Penalized maximum likelihood estimation for generalized linear point processes

Niels Richard Hansen[1], *University of Copenhagen*

Abstract:

A framework of generalized linear point process models (glppm) much akin to glm for regression is developed where the intensity depends upon a linear predictor process through a known function. In the general framework the parameter space is a Banach space. Of particular interest is when the intensity depends on the history of the point process itself and possibly additional processes through a linear filter, and where the filter is parametrized by functions in a Sobolev space. We show two main results. First we show that for a special class of models the penalized maximum likelihood estimate is in a finite dimensional subspace of the parameter space – if it exists. In practice we can find the estimate using a finite dimensional glppm framework. Second, for the general class of models we develop a descent algorithm in the Sobolev space. We conclude the paper by a discussion of additive model specifications.

1 Introduction

Statistics for point process models is by now a vast and mature subject with a range of applications. Survival analysis or more generally event history analysis is perhaps the most notable area of application of one-dimensional point processes – or in the one-dimensional case we could equivalently say counting processes – with a large body of well developed theory. An authoritative treatment is Andersen et al. (1993). Other classical references include Fleming & Harrington (1991) and Karr (1991). The setup for statistical analysis of event history models is characterized by observing the occurrence of events – or transitions between states – for a collection of individuals. The modeling is based on intensities and it is paramount to incorporate covariate effects and be able to handle censoring mechanisms.

Many other important applications of one-dimensional point processes exist such as queuing and telecommunication systems, Asmussen (2003), insurance mathematics, Mikosch (2004), earthquakes, Ogata & Katsura (1986), Ogata et al. (2003), neuronal activity, Brillinger (1992), Paninski (2004), Pillow et al. (2008), and high-frequency financial modeling Hautsch (2004), just to mention some.

A major motivation for the present paper comes from yet another application. With the sequencing of the human genome and the subsequent sequencing of many other genomes the ground has been laid for analyzing and interpreting the blueprints of life. We analyze the static genomes that consist of long DNA sequences and try to identify the collection of functional elements that are written is this DNA code. We find the protein coding genes

[1]Postal address: Department of Mathematical Sciences, University of Copenhagen Universitetsparken 5, 2100 Copenhagen Ø, Denmark.
Email address: richard@math.ku.dk

but also a myriad of other important features such as regulatory elements, Maston et al. (2006). In the analysis of the genomic data a typical question is whether the occurrence of a given feature or sequence motif is entirely random as opposed to being organized in some non-random way. The traditional use of point processes in this area is mostly limited to specifying a null or reference distribution for randomness – a common choice is here the homogeneous Poisson point process. Deviations for the data from the null distribution is taken as evidence for the existence of an organizational structure in the data of some biological significance.

One attempt to go beyond the Poisson process null model and actually model the occurrences of certain motifs in the DNA-sequences is found in Gusto & Schbath (2005), which was also an important inspiration for our further work. The linear Hawkes processes, as used in Gusto & Schbath (2005), and the general class of multivariate, non-linear Hawkes processes, as treated in Brémaud & Massoulié (1996), were considered in our further development of models appropriate for genomic organization. We noted a structural similarity of the models to the generalized linear models, and this has played a role in the implementation, Carstensen et al. (2010). The similarity, which we for sure are not the first to observe, is implicitly present in several of the popular models for survival analysis, such as Cox's regression model and Aalens additive model, where the intensity is specified through a fixed function of a linear combination of covariates. A more direct relation to the log-linear Poisson model is illustrated in Example VI.1.3 in Andersen et al. (1993). See also Whitehead (1980) and Aitkin et al. (2005). The terminology of a generalized linear point process model has, furthermore, been used recently for various Hawkes-type models of spike trains for neurons, Paninski (2004), Pillow et al. (2008), Toyoizumi et al. (2009). The models considered in Pillow et al. (2008) for multivariate spike trains share many components with our models of the occurrences of multiple transcription regulatory elements. In particular, the use of basis expansions for estimation of functional components, which may be combined with regularization in terms of penalized maximum-likelihood estimation. In Pillow et al. (2008) the basis functions chosen were raised cosines with a log-time transformation, whereas we used B-splines in Carstensen et al. (2010).

Motivated by the different applications described above we ask if there are theoretical results supporting any particular choice of basis functions. Or phrased differently, if we can understand a particular choice of basis functions as the solution of a more abstractly formulated problem. Clearly we have the classical result on smoothing splines in mind, which shows that splines appear as the solution of a particular penalized least squares problem, Theorem 2.4 in Green & Silverman (1994). To proceed we first develop a formal and abstract framework of generalized linear point process models (glppm) parametrized by a Banach space, and then we show two main results for a general class of models that includes the Hawkes processes as a special case. The first result we show is similar to the result on smoothing splines, and it states that the penalized maximum-likelihood estimator for a specific model is found in a finite-dimensional space spanned by an explicit set of basis functions. For the linear Hawkes process the solution is a spline. The second result is dif-

ferent. For the general model class considered we do not find an explicit finite-dimensional basis. In the alternative we derive an infinite-dimensional gradient, which suggest an iterative algorithm, and we establish a convergence result for this algorithm. The interpretation of the algorithm is as a sequence of finite-dimensional subspace approximations.

The purpose of the present paper is to provide the theoretical framework for the computation of penalized maximum-likelihood estimators for functional parameters in a one-dimensional point process setup. For a treatment of properties of penalized maximum-likelihood estimators we refer to Cox & O'Sullivan (1990). The focus is here on the representation and computation.

2 Setup

We consider a filtered probability space – a stochastic basis – $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ were the filtration is assumed to be right continuous. We will, in addition, assume that $(N_t)_{t\geq 0}$ is an adapted counting process, which, under P, is a homogeneous Poisson process with rate 1.

If $(\lambda_t)_{t\geq 0}$ is a positive, predictable process we can define the positive process, known as the likelihood process, by

$$\mathcal{L}_t = \exp\left(t + \int_0^t \log \lambda_s N(\mathrm{d}s) - \Lambda_t\right), \qquad \Lambda_t = \int_0^t \lambda_s \mathrm{d}s. \tag{1}$$

We will assume that $\Lambda_t < \infty$ P-a.s., in which case $(\mathcal{L}_t)_{t\geq 0}$ in general is a P-local martingale and a P-supermartingale with $\mathbb{E}_P(\mathcal{L}_t) \leq 1$ for all $t \geq 0$, Theorem VI.T2, Brémaud (1981). If $\mathbb{E}_P(\mathcal{L}_t) = 1$ we can define a probability measure Q_t on $\mathcal{F}$ by taking $\mathcal{L}_t$ to be the Radon-Nikodym derivative of Q_t w.r.t. P. That is,

$$Q_t = \mathcal{L}_t \cdot P. \tag{2}$$

We note that $\mathbb{E}_P(\mathcal{L}_t) = 1$ if and only if $(\mathcal{L}_s)_{0\leq s\leq t}$ is a true P-martingale. If $\mathbb{E}_P(\mathcal{L}_t) < 1$ we can not define a probability measure Q_t on the abstract space $(\Omega, \mathcal{F})$ by (2). With a more explicit, canonical choice of Ω it is possible always to construct a measure Q_t such that

$$Q_t = \mathcal{L}_t \cdot P + Q_t^{\perp}$$

where $Q_t^{\perp}(N_t < \infty) = 0$, see Jacod (1975) or Theorem 5.2.1(ii), Jacobsen (2006).

