling duration data.

In the nonparametric regression context with normal errors, when the penalty functional is proportional to the curvature of the regression function, calculus of variations shows that the solution of the optimization problem is a cubic spline, i.e., a piece-wise cubic polynomial with continuous derivative (e.g., Whaba, 1984). In the case of density estimation, however, there needs to be a nonlinear term in the likelihood function in order for the density to integrate to unity. Because of this, it is known in the i.i.d. case that the maximizer of the penalized likelihood according to calculus of variations satisfies a nonlinear fourth order differential equation in between data points, with jumps in the third derivative (e.g., O’Sullivan, 1988); an exact solution is hence discarded, in favor of an approximation.
4 A linear splines model for the log-hazards

We propose an approximate solution to the MPL problem, based on modeling the log-hazards via linear splines. An advantage of this postulation is that Weibull sojourn distributions are special cases, a feature that is lost if the hazard where modelled as piecewise constant. Additionally, the spline model gives flexibility to accommodate to peaks in the hazard.

For absolutely continuous distributions, there is an one to one correspondence between the distribution \( F \) and (equivalence classes of) the hazard rate function \( h \) given by the pair of equations

\[
\begin{align*}
  h(x) &= \frac{f(x)}{1 - F(x)} \quad \text{and} \quad f(x) = h(x) \exp \left[ - \int_0^x h(s) \, ds \right].
\end{align*}
\]

(4)

We assume that \( \ln(h) \) is a linear spline in \( \ln(x) \) with knots at \(-\infty < t_1, \ldots, t_k < \ln(T)\). That is, \( \ln(h) \) is a piecewise linear continuous function of \( \ln(x) \) with vertices at the \( t_i \). We select 5 knots placed in a data dependent manner at the log of the percentiles 10, 25, 50, 75 and 90 of the set of transition times that are not right truncated. This choice represents a compromise between the goals of capturing the essential features of the hazard function and keeping the number of parameters small. (For a discussion on the merits of different knot placement rules in density estimation see Kooperberg, 1991).

To be specific, we parameterize each log hazard by its height \( v_1 \) at the first knot and by the slopes \( b_0, \ldots, b_5 \). That is, \( v(\ln x) := \ln h_{ij}(x) \), with

\[
v(\ln x) = v_1 - b_0(t_1 - \ln x)1(\ln x < t_1) + \sum_{i=1}^5 b_i(\ln x \wedge t_{i+1} - t_i)1(\ln x \geq t_i)
\]

(5)

(with \( t_6 := \infty \)). For the density to be non-degenerate we need to impose a couple of conditions on the cumulative hazard, defined as \( H(x) := \int_0^x h(s) \, ds \). First, it must be unbounded above. Calling the height of the log-hazard at the knots \( v_i := \ln h(t_i) \) after some simple algebra on (5) we get

\[
\int_{\exp(t_5)}^{\infty} h(s) \, ds = \int_{\exp(t_5)}^{\exp(t_5)} \exp[v(s)] \, ds = \exp[v_4 - b_5 t_5] \int_{\exp(t_5)}^{\infty} s^{b_5} \, ds,
\]

(6)

which is \( \infty \) as long as \( b_5 \geq -1 \). Also, the integral at the left of the first knot must be finite. That is

\[
\int_{0}^{\exp(t_1)} h(s) \, ds = \int_{0}^{\exp(t_1)} \exp[v(s)] \, ds = \exp[v_1 - b_0 t_1] \int_{0}^{\exp(t_1)} s^{b_0} \, ds < \infty,
\]

(7)
which is satisfied for \( b_0 > -1 \). With conditions (6) and (7) it is guaranteed that \( H(x) \geq 0 \) for \( x \geq 0 \) and that \( \lim_{x \to \infty} H(x) = \infty \). This is sufficient for the density to be nonnegative and integrate to unity, since

\[
\int_0^\infty f(x)dx = \int_0^\infty h(x) \exp[-H(x)]dx = -\exp[-H(x)]|_0^\infty = 1.
\]

In order to obtain an expression for the likelihood of the semi-Markov process in terms of the hazard rates, we need now to obtain formulas for the density, the survival function and the mean. Integrating (5), the cumulative hazard can be obtained recursively by

\[
H(x) = \begin{cases} 
\exp[v_1 - b_0 t_1] \frac{x^{b_0+1}}{b_0 + 1} & \ln x \leq t_1 \\
H(e^{t_i}) + \exp[v_i + t_i] \frac{x^{b_i+1}}{b_i + 1} & t_i \leq \ln x \leq t_{i+1}, \ 1 \leq i \leq 5.
\end{cases}
\]

(8)

Integrating now \( E(X) = \int_0^\infty [1 - F(u)]du \) we get that \( \mu \) equals

\[
\int_0^{\exp(t_1)} \exp[-H(u)]du + \sum_{i=1}^4 \int_{\exp(t_i)}^{\exp(t_{i+1})} \exp[-H(u)]du + \int_{\exp(t_5)}^{\infty} \exp[-H(u)]du
\]

\[
= \exp \left[ -h(e^{t_1}) e^{t_1} \right] \int_0^{\exp(t_1)} \exp(-u^{b_0+1})du + \sum_{i=1}^4 \int_{\exp(t_i)}^{\exp(t_{i+1})} \exp[-H(u)]du
\]

\[
+ \exp \left[ -H(e^{t_5}) + \frac{-h(e^{t_5}) e^{t_5}}{b_5 + 1} \right] \int_{\exp(t_5)}^{\infty} \exp(-u^{b_5+1})du.
\]

(9)

Notice that all the terms in r.h.s. are of the form \( \int \exp(-u^\alpha)du \), but while \( \alpha \) is nonnegative for the first and last integral, in view of conditions (6) and (7), this is not necessarily true for the middle terms. This is relevant computationally because when \( \alpha \geq 0 \) the change of variables \( y = u^\alpha \) gives rise to incomplete gamma functions, for the approximation of which powerful standard numerical algorithms exist based on either series or continued fraction expansions. For the middle terms, however, approximate integration is needed.

### 4.1 Maximization algorithm

Recall that \( \theta \in \Theta \subset \{0, 1\}^{k \times k} : \forall i \theta_{ii} = 0 \), irreducible \} and with the spline approximation \( h_{ij} \in \mathcal{H} := \{ h : \ln(h) \text{cts, piecewise linear with knots at } t_1, \ldots, t_5 \} \). Maximizing the penalized log-likelihood over the sample is thus a finite dimensional
problem in the transition probabilities $\theta_{ij}$’s and the parameters that characterize each log-hazard spline: $v_1$ and $b_0, \ldots, b_5$. For the following reasons, estimators in closed form are not available and we need a numerical procedure: (i) the likelihood depends on the stationary probabilities of the Markov chain, (ii) it also depends on the mean sojourn times, which have to be calculated with numerical integration and algorithms for the incomplete gamma function, and (iii) the placement of the knots is data dependent.

We have developed a numerical method of optimization that essentially fits a continuous time Markov chain to the data (constant hazard) as initial estimator and then performs updates using a method of conjugate directions (e.g. Lange, 1998).

4.2 Numerical application

To illustrate, we apply the method to a 4-states semi-Markov process. The transitions are given in Figure 2, where we consider two transition probabilities are known, while the rest are estimated. All sojourn distributions are of Weibull type except the one corresponding to \(4 \rightarrow 1\), which has triangular hazard

\[
h_{41}(x) = \begin{cases} 
0.1 + 0.25x & 0 < x < 2 \\
1.1 - 0.25x & 2 \leq x < 4 \\
0.1 & x \geq 4,
\end{cases}
\]

(and notice its not in the space because its the log hazard that has to be linear).

We did the estimation out of 50 windows, non-penalized and with a penalty large
enough to force Weibull estimates. Results are presented in Figures 3 to 6. They show a clear advantage of the nonparametric model to fit the sojourn distribution from \(4 \to 1\), which has a peak in the hazard in the interior of the observation window. We also observe that the main way in which the estimators differ is that, especially for small samples, the spline estimator is more flexible to give less mass to the tails of the distribution.
5 Consistency

There are several papers in the literature providing asymptotic analyses of penalized likelihood or log-spline estimators for density functions or hazard rates. Cox and O’Sullivan (1990) study the consistency of penalized M-estimators for log-density, log-hazards and non-parametric logistic regressions under a family of spectral norms in Sobolev spaces, when the penalty constant converges to zero and the sample size increases. In their paper, they look at M-estimators for functions known to have at least one continuous derivative; no knots are placed and computational aspects are not considered. A different approach is followed by Stone (1990), who considers consistency when the log-density is constrained to be a spline with an increasing number of knots, placed in a data-independent manner. He does not penalize and his estimators are based on i.i.d. observations from a continuous density on a compact interval. His analysis relies on a parameterization based on $B$ splines, under which for data that are not nontruncated or censored, the densities form a finite parameter exponential family.

The situation for penalized log-hazard estimation in a semi-Markov process has some unique features. As mentioned in Stone(1990), the estimators are restricted to be splines with pre-set knots. But here censoring and truncation together with the stationarity preclude an exponential family structure in any $B$-spline parameterization. On the other hand, penalization is also considered.