Throughout we will fix an observation window $[0, t]$. The process $(\lambda_s)_{0\leq s\leq t}$ is called the (predictable) intensity process for the counting process $(N_s)_{0\leq s\leq t}$ under Q_t. The integrated intensity, $(\Lambda_s)_{0\leq s\leq t}$, is the compensator, and if $\mathbb{E}_P(\mathcal{L}_t) = 1$ the process $M_s = N_s - \Lambda_s$ for $s \in [0, t]$ is a Q_t-martingale, Theorem VI.T3, Brémaud (1981).

From a model building perspective the direct specification of the intensity process is natural as well as practical. Practical because the construction of probability models for a

parametrized family of intensity processes, $(\lambda_t(\beta))_{t\geq 0}$, for $\beta \in \Theta$, through the likelihood process construction above immediately yields the likelihood function $\mathcal{L}_t(\beta)$ for subsequent statistical inference. There is one small caveat though. For the specification of the probability models to lead to a statistical model dominated by P all the likelihood processes need to be true P-martingales. At least on $[0,t]$, which is equivalent to $\mathbb{E}_P(\mathcal{L}_t(\beta)) = 1$ for all $\beta \in \Theta$. This is a technical obstacle, and it does not seem to be easy to formulate a simple, general criteria. The problem is equivalent to checking whether the intensities specify non-exploding point processes on canonical spaces. When there is positive probability of explosion for some measures, and the model is thus not dominated by P, it is anyway sensible to compare two models, Q_t^β and $Q_t^{\beta'}$, in terms of the Radon-Nikodym derivatives

$$\frac{\mathrm{d}Q_t^\beta}{\mathrm{d}(Q_t^\beta + Q_t^{\beta'})} \quad \text{and} \quad \frac{\mathrm{d}Q_t^{\beta'}}{\mathrm{d}(Q_t^\beta + Q_t^{\beta'})},$$

see page 893, Kiefer & Wolfowitz (1956). If we have *non-exploding* data on $[0,t]$, this comparison is equivalent to comparing $\mathcal{L}_t(\beta)$ with $\mathcal{L}_t(\beta')$, and though $\mathcal{L}_t(\beta)$ is not necessarily a true likelihood function it provides a sensible relative measure of the models parametrized by β – but only for non-exploding data. It seems that we do not need to check if $\mathbb{E}_P(\mathcal{L}_t(\beta)) = 1$, but this is a delusion. If the process we observe is not exploding on $[0,t]$ – and there may often be subject matter reasons it is not – all models with $\mathbb{E}_P(\mathcal{L}_t(\beta)) < 1$ are misspecified in a fundamental way. Arguably, the models specified by $\tilde{Q}_t^\beta = \tilde{\mathcal{L}}_t(\beta) \cdot P$ with

$$\tilde{\mathcal{L}}_t(\beta) = \frac{\mathcal{L}_t(\beta)}{\mathbb{E}_P(\mathcal{L}_t(\beta))}$$

are more appropriate, which is equivalent to conditioning on non-explosion. However, $(\lambda_t(\beta))_{0\leq s\leq t}$ is no longer the intensity process under $\tilde{Q}_t^\beta$ and the likelihood, $\tilde{\mathcal{L}}_t(\beta)$, is only known up to a normalizing constant, which in general is complicated to compute. We will not pursue this direction any further.

We proceed with the general setup and let V denote a separable Banach space with the norm $||\cdot||$, and V^* is its dual space of continuous linear functionals equipped with the dual norm. The dual norm is also denoted $||\cdot||$, which turns V^* into a Banach space as well. Due to separability of V the dual space V^* is separable and second countable in the weak*-topology (see e.g. Exercise E.2.5.3, Pedersen (1989)). We equip V^* with the weak* Borel σ-algebra, which then coincides with the σ-algebra generated by the linear functionals

$$x \mapsto x\beta$$

for $\beta \in V$. We then consider an adapted, norm-càdlàg stochastic process $(X_s)_{0\leq s\leq t}$ with values in V^*. That is, X_s is an $\mathcal{F}_s$-measurable, random variable with values in V^* and we assume that the sample path of the process is càdlàg in the norm topology so that for all ω it holds that

$$\lim_{\varepsilon \to 0+} ||X_{s+\varepsilon}(\omega) - X_s(\omega)|| = 0,$$

and there is an $X_{s-}(\omega) \in V^*$ such that

$$\lim_{\varepsilon \to 0+} ||X_{s-\varepsilon}(\omega) - X_{s-}(\omega)|| = 0.$$

For any $\beta \in V$ the real valued process $(X_s\beta)_{0 \leq s \leq t}$ is then adapted and càdlàg, and $(X_{s-}\beta)_{0 \leq s \leq t}$ is predictable, cf. Proposition 2.6 in Jacod & Shiryaev (2003). We call $(X_{s-}\beta)_{0 \leq s \leq t}$ the *linear predictor process*. If $D \subseteq \mathbb{R}$ we introduce the set

$$\Theta(D) = \{\beta \in V \mid X_{s-}\beta \in D \text{ for all } s \in [0,t] \ P\text{-a.s.}\}.$$

Definition 2.1. *Assume that $\varphi : D \to [0, \infty)$ and assume in addition that $(Y_s)_{0 \leq s \leq t}$ is a predictable, càdlàg process with values in $[0, \infty)$. We define a generalized linear point process model on $[0,t]$ to be the statistical model for a point process on $[0,t]$ with parameter space $\Theta(D)$ such that for $\beta \in \Theta(D)$ the point process has intensity*

$$\lambda_s = Y_s \varphi(X_{s-}\beta)$$

for $s \in [0,t]$.

For $\beta \in \Theta(D)$ we have the likelihood process $\mathcal{L}_t(\beta)$ given by (1) in terms of the intensity defined above. For the general definition we do not require it to be a martingale, and it plays no role for the results and computations in the present paper. However, for interpretations and to obtain sensible models via penalized maximum-likelihood estimation we certainly need to be able to verify if the process is a martingale. We discuss some possibilities below.

The Y-process in the definition serves the same purpose as in survival analysis, that is, it can be a simple *at risk* indicator process, but we keep it in the definition as a general, predictable, non-negative process. Note that if φ is one-to-one with inverse $m = \varphi^{-1} : \varphi(D) \to D$ then in the absence of the Y-process we have

$$X_s\beta = m(\lambda_s).$$

Drawing an analogy to ordinary generalized linear models it seems natural at this point to call m the link function – it transforms the intensity process into a process that is linear in the parameter β. With this terminology we would call φ the inverse link function. However, in general there is no reason to require φ to be one-to-one, and we will not use the terminology.

Whether the intensity in Definition 2.1 gives rise to a likelihood process, which is a true martingale, can depend quite heavily on the choice of φ. If φ is bounded the martingale condition is easy to verify, cf. Theorem VI.T4 in Brémaud (1981). If, on the other hand, $(X_s)_{0 \leq s \leq t}$ is independent of $(N_s)_{0 \leq s \leq t}$ under P and $(Y_s)_{0 \leq s \leq t}$ is bounded, say, then $E_P(\mathcal{L}_t) = 1$ disregarding the choice of φ. To give one additional criteria we assume for simplicity that $Y_s = 1$ and Ω is the canonical space of counting processes. Then we can,

as mentioned above, construct a measure Q such that $\lambda_s = \varphi(X_{s-}\beta)$ is the intensity for the counting process under Q. If $||X_{s-}|| \leq CN_{s-} + D$ and $\varphi(x) \leq c|x| + d$ it follows that

$$\lambda_s \leq \alpha N_{s-} + \gamma.$$

According to Example 4.4.5 in Jacobsen (2006) the counting process is not exploding under Q and by Theorem 5.2.1(ii) in Jacobsen (2006) the likelihood process is a true martingale. A more refined treatment for the class of non-linear Hawkes processes focusing on stability in the sense of (asymptotic) stationarity is found in Brémaud & Massoulié (1996).