To prove consistency for knots placed independently of the data, we embed the problem into the frame of Grenander’s method of sieves for Maximum Likelihood estimation (e.g. Grenander 1981). We present next our theoretical framework:

We observe $n$ realizations $w_1, w_2, \ldots, w_n$ of a stationary semi-Markov process on $[0, T]$. The transition matrix $\theta \in \Theta \subset \{[0, 1]^{k \times k} : \forall i \theta_{ii} = 0, \text{irreducible}\}$ and $h_{ij} \in \mathcal{H} := \{h : [0, \infty) \rightarrow \mathbb{R}^+ : \ln(h) \text{ continuous, and constant on } [T, \infty)\}$ so that we are assuming that the true hazards can be arbitrary, but with exponential tails. Letting now $k^* := \sharp\{\theta_{ij} > 0\}$, the parameter space is the Cartesian product $\mathcal{A} := \Theta \times \mathcal{H}^{k^*}$. For the value space we have that each window is a sample path

\[ w \in \mathcal{W} := \{w : [0, T] \rightarrow \{1, 2, \ldots, k\}, \text{cadlag, piecewise constant}\}. \]
The $n$ windows are i.i.d. observations on the space $(\mathcal{W}, \mathcal{B}_W, P_\eta)$ where the $\sigma$-field $\mathcal{B}_W$ is generated by the Skorohod metric on $\mathcal{W}$, and the measures $P_\eta$ range over $\eta \in \mathcal{A}$. Since maximization over $\mathcal{A}$ is indeterminate, we consider instead a sequence of subsets of $\mathcal{H}_m \subset \mathcal{H}$ that we define as the class of all functions $h \in H$ such that $\ln(h)$: (i) is piecewise linear in $\ln(x)$, (ii) it has knots at $t_i := \ln(Ti/m_n)$ for $0 < i < m_n$, (iii) it is constant on $(-\infty, t_1)$ and $(t_{m-1}, \infty)$, and (iv) it satisfies the bounds $\sup\{0 \leq x \leq T : |\ln h(\ln x)|\} \leq r(m)$, and (v)
\[
\int_0^T \left[ \frac{d}{d \ln u} \ln h(\ln u) \right]^2 du \leq s(m),
\]
with $m = m(n)$, $r = r(n)$ and $s = s(n)$ increasing.

Consider now a sequence of estimators $\hat{\eta}_m$ as those which maximize the log-likelihood in $\Theta \times \mathcal{H}_m^k$. To get some intuition about the meaning of this constrained optimization, notice that this problem has a Lagrangian equivalent of maximizing

\[
\tilde{l}_m(w_1, w_2, \ldots, w_n, \eta) := \frac{1}{n} \sum_{i=1}^n \ln dP_\eta(w_i) + \lambda \int_0^T \left[ \frac{d}{d \ln u} \ln h(u) \right]^2 du,
\]
subject to each $\ln h_{ij}$ being a linear spline in $\ln(x)$ with knots at $\ln(iT/m_n)$, $i = 1, 2, \ldots, m_n - 1$, when it is known that the constraint is tight. Our treatment hence combines the log-spline and penalization approaches.

Now for $\eta_0 \in \mathcal{A}$, we want to establish the strong consistency $\hat{\eta}_m \to \eta_0$ as $n \to \infty$. This requires to select appropriate functions $m(n) \to \infty$, $r(m) \to \infty$, and $s(m) \to \infty$, together with a metric on $\mathcal{A}$. In the literature, the choice of the metric is tied in with the assumptions one is willing to make about $\mathcal{H}$, with Sobolev type norms allowed when continuity of a certain number of derivatives is assumed. For our development, we do not make those type of assumptions and use Hellinger distance on $\mathcal{A}$. This choice has a couple of advantages. First, it is known that under identifiability for a given metric $\tau$, Hellinger-convergence implies convergence in $\tau$ (e.g. Van de Geer, 1993). E.g. we could choose
\[
\tau(\eta, \bar{\eta}) := ||\theta - \bar{\theta}|| + \max_{i,j} ||h_{ij} - \bar{h}_{ij}||_\infty.
\]
Second, Hellinger distance is dominated by Kullback Leibler pseudo-distance, so known results in Kullback Leibler convergence can be applied instead of dealing with Hellinger distance directly. A disadvantage, however, is that presumably better rates could be obtained form a direct approach.
6 Consistency Proof