When the likelihood process is a martingale it is evident from (1) that as a statistical model with parameter space $\Theta(D) \subseteq V$ the minus-log-likelihood function for observing $(N_s)_{0\leq s\leq t}$ is

$$l_t(\beta) = \int_0^t Y_s\varphi(X_{s-}\beta)\mathrm{d}s - \int_0^t \log(Y_s\varphi(X_{s-}\beta))N(\mathrm{d}s) \tag{3}$$

for $\beta \in \Theta(D)$. Note that if $\Delta N_s = 1$ but $Y_s = 0$ then $l_t(\beta) \equiv \infty$, which simply tells us that the model as formulated is inappropriate for the data. In the following we therefore assume that this is not the case, that is, $Y_s > 0$ for all s with $\Delta N_s = 1$.

For practical applications – even when V is finite dimensional – the maximum-likelihood estimator may not be well defined. One solution is to introduce a penalty function $J : \Theta(D) \to \mathbb{R}$ and then to minimize the function

$$l_t(\beta) + J(\beta)$$

instead. We provide examples below.

The minus-log-likelihood function is simple as a function of β and if φ is convex and log-concave we see that l_t is convex as well. The penalty function is typically also chosen to be convex.

The generalized linear models for point processes have value even when V is finite dimensional, but we emphasize that models of considerably greater generality fit into the model class above for a suitable choice of X-process. This is at least true from a practical point of view where finite basis expansions can be used to approximate non-parametric components, and we also show one result in Section 3 where penalized maximum-likelihood estimation in an infinite dimensional function space reduces to penalized maximum likelihood estimation for a generalized linear model with a finite dimensional parameter space. Here we give a simple but well known example of how Cox's regression model fits into the framework.

Example 2.2. The Cox proportional hazards model can be (re)formulated as a generalized linear point process model. We take $X_s \in \mathbb{R}^d$ to be an adapted, d-dimensional càdlàg process, which is independent of $(N_s)_{0\leq s\leq t}$ under P. Then X_s^T is a process in $(\mathbb{R}^d)^*$ and the Cox model is specified by the intensity

$$\lambda_s = \exp(X_{s-}^T\beta)\alpha(s)$$

where $\alpha(s)$ is the baseline intensity and $\beta \in \mathbb{R}^d$. If

$$\log \alpha(s) = B_{s-}\beta_\alpha$$

where $B_s \in V^*$ is a known, adapted, norm-càdlàg process with values in the dual of V and $\beta_\alpha \in V$ we can rewrite the intensity as

$$\lambda_s = \exp\left((X_{s-}^T \; B_{s-}) \begin{pmatrix} \beta \\ \beta_\alpha \end{pmatrix} \right) = \exp(X_{t-}^T\beta + B_{s-}\beta_\alpha) = \exp(X_{s-}^T\beta)\exp(B_{s-}\beta_\alpha),$$

which is a generalized linear point process model with $\varphi = \exp$, with domain $D = \mathbb{R}$, and with parameter space $\mathbb{R}^d \times V$.

It is hardly conceivable that we can estimate the parameters in a sensible way for a single observation of the counting process, and in practice we will use the model with independent replications and corresponding intensities, $\lambda^1, \ldots, \lambda^n$, possibly even multiplied by an *at risk* indicator processes. Even so, it may still be desirable to penalize the β_α parameter to obtain a smooth fit of α, and a possible choice of penalty function is

$$J(\beta) = \lambda ||\beta_\alpha||$$

for $\lambda > 0$. In practice we may have $V = \mathbb{R}^{d'}$ and $B_t = (B_{t,1}, \ldots, B_{t,d'})$ where $B_{t,1}, \ldots, B_{t,d'}$ are known (deterministic) basis functions. If the basis functions are C^2 in t a natural norm on $\mathbb{R}^{d'}$ is given by the quadratic form K where

$$K_{i,j} = \int_0^t B''_{s,i} B''_{s,j} \mathrm{d}s.$$

In this case,

$$J(\beta) = \lambda \beta_\alpha^T K \beta_\alpha = \lambda \int_0^t \left([\log \alpha(s)]'' \right)^2 \mathrm{d}s,$$

which is a popular choice of penalty term that we will consider below.

Before turning to more concrete models we make one general observation about the derivatives of the minus-log-likelihood under the assumption that φ is differentiable.

Proposition 2.3. *If $D \subseteq \mathbb{R}$ is open and if φ is C^1 on D then l_t is Gâteaux differentiable in $\beta \in \Theta(D)^\circ$ if $l_t(\beta) < \infty$ with derivative*

$$Dl_t(\beta) = \int_0^t Y_s \varphi'(X_{s-}\beta) X_{s-} \mathrm{d}s - \int_0^t \frac{\varphi'(X_{s-}\beta)}{\varphi(X_{s-}\beta)} X_{s-} N(\mathrm{d}s). \tag{4}$$

Moreover, if φ is C^2 the second Gâteaux derivative is

$$\begin{aligned} D^2 l_t(\beta) \;=\; & \int_0^t Y_s \varphi''(X_{s-}\beta) X_{s-} \otimes X_{s-} \mathrm{d}s \\ & - \int_0^t \frac{\varphi''(X_{s-}\beta)\varphi(X_{s-}\beta) - \varphi'(X_{s-}^T\beta)^2}{\varphi(X_{s-}\beta)^2} X_{s-} \otimes X_{s-} N(\mathrm{d}s) \end{aligned} \tag{5}$$

If J is Gâteaux differentiable the penalized maximum likelihood estimator in $\Theta(D)^\circ$, if it exists, is then a solution to the equation $Dl_t(\beta) + \lambda DJ(\beta) = 0$.

One way to interpret the process $(X_{s-}\beta)_{0\leq s\leq t}$ is as a predictable, linear filter of the Banach space valued process $(X_s)_{0\leq s\leq t}$. The possible linear filters are parametrized by $\beta \in \Theta(D)$, and the objective from a statistical point of view is the estimation of β.

In Section 3 below we restrict our attention to stochastic processes with values in a reproducing kernel Hilbert space (RKHS), which are given through stochastic integration w.r.t. an ordinary real valued stochastic process. In Section 4 we generalize the class of models to an additive model framework, where the parameter space is a product of RKHSs. The product space can be equipped with an inner product that turns it into a Hilbert space, but it can also be equipped with a 1-norm, which turns the product space into a Banach space. In the latter case we discuss how the natural penalization lead to an infinite dimensional version of a lasso estimator.

3 Linear filters from stochastic integration

Let $g : [0, \infty) \to \mathbb{R}$ be a measurable function and $(Z_s)_{0\leq s\leq t}$ a càdlàg semi-martingale. If g is e.g. locally bounded (which is sufficient for our purposes) the stochastic process

$$\int_0^s g(s-u)\mathrm{d}Z_u$$

is a well defined càdlàg process. The process is sometimes called a homogeneous linear filter or a moving average.

The parameter space we will consider is $V = W^{m,2}([0,t])$, that is, V is the Sobolev space of functions that are m times weakly differentiable with the m'th derivative in $L_2([0,t])$. For this concrete parameter space we will use g for the generic parameter – in contrast to the abstract notation where we use β.