We re-express the sets $\mathcal{H}_m$ using a basis of $B$-linear splines, under which for functions $h \in \mathcal{H}_m$, its log takes the form $\ln h(\ln x) = \sum_{i=1}^{m-1} b_i B_i(\ln x)$ (e.g. De Boor, 1978). Notice that $|\ln h| \leq r(m)$ implies that $\max_i |b_i| \leq r(m)$ and in turn $\max_i |b_i| \leq r(m)$ implies that $\sup |\ln h| \leq 2r(m)$, for it is a property of the linear spline basis that at most two of the $B_j$’s can be nonzero at any point. Also, in $\mathcal{H}_m$, the penalty functional becomes

$$\int_0^T \left[ \frac{d}{d \ln u} \ln h(\ln u) \right]^2 du = \sum_{i=1}^{m-2} (b_{i+1} - b_i)^2 \frac{1}{m} T \leq C r(m)^2.$$  

Then if we take $s(m) = C r(m)^2$, the smoothness constraint becomes asymptotically “untight” or redundant as $n$ increases.

We also need to take care that the parameter space for the transition matrix is not compact. For $\theta$ to be irreducible, it is necessary that the algebraic multiplicity of the unit eigenvalue be one. Letting $P_{\theta}(\lambda) = |\theta - \lambda I|$ be the characteristic polynomial, this condition implies

$$\lim_{\lambda \to 1} \frac{|\theta - \lambda I|}{\lambda - 1} = \left( \frac{d}{d \lambda} P_{\theta}(\lambda) \right)_{\lambda=1} \neq 0.$$  

In addition, for irreducibility the states must all communicate and the stationary probabilities will be positive. So we estimate $\theta$ in the sequence of closed sets $\Theta_m := \{ \theta \in \Theta : \theta_{ij}, m^{-\beta} \leq \theta_{ij}, \pi_i \leq 1 - m^{-\beta}, \quad |P_{\theta}(1)| \geq m^{-1} \}$, with $0 < \beta < 1$.

The estimators $\hat{\eta}_m := (\hat{\theta}_m, \hat{h}_m)$ that maximize the log-likelihood of the sample in $\mathcal{A}_m = \Theta_m \times \mathcal{H}_m$ are called sieved maximum likelihood estimators. In the appendix we prove their consistency by verifying conditions are satisfied to apply a theorem given as Theorem 2 in Geman and Hwang (1982). The main result is that

$$\| \eta_m - \eta_0 \|_{\text{Hellinger}} = O(n^{1-\varepsilon}),$$

establishing the consistency of the spline maximum penalized likelihood estimators. It is relevant to mention that we have not attempted to achieve the best possible rate. In the literature of density estimation, the optimal rate depends
on such factors as the smoothness one is willing to assume for the true densities or hazards and the choice of the metric in the parameter space (e.g. Cox and O’Sullivan, 1990), and the smoothing is achieved through either penalization or placing knots at a subset of the sample. In our case, we could additionally exploit the combination of different knot placement rules and different penalty bounds with the aim of improving rates. Doing so would presumably be require the machinery of empirical processes, as in Shen and Wong (1994).

It is also noteworthy that in the consistency proof we intentionally skipped the fact that the knot are placed in a data dependent way. The question of how the asymptotic properties behave when the smoothing is data-dependent is an open problem even in the i.i.d. case, where it is known that data dependency can preclude consistency even in simple examples (e.g., Chow et. al, 1983).

Acknowledgement

This work forms part of the author’s Ph.D. dissertation, presented to the Department of Statistics, University of Michigan (December, 2002). The author wishes to thank his advisor, Robert W. Keener, for superb guidance throughout the development of the thesis work.

Appendix

A Stationarity

A.1 Proof of Theorem 2.2

Proof. For fixed $j$ and $k$, let $\Psi(t) := P\{W(t) = j, J_{N(t)+1} = k, Y(t) > x\}$. We will see that with the choice in equation (2) of Theorem 2.2, $\Psi(t)$ is a constant. Recall that the processes $\{N_j(s), s \geq 0\}$ that count transitions into each state $j$ are of renewal type with arrival times (say) $S_1^{(j)}, S_2^{(j)}, \ldots$ and $N_j(t) < \infty$ w.p.1. Observe further that

$$
\Psi(t) = P \{W(t) = j, J_{N(t)+1} = k, Y(t) > x, N_j(t) = 0\}
+ \sum_{n=1}^{\infty} P \{W(t) = j, J_{N(t)+1} = k, Y(t) > x, N_j(t) = n\},
$$

(11)
and the first term in the l.h.s. is
\[ P \{ W(0) = j, N(t) = k, X_0 > t + x \} = \frac{\theta_{jk}}{\xi_{jj}} \int_{t+x}^{\infty} [1 - F_{jk}(t)] dt. \]