We will need to interpret the stochastic integral above as a stochastic process with values in V^*. Since the stochastic integral is not defined pathwise in general, it is in fact not obvious that

$$g \mapsto X_s g := \int_0^s g(s-u)\mathrm{d}Z_u$$

for a fixed sample path is even a well defined linear functional – let alone continuous. For the pathwise definition of the stochastic integral as a linear functional we note that functions in $W^{m,2}([0,t])$ for $m \geq 1$ are weakly differentiable with L_2-derivatives. Hence by integration by parts, see e.g. Definition 4.45 and Proposition 4.49(b) in Jacod & Shiryaev (2003), we have that

$$\int_0^s h(u)\mathrm{d}Z_u \;=\; h(s)Z_s - h(0)Z_0 - \int_0^s Z_{u-}h'(s)\mathrm{d}u \tag{6}$$

for $h \in W^{m,2}([0,t])$. This equality is in general valid up to evanescence. The right hand side is pathwise well defined, thus we simply use the version of the stochastic integral defined by the right hand side above. The integral then obviously becomes a linear functional in h for a concrete realization of the Z-process. Combined with Corollary A.2 this shows that we can regard $(X_s)_{0\leq s\leq t}$ as a stochastic process with values in V^*. Lemma A.3 shows, moreover, that $(X_s)_{0\leq s\leq t}$ is norm-càdlàg with X_{s-} obviously given as

$$X_{s-}g = \int_0^{s-} g(s-u)\mathrm{d}Z_u.$$

If the function $\varphi : D \to [0,\infty)$ is given we find that $\Theta(D)$ consists of those g such that

$$\int_0^{s-} g(s-u)\mathrm{d}Z_u \in D \text{ for all } s \in [0,t] \quad P\text{-a.s.} \tag{7}$$

The Sobolev space $W^{m,2}([0,t])$ can be equipped with several inner products that give rise to equivalent norms and turn the space into a RKHS, Wahba (1990), Berlinet & Thomas-Agnan (2004). For each inner product there is an associated kernel, the reproducing kernel, and we assume here that one inner product is chosen with the corresponding norm denoted $||\cdot||$ and corresponding kernel denoted $R : [0,t]\times[0,t] \to \mathbb{R}$. Moreover, we fix $\varphi_1,\ldots,\varphi_l \in W^{m,2}([0,t])$ and denote by P the orthogonal projection onto $\mathrm{span}\{\varphi_1,\ldots,\varphi_l\}^\perp$. One of the defining properties of the kernel R is that for fixed $s \in [0,t]$, $R(s,\cdot) \in W^{m,2}([0,t])$, hence $PR(s,\cdot)$ is a well defined function. This give rise to the projected kernel, which we denote $R^1 = PR$. With this setup the penalty function we choose is $J(g) = \lambda||Pg||^2$ for $\lambda > 0$, and the penalized minus-log-likelihood function reads

$$l_t(g) + \lambda||Pg||^2 \tag{8}$$

for $g \in \Theta(D)$ where

$$l_t(g) = \int_0^t Y_s\varphi\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right)\mathrm{d}s - \int_0^t \log(Y_s\varphi\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right))N(\mathrm{d}s).$$

With $\tau_1,\ldots,\tau_{N_t}$ denoting the jump times for the counting process $(N_s)_{0\leq s\leq t}$ we can state one of the main theorems.

Theorem 3.1. *If $\varphi(x) = x + d$ with domain $D = [-d,\infty)$ then a minimizer of (8) over $\Theta(D) \subseteq W^{m,2}([0,t])$, $m \geq 1$, belongs to the finite dimensional subspace of $W^{m,2}([0,t])$ spanned by the functions $\varphi_1,\ldots,\varphi_l$, the functions*

$$h_i(r) = \int_0^{\tau_i-} R^1(\tau_i - u, r)\mathrm{d}Z_u$$

for $i = 1,\ldots,N_t$ together with the function

$$f(r) = \int_0^t Y_s \int_0^{s-} R^1(s-u,r)\ \mathrm{d}Z_u\mathrm{d}s.$$

Remark 3.2. A practical consequence of Theorem 3.1 is that by collecting $\varphi_1(r), \ldots, \varphi_l(r)$, $f(r)$ and $h_i(r)$, $i = 1, \ldots, N_t$, in an $l+1+N_t$ dimensional vector we reduce the estimation problem to a finite dimensional optimization problem. For the concrete realization we may of course choose whichever basis that is most convenient for this function space. For the practical computation of f we note that by Lemma A.5 we can interchange the order of the integrations so that

$$f(r) = \int_0^t \int_u^t Y_s R^1(s-u, r) \, \mathrm{d}s \mathrm{d}Z_u. \tag{9}$$

Remark 3.3. It is a common trick to construct a model conditionally on the entire outcome of a process $(Z_s)_{0 \leq s \leq t}$ by assuring that Z_s is $\mathcal{F}_0$-measurable for all $s \in [0, t]$. In this case the process

$$\int_0^t g(|s-u|) \mathrm{d}Z_u$$

for $s \in [0, t]$ is obviously predictable. Theorem 3.1 still holds with the modification that

$$h_i(r) = \int_0^t R^1(|\tau_i - u|, r) \mathrm{d}Z_u$$

for $i = 1, \ldots, N_t$ and

$$f(r) = \int_0^t Y_s \int_0^t R^1(|s-u|, r) \mathrm{d}Z_u \mathrm{d}s.$$

When we model events that happen in time it is most natural that the intensity at a given time t only depends on the behavior of the Z-process up to just before t. This corresponds to the formulation chosen in Theorem 3.1. However, if we model events in a one-dimensional space it is often more natural to take the approach in this remark.

One useful choice of inner product on $W^{m,2}([0, t])$ is given as follows. Take

$$\mathcal{H}_1 = \{f \in W^{m,2}([0, t]) \mid f(0) = Df(0) = \ldots = D^{m-1} f(0) = 0\},$$

which we equip with the inner product

$$< f, g >= \int_0^t D^m f(s) D^m g(s) \mathrm{d}s.$$

This turns $\mathcal{H}_1$ into a reproducing kernel Hilbert space for $m \geq 1$ with reproducing kernel $R^1 : [0, t] \times [0, t] \to \mathbb{R}$ given as

$$R^1(s, r) = \int_0^{s \wedge r} \frac{(s-u)^{m-1}(r-u)^{m-1}}{((m-1)!)^2} \mathrm{d}u,$$

see Wahba (1990). Furthermore, define $\varphi_k(t) = t^{k-1}/(k-1)!$ for $k = 1, \ldots, m$ and

$$\mathcal{H}_0 = \mathrm{span}\{\varphi_1, \ldots, \varphi_m\},$$

which we equip with the inner product

$$< \sum_i a_i \varphi_i, \sum_j b_j \varphi_j >= \sum_{i,j} a_i b_j,$$

so that $\varphi_1, \ldots, \varphi_m$ is an orthonormal basis for $\mathcal{H}_0$. Then $\mathcal{H}_0$ is also a reproducing kernel Hilbert space with reproducing kernel $R^0 : [0,t] \times [0,t] \to \mathbb{R}$ defined by

$$R^0(s,r) = \sum_{k=1}^{m} \varphi_k(s)\varphi_k(r).$$

Then the Sobolev space $W^{m,2}([0,t]) = \mathcal{H}_0 \oplus \mathcal{H}_1$ is a reproducing kernel Hilbert space with reproducing kernel $R(s,r) = R^0(s,r) + R^1(s,r)$, $\mathcal{H}_0 \perp \mathcal{H}_1$, and with P the orthogonal projection onto $\mathcal{H}_1$, $PR = R^1$ and

$$J(g) = \int_0^t (D^m g(s))^2 \mathrm{d}s.$$

It follows by the definition of R that $R^1(s,\cdot)$ for fixed s is a piecewise polynomial of degree $2m-1$ with continuous derivatives of order $2m-2$, that is, $R(s,\cdot)$ is an order $2m$ spline. We find that e.g. the h_i-functions for the basis in Theorem 3.1 are given as stochastic integrals of order $2m$ splines.

Example 3.4. If $(Z_s)_{0\leq s\leq t}$ itself is a counting process and $\varphi(x) = x + d$ as in Theorem 3.1 we can give a more detailed description of the minimizer of (8) over $\Theta(D)$. We will also assume that the Y-process is identically 1. If $\sigma_1, \ldots, \sigma_{Z_t}$ denote the jump times for $(Z_s)_{0\leq s\leq t}$ we find that

$$h_i(r) = \sum_{j:\sigma_j<\tau_i} R^1(\tau_i - \sigma_j, r).$$

Collectively, the h_i basis functions are order $2m$ splines with knots in

$$\{\tau_i - \sigma_j \mid i = 1, \ldots, N_t,\ j : \sigma_j < \tau_i\}.$$

Due to (9) the last basis function, f, is seen to be an order $2m+1$ spline with knots in

$$\{t - \sigma_j \mid i = 1, \ldots, Z_t\}.$$

The cubic splines, $m = 2$, are the splines mostly used in practice. Here

$$R(s,r) = \int_0^{s\wedge r} (s-u)(r-u)\mathrm{d}u = sr(s\wedge r) - \frac{(s+r)(s\wedge r)^2}{2} + \frac{(s\wedge r)^3}{3}$$

and we can compute the integrated functions that enter in f as follows. If $t - u < r$

$$\int_u^t R(s-u,r)\mathrm{d}s = \int_0^{t-u} R(s,r)\mathrm{d}s \quad = \quad \frac{r(t-u)^3}{6} - \frac{(t-u)^4}{24}$$

and if $t-u \geq s$

$$\begin{aligned}\int_u^t R(s-u,r)\mathrm{d}s &= \int_0^{t-u} R(s,r)\mathrm{d}s = \frac{3r^4}{24} + \int_r^{t-u} R(s,r)\mathrm{d}r \\ &= \frac{r^4}{24} + \frac{r^2(t-u)^2}{4} - \frac{r^3(t-u)}{6}.\end{aligned}$$

Thus the function f is a sum of functions, the j'th function being a degree 4 polynomial on $[0, t-\sigma_j]$ and an affine function on $(t-\sigma_j, t]$.

If $Z_s = N_s$ the process $(N_s)_{0\leq s\leq t}$ is under Q_t known as a *linear Hawkes process*, in which case the set of knots for the h_i-functions equals the collection of interdistances between the points.

Proposition 3.5. *If φ is continuously differentiable and $g \in \Theta(D)^\circ$ we define η_i for $i = 1, \ldots, N_t$ as*

$$\eta_i(r) = \int_0^{\tau_i-} R(\tau_i - u, r)\mathrm{d}Z_u$$

and

$$f_g(r) = \int_0^t \int_u^t Y_s \varphi'\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right) R(s-u,r)\mathrm{d}s\mathrm{d}Z_u.$$

Then the gradient of l_t in g is

$$\nabla l_t(g) = f_g - \sum_{i=1}^{N_t} \frac{\varphi'\left(\int_0^{\tau_i-} g(\tau_i-u)\mathrm{d}Z_u\right)}{\varphi\left(\int_0^{\tau_i-} g(\tau_i-u)\mathrm{d}Z_u\right)}\eta_i.$$

The explicit derivation of the gradient above has several interesting consequences. First, a necessary condition for $g \in \Theta(D)^\circ$ to be a minimizer of the penalized minus-log-likelihood function is that g solves $\nabla l_t(g) + 2\lambda P g = 0$, which yields an integral equation in g. The integral equation is hardly solvable in any generality, but for $\varphi(x) = x + d$ it does provide the same information as Theorem 3.1 for interior minimizers – that is, a minimizer must belong to the given finite dimensional subspace of $W_2^m([0,t])$. The gradient can be used for descent algorithms. Inspired by the gradient expression we propose a generic algorithm, Algorithm 3.6, for subspace approximations. We consider here only the case where $D = \mathbb{R}$ so that $\Theta(D) = W^{m,2}([0,t])$. The objective function that we attempt to minimize with Algorithm 3.6 is

$$\Lambda(g) = l_t(g) + \lambda||Pg||^2$$

with gradient $\nabla\Lambda(g) = \nabla l_t(g) + 2\lambda P g$. We assume here that φ is continuously differentiable. To show a convergence result we need to introduce a condition on the steps of the algorithm, and for this purpose we introduce for $0 < c_1 < c_2 < 1$ and $\delta \in (0,1)$ fixed and $g \in W^{m,2}([0,t])$ the subset

$$W(g) = \left\{ \tilde{g} \in W^{m,2}([0,t]) \;\middle|\; \begin{array}{rcl} \Lambda(\tilde{g}) & \leq & \Lambda(g) + c_1 < \nabla\Lambda(g), \tilde{g} - g > \\ < \nabla\Lambda(\tilde{g}), \tilde{g} - g > & \geq & c_2 < \nabla\Lambda(g), \tilde{g} - g > \\ - < \nabla\Lambda(g), \tilde{g} - g > & \geq & \delta||\nabla\Lambda(g)||\,||\tilde{g} - g|| \end{array} \right\}$$

The two first conditions determining $W(g)$ above are known as the *Wolfe conditions* in the literature on numerical optimization, Nocedal & Wright (2006). The third is an *angle condition*, which is automatically fulfilled if $\tilde{g} - g = -\alpha\nabla\Lambda(g)$ for $\alpha > 0$. In Algorithm 3.6 we need to iteratively choose $\hat{g}_h$, and we show that if $\nabla\Lambda(\hat{g}_{h-1}) \neq 0$ then under the assumptions in Theorem 3.7 below

$$W(\hat{g}_{h-1}) \cap \text{span}\{\hat{g}_{h-1}, \nabla\Lambda(\hat{g}_{h-1})\} \neq \emptyset, \tag{10}$$

which makes the iterative choices possible.

Algorithm 3.6. Initialize; fix c_1, c_2 with $0 < c_1 < c_2 < 1$ and $\delta \in (0, 1)$, set

$$f_0(r) = \int_0^t \int_u^t Y_s R^1(s-u, r)\mathrm{d}s\mathrm{d}Z_u,$$

let $\hat{g}_0 \in \text{span}\{\eta_1, \ldots, \eta_{N_t}, f_0\}$ and set $h = 1$.

1. Stop if $\nabla\Lambda(\hat{g}_{h-1}) = 0$. Otherwise choose

   $$\hat{g}_h \in W(g_{h-1}) \cap \text{span}\{\eta_1, \ldots, \eta_{N_t}, f_0, \ldots, f_{h-1}\}$$

   where $W(g_{h-1})$ as defined above depends on c_1, c_2 and δ.

2. Compute

   $$f_h(r) = \int_0^t \int_u^t Y_s \varphi'\left(\int_0^{s-} \hat{g}_h(s-u)\mathrm{d}Z_u\right) R^1(s-u, r)\mathrm{d}s\mathrm{d}Z_u.$$

3. Set $h = h + 1$ and return to 1.

Note that the computation of f_h is just as in (9) except that the Y-process is iteratively updated.