Since also \( \{ W(t) = j, N_j(t) = n_j \} = \{ S_{n_j}^{(j)} \leq t, X_{n_j} > t - S_{n_j}^{(j)} \} \), the second term in (11) becomes after conditioning on \( S_{N_i(t)} = s \),
\[
\sum_{n=1}^{\infty} \int_0^t P \{ W(t) = j, N(t) = k, X > t - s + x, S_{N_i(t)} = s \} F_{s_n^{(j)}}(ds) = \int_0^t \theta_{jk} [1 - F_{jk}(t - s + x)] \nu_r(ds),
\]
where the renewal measure \( \nu_r([0, s]) = s/\xi_{jj} \) for a stationary renewal process (e.g., Resnick 1992, chapter 3). With this we get
\[
\int_0^t \theta_{jk} [1 - F_{jk}(t - s + x)] \nu_r(ds) = \frac{\theta_{jk}}{\xi_{jj}} \int_x^{t+x} [1 - F_{jk}(u)] du. \tag{12}
\]
Finally, summing the two integrals,
\[
\Psi(t) = \frac{\theta_{jk}}{\mu_{jj}} \int_x^{\infty} [1 - F_{jk}(u)] du = \Psi(0).
\]

This indeed gives stationarity in accordance to definition 2.1, for the process \( W(t) \) was constructed solely based on the Q’s and an initial state. For a rigorous treatment see Alvarez (2003a) or Alvarez (2003b) for the proof in the special case of \( K = \{0, 1\} \).

### B Consistency

In order to prove consistency, we will verify that the conditions of the following theorem, given in Geman and Hwang (1982), are satisfied.

**Theorem B.1.** Assume that the following two conditions hold:

**Condition 1** With \( L_n(w, A_m) := \sup_{\eta \in A_m} L_n(w, \eta) \), the set
\[
M^*_{m} := \{ \eta \in A_m : L_n(w, \eta) = L_n(w, A_m) \}
\]
is nonempty.
Condition 2 a) If, for some sequence \( \eta_m \in A_m \),
\[
E_{\eta_0} \left( \ln \frac{l(w, \eta_0)}{l(w, \eta_m)} \right) \to 0,
\]
then \( \eta_m \to \eta_0 \) in the chosen metric.

b) There exists a sequence \( \eta_m \) such that (13) holds.

For each \( \delta > 0 \) and each \( m \) define
\[
D_m := \left\{ \eta \in A_m : E_{\eta_0} \ln \frac{l(w, \eta)}{l(w, \eta_m)} \leq -\delta \right\}
\]
where \( \eta_m \) is the sequence in Condition 2-(b). Given \( l \) sets \( O_1, O_2, \ldots, O_l \) in \( A_m \) such that the likelihood function is measurable for each, define
\[
\rho_m := \sup_k \inf_{t \geq 0} E_{\eta_0} \left[ t \ln \frac{l(w, O_k)}{l(w, \eta_m)} \right].
\]

Let \( m_n \) be a sequence diverging to \( \infty \) and suppose that for each \( \delta > 0 \), we can find \( O_1^m, O_2^m, \ldots, O_l^m \) in \( A_m \), \( m = 1, 2, \ldots \), such that (i) \( D_m \subset \bigcup_{k=1}^{l_m} O_k^m \), (ii) \( l(w, O_k^m) \) is measurable, and (iii) \( \sum_{n=1}^{\infty} l_m(\rho)^n \leq \infty \).

Then, \( \eta_m \to \eta_0 \) a.e.

B.1 Checking the conditions

In checking the conditions of the theorem and constructing the cover of \( D_m \) we will obtain also a rate. Condition 1 follows because the function \( \ln l(w, \cdot) \) is continuous on the compact set \( A_m \), and Condition 2-a follows because Kullback-Leibler pseudo-distance dominates Hellinger distance.

Condition 2-b Let us exhibit a sequence \( \eta_m = (\theta_m, h_m) \) such that equation (13) holds. Since \( \Theta^0 = \bigcup_m \Theta_m \), for any \( \theta_0 \in \Theta \) there exists a sequence \( \{\theta_m\} \) with \( \theta_m \in \Theta_m \) for which \( ||\theta_m - \theta_0|| \to 0 \). On the other hand, for each pair of states \( \{i, j\} \) the log-hazard (call it \( v := \ln h \)) is continuous, constant on \( [\ln T, \infty) \), and with horizontal left sided asymptote \( \lim_{y \to -\infty} v(y) = \ln h(0) \in \mathbb{R} \). So that by taking continuous piecewise linear functions \( v_m \) that agree with \( v \) at the knots
\[
\left\{ \ln \left( \frac{1}{m} T \right), \ldots, 0, \ldots, \ln \left( \frac{m-1}{m} T \right) \right\}
\]
and that are constant on the tails we get \( ||v_m - v||_\infty \to 0 \). By continuity, we get the required convergence of the expected log-likelihoods.
We need now to define sequences of sets \( \mathcal{O}_1^m, \mathcal{O}_2^m, \ldots, \mathcal{O}_m^m \) covering \( D_m \) in (14), with \( \eta_m \) being the sequence in Condition 2-b. Consider thus the set of functions \( \beta \in \mathcal{H}_m \) whose representation \( \ln \beta(x) = \sum_{i=1}^{m-1} b_i B_i(x) \) has coefficients of the form \( b_i = -\alpha \ln m + pm^{-2} \) for some \( p \in \mathbb{Z} \). Since \( |b_i| \leq \alpha \ln m \), there are no more than

\[
\left( \frac{2 \ln m}{m^2} \right)^m \leq (cm)^m
\]

of such \( \beta \)’s. Here, as in what follows we use \( c \) to denote generic constants (i.e., \( c \) need not be the same each time it appears).

Now associate with each of these \( \beta \)’s, the set of all functions \( \gamma \in \mathcal{H}_m \) satisfying \( \| \ln \beta - \ln \gamma \|_\infty \leq (2m)^{-1} \). Call the resulting collection of sets \( \hat{\mathcal{Q}}_1^m, \hat{\mathcal{Q}}_2^m, \ldots, \hat{\mathcal{Q}}_m^m \), where \( j_m \leq (cm)^m \). That they are a cover since for any \( \xi \in \mathcal{H}_m \), its logarithm has a representation \( \ln \xi(x) = \sum_{i=1}^{m-1} b_i B_i(x) \) and we can find coefficients \( a_1, a_2, \ldots, a_m \) of the from \( a_i = -\alpha \ln m + pm^{-2} \) for which \( |a_i - b_i| \leq m^{-2} \). Thus, with \( \ln \alpha(\ln x) = \sum_{i=1}^{m-1} a_i B_i(x) \), \( \| \ln \alpha - \ln \xi \|_\infty \leq 2m^{-2} \).

Next, we partition \([1/m, 1 - 1/m] \) into \( m - 2 \) closed intervals \([1/m, 2/m], [2/m, 3/m], \ldots, [(m - 2)/m, (m - 1)/m]\) and we construct \( \ell_m = [(m - 2) \times j_m]^k \) sets \( \hat{\mathcal{O}}_1^m, \hat{\mathcal{O}}_2^m, \ldots, \hat{\mathcal{O}}_m^m \) by taking all Cartesian products between those intervals and the \( \hat{\mathcal{Q}}^m \)’s for all \( k \) possible transitions \( \{i, j\} \). Finally, we define \( \mathcal{O}_k^m := \hat{\mathcal{O}}_k^m \cap D_m \). These cover \( D_m \).

At this point we need to obtain \( \rho_m \) as defined in equation (15). Take \( \eta \) and \( \tilde{\eta} \) in \( \mathcal{O}_k^m \). For each \( \{i, j\} \), \( \text{sup} \{ \eta, \tilde{\eta} \in \mathcal{O}_k^m : \| \ln h_{ij} - \ln \tilde{h}_{ij} \|_\infty \} \leq m^{-1} \). This translates into approximation rates for relevant functionals in the log-likelihood, as follows. First notice that

\[
\|h_{ij} - \tilde{h}_{ij}\|_\infty = \left\| 1 - \exp \left( \ln h_{ij} - \ln \tilde{h}_{ij} \right) \right\|_\infty \exp \left( \ln \tilde{h}_{ij} \right) \| \exp [r(m)].
\]

Choose now \( r(m) = \alpha \ln m \) for some \( 0 < \alpha < 1 \), so that \( \|h_{ij} - \tilde{h}_{ij}\|_\infty = O(m^{\alpha-1}) \). Now, the difference in cumulative hazards satisfies

\[
\|H_{ij}(x) - \tilde{H}_{ij}(x)\|_\infty \leq \int_0^x \left| h_{ij}(s) - \tilde{h}_{ij}(s) \right| ds = O(m^{\alpha-1}),
\]