Theorem 3.7. *If $D = \mathbb{R}$, if φ is strictly positive, twice continuously differentiable and if the sublevel set*

$$\mathcal{S} = \{g \in \Theta(D) \mid \Lambda(g) \leq \Lambda(\hat{g}_0)\}$$

is bounded then Algorithm 3.6 is globally convergent in the sense that

$$||\nabla\Lambda(\hat{g}_h)|| \to 0$$

for $h \to \infty$.

If we for instance have strict convexity of Λ then under the assumptions in Theorem 3.7 we have a unique minimizer in $\mathcal{S}$. Then we can strengthen the conclusion about convergence and get weak convergence of $\hat{g}_h$ towards the minimizer. In particular, we have the following corollary.

Corollary 3.8. *If there is a unique minimizer, $\hat{g}$, of Λ in $\mathcal{S}$ then under the assumptions in Theorem 3.7*

$$\hat{g}_h(s) \to \hat{g}(s)$$

for $h \to \infty$ for all $s \in [0,t]$.

4 Additive models

We give in this section a brief treatment of how the setup in the previous section extends to the setup where the intensity is given in terms of a sum of linear filters. We restrict the discussion to the situation where $V = W^{m,2}([0,t])^d$ and $(Z_s)_{0\leq s\leq t}$ is a d-dimensional semi-martingale. Perceiving $g \in V$ as a function $g : [0,1] \to \mathbb{R}^d$ with coordinate functions in $W^{m,2}([0,t])$ we write

$$\int_0^s g(s-u)\mathrm{d}Z_u = \sum_{j=1}^d \int_0^s g_j(s-u)\mathrm{d}Z_{j,u}$$

and just as above, by Corollary A.2,

$$g \mapsto X_s g := \int_0^s g(s-u)\mathrm{d}Z_u$$

is a continuous linear function on V when equipped with the product topology. The inner product $< g, h >= \sum_{j=1}^d < g_i, h_i >$ with corresponding norm $||g||^2 = \sum_{j=1}^d ||g_i||^2$ obviously turns V into a Hilbert space.

The minus-log-likelihood function is given just as in the previous section, but we will consider the more general penalization term

$$J(g) = \lambda r(||Pg_1||^2, \ldots, ||Pg_d||^2)$$

where $\lambda > 0$, P is the orthogonal projection on $\mathrm{span}\{\varphi_1, \ldots, \varphi_l\}^\perp$ and $r : [0,\infty)^d \to [0,\infty)$ is coordinate-wise increasing. Theorem 3.1 easily generalizes with the following modification. If $\varphi(x) = x + d$ then with

$$h_{i,j}(r) = \int_0^{\tau_i -} R^1(\tau_i - u, r)\mathrm{d}Z_{j,u}$$

for $i = 1, \ldots, N_t$ and $j = 1, \ldots, d$ a minimizer of the penalized minus-log-likelihood functions has j'th coordinate in the space spanned by $\varphi_1, \ldots, \varphi_l$ together with $h_{1,j}, \ldots, h_{N_t,j}$ and f given by

$$f(r) = \int_0^t Y_s \sum_{j=1}^d \int_0^{s-} R^1(s-u,r)\mathrm{d}Z_{j,u}\mathrm{d}s = \sum_{j=1}^d \int_0^t \int_u^t Y_s R^1(s-u,r)\mathrm{d}s\mathrm{d}Z_{j,u}.$$

Theorem 3.5 also generalizes similarly and if r is smooth, for instance if $r(x_1, \ldots, x_d) = \sum_{j=1}^d x_j$, Algorithm 3.6 generalizes as well.

In the alternative, we can choose $r(x_1, \ldots, x_d) = \sum_{j=1}^d \sqrt{x_j}$ leading to the penalty term

$$J(g) = \lambda \sum_j^d ||Pg||,$$

which gives an infinite dimensional version of grouped lasso. Since r is not differentiable, Algorithm 3.6 does not work directly. However, a cyclical descent algorithm may be suggested, where we cyclically decide if the coordinate function g_j should be equal to 0 or should be updated to decrement the objective function. The idea is then to initialize the algorithm with a large λ and all g_j-functions equal to 0, and then in an outer loop decrease λ in small steps and for each choice of λ provide a warm start for the descent algorithm by using the previously estimated g. This strategy has been investigated thoroughly in Friedman et al. (2010) for the ordinary lasso and its generalizations showing very promising performance results.

5 Discussion

The problem that initially motivated the present work was the estimation of the linear filter functions entering in the specification of a non-linear Hawkes model with an intensity specified as

$$\varphi\left(\sum_{j=1}^d \int_0^{s-} g_j(s-u)N_j(\mathrm{d}u)\right)$$

where N_j for $j = 1, \ldots, d$ are counting processes, Brémaud & Massoulié (1996). We have provided structural and algorithmic results for the penalized maximum-likelihood estimator of g_j in a Sobolev space and we have showed that these results can be established in a generality where the stochastic integrals are with respect to any semi-martingale. The representations of basis functions and the gradient are useful for specific examples such as counting processes, but of little analytic value for general semi-martingales. In practice we can only expect to observe a general semi-martingale discretely and numerical approximations to the integral representations and thus the minus-log-likelihood function must be

used. If the semi-martingale is coarsely observed it is unknown how reliable the resulting approximation of the penalized maximum-likelihood estimator is.

For practical applications the R-package `ppstat` contains an implementation of finite-dimensional glppm's with a formula based model specification of additive models. Currently the implementation only supports a quadratic penalization term, but work is ongoing to support grouped lasso penalization as described above. The package is available from `http://www.math.ku.dk/~richard/ppstat/`.

Another point worth mentioning is the similarity between Algorithm 3.6 and the *functional gradient descent* algorithm from the boosting literature, Bühlmann & Hothorn (2007). As in the boosting algorithm the functional estimate is iteratively updated by an additive component, and in one incarnation of Algorithm 3.6 this component is a scalar multiple of the gradient. The main difference is that we propose to compute the gradient in the functional space, which utilizes the inner product in that space, whereas the functional gradient descent algorithm computes the gradient in an ordinary euclidean space and subsequently computes an approximating functional component by a *base procedure*. Details are found in Bühlmann & Hothorn (2007).

6 Acknowledgments

The author was supported by the Danish Natural Science Research Council, grant 09-072331, "Point process modelling and statistical inference".

A Proofs

The Sobolev space $W^{m,2}([0,t])$ has already been equipped with one inner product denoted $<,>$ and the corresponding norm $||\cdot||$. An alternative useful inner product on $W^{m,2}([0,t])$ is

$$< f, g >_m = \sum_{k=0}^{m} \int_0^t D^k f(s) D^k g(s) \mathrm{d}s$$

and the corresponding norm is given by

$$||f||_{m,2}^2 = < f, f >_m = \sum_{k=0}^{m} \int_0^t D^k f(s)^2 \mathrm{d}s.$$

It is straight forward to show that $||\cdot||$ and $||\cdot||_{m,2}$ are equivalent norms. We will use whichever norm is most convenient in the proofs below. Note that the embedding $W^{m,2}([0,t]) \hookrightarrow W^{k,2}([0,t])$ for $m < k$ is continuous, which is straight forward using the norms $||\cdot||_{m,2}$ and $||\cdot||_{k,2}$. The continuity of the embedding holds even when $k = 0$ where $W^{0,2}([0,t]) = L_2([0,t])$, which is not a reproducing kernel Hilbert space.