18
for all $x \in \mathbb{R}$. For the survival function we get
\[
\| \exp (-H_{ij}(x)) - \exp (-\tilde{H}_{ij}(x)) \|_\infty \\
= \| 1 - \exp (-\tilde{H}_{ij}(x) + H_{ij}(x)) \|_\infty \| \exp (-H_{ij}(x)) \|_\infty \\
\leq | 1 - \exp (-cm^{\alpha-1}x) | ,
\]
implying that $\| S_{ij}(x) - \tilde{S}_{ij}(x) \|_\infty = O(m^{\alpha-1})$. This also translates into rates for the mean inter-arrival times. First notice that for $\eta \in A_m,$
\[
| \ln h | \leq \alpha \ln m \Rightarrow \frac{1}{m^\alpha} \leq h \leq m^\alpha \\
\Rightarrow \frac{1}{m^\alpha} x \leq H(x) \leq m^\alpha x \\
\Rightarrow \exp (-m^\alpha x) \leq \exp [-H(x)] \leq \exp \left[ - \frac{1}{m^\alpha} x \right] \\
\Rightarrow \frac{1}{m^\alpha} \leq \mu \leq m^\alpha. \tag{17}
\]
Thus decompose
\[
| \mu_{ij} - \tilde{\mu}_{ij} | = \left| \int_0^\infty \exp (-H_{ij}(s)) \, ds - \int_0^\infty \exp (-\tilde{H}_{ij}(s)) \, ds \right| \\
= \left| \int_0^{m^{-1}} \left[ \exp (-H_{ij}(s)) - \exp (-\tilde{H}_{ij}(s)) \right] \, ds \\
+ \int_T^{m^{-1}} \left[ \exp (-H_{ij}(s)) - \exp (-\tilde{H}_{ij}(s)) \right] \, ds \\
+ \int_T^\infty \left[ \exp (-H_{ij}(s)) - \exp (-\tilde{H}_{ij}(s)) \right] \, ds \right| \\
\leq O(m^{\alpha-1}) + \int_T^\infty \exp (-H_{ij}(s)) - \exp (-\tilde{H}_{ij}(s)) \, ds. \tag{18}
\]
But on $[T, \infty)$, since the log-hazard is constant, $H_{ij}(x) = H_{ij}(T) + e^b (x - T)$. Consequently, let $b' := \exp(b)$ and obtain
\[
I : \ = \int_T^\infty \exp (-H_{ij}(s)) \, ds = \exp \left\{ -H_{ij}(T) \right\} \frac{1}{b'}, \tag{19}
\]
\[
\tilde{I} : \ = \int_T^\infty \exp (-\tilde{H}_{ij}(s)) \, ds = \exp \left\{ -\tilde{H}_{ij}(T) \right\} \frac{1}{c'}. \tag{20}
\]
Now notice that, $\ln I - \ln \tilde{I} = \tilde{H}_{ij}(T) - H_{ij}(T) + (c - b) = O(m^{\alpha-1})$. Since $I \leq \mu \leq m^\alpha$, this implies that
\[
\left| \int_T^\infty \left\{ \exp [-H_{ij}(s)] - \exp [-\tilde{H}_{ij}(s)] \right\} \, ds \right| = O(m^{2\alpha-1}),
\]
and thus, plugging this back into (18), we get $|\mu_{ij} - \tilde{\mu}_{ij}| = O(m^{2\alpha - 1})$.

Now we are going to find a bound for the likelihood ratio, for which we need a few more results. Recall that one term that appears in the likelihood is the mean inter-visit time $\mu_{ii}$, which is obtained in two steps.

1. First,
   \[
   \mu_i = \sum_j \theta_{ij} \mu_{ij} \geq \frac{1}{m^\alpha},
   \]

2. and then,
   \[
   \xi_{ii} = \sum_j \pi_j \mu_j \geq \frac{m^{-\alpha}}{m^{-\beta}} = O(m^{\beta - \alpha}).
   \]

So that for all $j$,
   \[
   \left| \ln \frac{\xi_{jj}}{\xi_{ij}} \right| \leq \frac{|\xi_{jj} - \tilde{\xi}_{jj}|}{\xi_{jj} \land \xi_{ij}} \leq C m^{2\alpha - 1} m^{-\beta} = O(m^{2\alpha - \beta - 1}).
   \]

Also,
   \[
   \left| \ln \frac{\theta_{ij}}{\tilde{\theta}_{ij}} \right| \leq \frac{|\theta_{ij} - \tilde{\theta}_{ij}|}{\theta_{ij} \land \tilde{\theta}_{ij}} \leq m^{-1} m^{-\beta} = O(m^{-1}).
   \]