We note that the characterizing property of a reproducing kernel Hilbert space is that the function evaluations are continuous linear functionals. If δ_s denotes the evaluation in s, that is, $\delta_s f = f(s)$, then $R(s,\cdot)$ as a function in $W^{m,2}([0,t])$ represents δ_s by

$$f(s) =< f, R(s,\cdot) > .$$

By Cauchy-Schwarz' inequality $||\delta_s|| = R(s,s)$ and since R is a continuous function of both variables $R(s,s)$ is bounded for s in a compact set.

We have already argued that the stochastic integration of deterministic functions from $W^{m,2}([0,t])$ can be regarded as a pathwise, linear functional defined on $W^{m,2}([0,t])$ for $m \geq 1$. The next lemma anc following corollary states the this functional is continuous.

Lemma A.1. *Let $0 \leq s \leq t$. Then the linear functional $X_s : W^{1,2}([0,t]) \to \mathbb{R}$ defined by*

$$X_s h = \int_0^s h(u) \mathrm{d}Z_u$$

is continuous. More precisely, we have the bound

$$||X_s|| \leq |Z_s|(1+s) + |Z_0| + \left(\int_0^s Z_{u-}^2 \mathrm{d}s \right)^{1/2} < \infty.$$

Proof: Note that for $h \in W^{1,2}([0,t])$ we have

$$||h||^2 = |h(0)|^2 + ||h'||_2^2$$

and in particular

$$||h' 1_{[0,s]}||_2 \leq ||h'||_2 \leq ||h||.$$

Using (6) and Cauchy-Schwarz' inequality

$$\begin{aligned}
|X_s h| &\leq |h(s)Z_s| + |h(0)Z_0| + \int_0^s |Z_{u-} h'(u)| \mathrm{d}u \\
&\leq |Z_s|\,|h(s)| + |Z_0|\,|h(0)| + \left(\int_0^s Z_{u-}^2 \mathrm{d}s \right)^{1/2} ||h' 1_{[0,s]}||_2 \\
&\leq \left(|Z_s|\,||\delta_s|| + |Z_0|\,||\delta_0|| + \left(\int_0^s Z_{u-}^2 \mathrm{d}s \right)^{1/2} \right) ||h|| \\
&\leq \left(|Z_s|(1+s) + |Z_0| + \left(\int_0^s Z_{u-}^2 \mathrm{d}s \right)^{1/2} \right) ||h||,
\end{aligned}$$

which shows the desired bound. Here we have used that for $m = 1$ we have $R(s,s) = 1 + s$ and that Z is càdlàg, hence bounded and hence in $L_2([0,s])$ for any s. □

As the embedding $W^{m,2}([0,t]) \hookrightarrow W^{1,2}([0,t])$ is continuous we get the following immediate corollary.

Corollary A.2. *The linear functional* $X_s : W^{m,2}([0,t]) \to \mathbb{R}$ *defined by*

$$X_s h = \int_0^s h(u)\mathrm{d}Z_u$$

is continuous.

Corollary A.2 shows that $X_s \in W^{m,2}([0,t])^*$ for $s \geq 0$. We now show that it is also norm-càdlàg.

Lemma A.3. *The process* $(X_s)_{0\leq s\leq t}$ *is a norm-càdlàg stochastic process.*

Proof: For $\varepsilon > 0$

$$|X_{s+\varepsilon}h - X_s h| = \left| \int_{s+}^{s+\varepsilon} h(u)\mathrm{d}Z_u \right|.$$

Again by integration by parts

$$\int_{s+}^{s+\varepsilon} h(u)\mathrm{d}Z_u = h(s+\varepsilon)Z_{s+\varepsilon} - h(s)Z_s - \int_{s+}^{s+\varepsilon} Z_{u-}h'(u)\mathrm{d}u$$

and arguments similar to those in the proof of Lemma A.1 gives that

$$|X_{s+\varepsilon}h - X_s h| \leq \left(||Z_{s+\varepsilon}\delta_{s+\varepsilon} - Z_s\delta_s|| + \left(\int_{s+}^{s+\varepsilon} Z_{u-}^2 \mathrm{d}u \right)^{1/2} \right) ||h||.$$

This shows that

$$||X_{s+\varepsilon} - X_s|| \leq ||Z_{s+\varepsilon}\delta_{s+\varepsilon} - Z_s\delta_s|| + \left(\int_{s+}^{s+\varepsilon} Z_{u-}^2 \mathrm{d}u \right)^{1/2}$$

and letting $\varepsilon \to 0+$ the right hand side tends to 0 by an application of dominated convergence and because $Z_{s+\varepsilon} \to Z_s$ and $\delta_{s+\varepsilon} \to \delta_s$. This proves that the process is continuous from the right in norm.

Defining X_{s-} by

$$X_{s-}h = \int_0^{s-} h(u)\mathrm{d}Z_u$$

a similar argument shows that $||X_{s-\varepsilon} - X_{s-}|| \to 0$ for $\varepsilon \to 0+$, which shows that the process has limits from the left in norm. □

To give the proof of Theorem 3.1 we will use the following general lemma.

Lemma A.4. *If* $(H_t)_{t\geq 0}$ *is a norm-càdlàg stochastic process with values in* V^* *then for* $t \geq 0$ *the integral* $\int_0^t Y_s H_{s-}\mathrm{d}s$ *defined by*

$$\beta \mapsto \int_0^t Y_s H_{s-}\beta \mathrm{d}s \tag{11}$$

is in V^* with

$$\left|\left|\int_0^t Y_s H_{s-} \mathrm{d}s\right|\right| \leq \int_0^t |Y_s| \, ||H_{s-}|| \mathrm{d}s$$

Proof: Clearly (11) defines for a fixed $t \geq 0$ a linear functional on V. Moreover, since $|H_{s-}\beta| \leq ||H_{s-}|| \, ||\beta||$

$$\begin{aligned} \left|\int_0^t Y_s H_{s-} \beta \mathrm{d}s\right| &\leq \int_0^t |Y_s H_{s-} \beta| \, \mathrm{d}s \\ &\leq \int_0^t |Y_s| \, ||H_{s-}|| \mathrm{d}s \; ||\beta||. \end{aligned}$$

Now as $(H_t)_{t\geq 0}$ is assumed norm-càdlàg it follows by continuity of the norm that $||H_{s-}||$ for $s \in [0,t]$ is bounded, and the integral is finite and a bound on the norm of the functional. □

Proof: (Theorem 3.1) When $\varphi(x) = x + d$ we have that

$$\begin{aligned} l_t(g) &= \int_0^t Y_s \int_0^{s-} g(s-u) \mathrm{d}Z_u + dY_s \mathrm{d}s - \int_0^t \log\left(Y_s \int_0^{s-} g(s-u)\mathrm{d}Z_u + dY_s\right) N(\mathrm{d}s) \\ &= \int_0^t Y_s \int_0^{s-} g(s-u)\mathrm{d}Z_u \mathrm{d}s + d\int_0^t Y_s \mathrm{d}s - \sum_{i=1}^{N_t} \log\left(Y_{\tau_i} \int_0^{\tau_i -} g(\tau_i - u) + dY_{\tau_i}\right) \mathrm{d}Z_u. \end{aligned}$$

It follows from Corollary A.2 that

$$g \mapsto \int_0^{\tau_i -} g(\tau_i - u)\mathrm{d}Z_u$$

are continuous, linear functionals on $W^{m,2}([0,t])$. The i'th of these continuous linear functionals is represented by $\eta_i \in W^{m,2}([0,t])$ given as

$$\eta_i(s) = \int_0^{\tau_i -} R(\tau_i - u, s)\mathrm{d}Z_u.$$

such that

$$< \eta_i, g >= \int_0^{\tau_i -} g(\tau_i - u)\mathrm{d}Z_u.$$

Hence $h_i = P\eta_i$.