Now consider the log-likelihood ratio
   \[
   \ln \frac{l(w, \eta)}{l(w, \tilde{\eta})} = \left\{ \begin{array}{ll}
   \ln \frac{\theta_{j_{0j1}}}{\tilde{\theta}_{j_{0j1}}} - \ln \xi_{j_{0j0}} - \left[ H_{j_{0j1}}(x_0) - \tilde{H}_{j_{0j1}}(x_0) \right] \\
   \sum_{i=1}^{\tau-1} \left( \ln \frac{\theta_{j_{i+1j_i}}}{\theta_{j_{i+1i}}} + \ln \frac{h_{j_{i+1j_i}}(x_i)}{h_{j_{i+1j_i}}(x_i)} + \left[ H_{j_{0j1}}(x_i) - \tilde{H}_{j_{0j1}}(x_i) \right] \right) \\
   + \ln \left( \frac{\sum_{k=1}^{j} \theta_{j_{r-1k}} \exp \left[ -H_{j_{r-1k}}(x_r) \right] }{\sum_{k=1}^{j} \tilde{\theta}_{j_{r-1k}} \exp \left[ -\tilde{H}_{j_{r-1k}}(x_r) \right] } \right) 1 (\tau > 0) \\
   + \left\{ \ln \frac{\sum_{k=1}^{j} \theta_{j_{0k}} \exp \left[ -H_{j_{0k}}(x_0) \right]}{\sum_{k=1}^{j} \tilde{\theta}_{j_{0k}} \exp \left[ -\tilde{H}_{j_{0k}}(x_0) \right]} - \ln \frac{\xi_{j_{0j0}}}{\xi_{j_{0j0}}} \right\} 1 (\tau = 0). 
   \end{array} \right.
   \]

We also have a bound for
   \[
   \ln \frac{\sum_{j=1}^{J} \theta_{ij} \exp \left[ -H_{ij}(x) \right]}{\sum_{j=1}^{J} \tilde{\theta}_{ij} \exp \left[ -\tilde{H}_{ij}(x) \right]} \leq C m^{\alpha - 1} m^{-\beta} m^{-\alpha} = O(m^{2\alpha + \beta - 1}),
   \]
and since also $E_0 \tau < \infty$, we get that when $\eta, \tilde{\eta} \in \mathcal{O}_k^n$, 
\[
E_0 \ln \frac{l(w, \eta)}{l(w, \tilde{\eta})} = O(m^{\beta-1}) + O(m^{3\alpha-\beta-1}) + O(m^{\alpha-1}) + O(m^{2\alpha+\beta-1}) \\
= O(m^{(2\alpha+\beta-1)\vee(3\alpha-\beta-1)}) .
\]
This converges to zero for $(\alpha, \beta)$ in the polyhedron with interior 
\[
0 < \alpha < \frac{1}{2}, \quad 0 < \beta < 1 - 2\alpha, \text{ and } \beta > 3\alpha - 1 .
\]
Call $\gamma = \gamma(\alpha, \beta) := (2\alpha + \beta - 1) \vee (3\alpha - \beta - 1)$ and observe that for instance, if $\alpha = \beta = 0.25$ we get $\gamma = -0.25$.

Now consider sequences $\eta_k \in \mathcal{O}_k^n$ and $\eta_m \in \mathcal{O}_k^n$, chosen so that (13) holds. Then,
\[
E_0 \left\{ \ln \frac{l(w, \mathcal{O}_k^m)}{l(w, \eta_m)} \right\} = E_0 \left\{ \ln \frac{l(w, \mathcal{O}_k^m)}{l(w, \eta_k)} \right\} + E_0 \left\{ \ln \frac{l(w, \eta_k)}{l(w, \eta_m)} \right\} \leq Cm^\gamma - \delta.
\]
In order to calculate $\rho_m$, let 
\[
\phi(t) := E_m \exp \left[ t \ln \frac{f(w, \mathcal{O}_k^m)}{f(w, \eta_m)} \right]
\]
and note since $\phi(0) = 1$ and $\phi'(0) \leq Cm^\gamma - \delta$, there is a bound $\rho_m \leq 1 + Cm^\gamma - \delta$.

We can hence finally bound the series in the theorem as 
\[
\sum_{n=1}^{\infty} l_{m, \rho_m}^m(n) \leq \sum_{n=1}^{\infty} (cm)^m [1 + Cm^\gamma - \delta]^n ,
\]
which converges for $m = O(n^{1-\varepsilon})$ by Cauchy’s convergence test.

**References**


Figure 3: Non-penalized Estimates of the Log-hazard functions in a General SMP with 50 windows

Solid line := True model, Dashed line:= Spline Estimate, Dotted line:= Weibull Estimate
Figure 4: (cont.) Non-penalized Estimates of the Log-hazard functions in a General SMP with 50 windows

Solid line := True model, Dashed line:= Spline Estimate, Dotted line:= Weibull Estimate
Figure 5: Non-penalized Estimates of the density functions in a General SMP with 50 windows

Solid line := True model, Dashed line:= Spline Estimate, Dotted line:= Weibull Estimate
Figure 6: (cont.) Non-penalized Estimates of the density functions in a General SMP with 50 windows

Solid line := True model, Dashed line:= Spline Estimate, Dotted line:= Weibull Estimate