Combining Lemma A.3 and A.4 we conclude that

$$g \mapsto \int_0^t Y_s \int_0^{s-} g(s-u)\mathrm{d}Z_u \mathrm{d}s$$

is a continuous linear functional and η is the representer given by

$$\eta(r) = \int_0^t Y_s \int_0^{s-} R(s-u, r)\mathrm{d}Z_u \mathrm{d}s$$

then $f = P\eta$.

Thus $l_t(g)$ is a function of a finite number of continuous, linear functionals on $W^{m,2}([0,t])$,

$$l_t(g) =< \eta, g > - \sum_{i=1}^{N_t} \log(Y_{\tau_i} < \eta_i, g > +d) + K$$

where $K = d\int_0^t Y_s \mathrm{d}s$ does not depend upon g. Assume that $g \in \Theta(D) \subseteq W^{m,2}([0,t])$ and write $g = g_0 + \rho$ where $\rho \in \mathrm{span}\{\varphi_1, \ldots, \varphi_m, h_1, \ldots, h_{N_t}, f\}^{\perp}$, then $\rho \perp \eta_i$ for $i = 1, \ldots, N_t$, $\rho \perp \eta$, $P\rho = \rho$ and

$$\begin{aligned} l_t(g) + \lambda||Pg||^2 &= < \eta, g > - \sum_{i=1}^{N_t} \log(Y_{\tau_i} < \eta_i, g > +d) + K + \lambda||Pg||^2 \\ &= < \eta, g_0 > - \sum_{i=1}^{N_t} \log(Y_{\tau_i} < \eta_i, g_0 > +d) + K + \lambda||Pg_0||^2 + \lambda||\rho||^2 \\ &\geq l_t(g_0) + \lambda||Pg_0||^2 \end{aligned}$$

with equality if and only if $\rho = 0$. Thus a minimizer of $l_t(g) + \lambda||Pg||^2$ over $\Theta(D)$ must be in $\mathrm{span}\{\varphi_1, \ldots, \varphi_m, h_1, \ldots, h_{N_t}, f\}$. □

We have used the Fubini theorem below to give an alternative representation of the basis function f from Theorem 3.1. The result is a consequence of Theorem 45 in Protter (2005). With the pathwise definition of stochastic integrals, as given by (6), that we have used throughout, we can give an elementary proof.

Lemma A.5. *With $(Z_s)_{0\leq s\leq t}$ a semi-martingale and $(Y_s)_{0\leq s\leq t}$ a predictable, càdlàg process then*

$$\int_0^t Y_s \int_0^{s-} g(s-u)\mathrm{d}Z_u \mathrm{d}s = \int_0^t \int_u^t Y_s g(s-u)\mathrm{d}s\mathrm{d}Z_u.$$

Proof: Using (6) and Fubini

$$\begin{aligned} \int_0^t Y_s \int_0^{s-} g(s-u)\mathrm{d}Z_u\mathrm{d}s &= g(0)\int_0^t Z_{s-}Y_s\mathrm{d}s - Z_0\int_0^t g(s)Y_s\mathrm{d}s + \int_0^t Y_s\int_0^{s-} Z_{u-}g'(s-u)\mathrm{d}u\mathrm{d}s \\ &= g(0)\int_0^t Z_{s-}Y_s\mathrm{d}s - Z_0\int_0^t g(s)Y_s\mathrm{d}s + \int_0^t Z_{u-}\int_u^t Y_s g'(s-u)\mathrm{d}s\mathrm{d}u. \end{aligned}$$

To use (6) for the right hand side above we first need to verify that the integrand is sufficiently regular. Defining

$$G(u) = \int_u^t Y_s g(s-u)\mathrm{d}s$$

for $g \in W^{1,2}([0,t])$ then G is weakly differentiable with derivative

$$G'(u) = -\int_u^t Y_s g'(s-u)\mathrm{d}s - Y_u g(0),$$

which is verified simply by checking that $G(u) = -\int_u^t G'(v)\mathrm{d}v$. Using this, we get for the right hand side above that

$$\begin{aligned}\int_0^t \underbrace{\int_u^t Y_s g(s-u)\mathrm{d}s}_{G(u)}\mathrm{d}Z_u &= G(t)Z_t - G(0)Z_0 - \int_0^t Z_{u-}G'(u)\mathrm{d}u \\ &= -G(0)Z_0 + \int_0^t Z_{u-}\left[\int_u^t Y_s g'(s-u)\mathrm{d}s + Y_u g(0)\right]\mathrm{d}u \\ &= g(0)\int_0^t Z_{s-}Y_s\mathrm{d}s - Z_0\int_0^t g(s)Y_s\mathrm{d}s + \int_0^t Z_{u-}\int_u^t Y_s g'(s-u)\mathrm{d}s\mathrm{d}u.\end{aligned}$$

□

Proof: (Theorem 3.5) The Gâteaux derivative of l_t in the direction of $h \in W^{m,2}([0,t])$ for $g \in \Theta(D)^\circ$ is by Proposition 2.3

$$\begin{aligned}Dl_t(g)h &= \int_0^t Y_s\varphi'\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right)\int_0^{s-} h(s-u)\mathrm{d}Z_u\mathrm{d}s \\ &\quad - \int_0^t \frac{\varphi'\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right)}{\varphi\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right)}\int_0^{s-} h(s-u)\mathrm{d}Z_u N(\mathrm{d}s).\end{aligned}$$

Now just as in the proof of Theorem 3.1 – replacing Y_s by $Y_s\varphi'\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right)$ – it follows that the first term above is a continuous, linear functional on $W^{m,2}([0,t])$ with representer f_g. Moreover, with η_i as defined in Theorem 3.5 the second term above is seen to be a continuous, linear functional on $W^{m,2}([0,t])$ with representer

$$\zeta_g = \sum_{i=1}^{N_t}\frac{\varphi'\left(\int_0^{\tau_i-} g(\tau_i-u)\mathrm{d}Z_u\right)}{\varphi\left(\int_0^{\tau_i-} g(\tau_i-u)\mathrm{d}Z_u\right)}\eta_i.$$

In conclusion, the gradient of l_t in g is $\nabla l_t(g) = f_g - \zeta_g$. □

Lemma A.6. *If $D = \mathbb{R}$ and φ is strictly positive, twice continuously differentiable then the gradient $\nabla\Lambda : W^{m,2}([0,t]) \to W^{m,2}([0,t])$ is Lipschitz continuous on any bounded set.*

Proof: Let $B(0,L)$ denote the ball with radius L in $W^{m,2}([0,t])$. Corollary A.2 shows that X_s is a continuous functional and $g \mapsto X_{s-}g = \int_0^{s-} g(s-u)\mathrm{d}Z_u$ is likewise continuous. That is, $|X_{s-}g| \leq ||X_{s-}||||g||$, and $s \mapsto ||X_{s-}||$ is, moreover, bounded on $[0,t]$. This means

that there is an $M > 0$ such that $X_{s-}g \in [-M, M]$ for all $g \in B(0, L)$ and $s \in [0, t]$. Since φ is twice continuously differentiable we have that φ' is Lipschitz continuous on $[-M, M]$ with Lipschitz constant K, say. With f_g for $g \in W^{m,2}([0, t])$ as in Theorem 3.7 we find that for $g, g' \in W^{m,2}([0, t])$

$$f_g - f_{g'} = \int_0^t Y_s\big(\varphi'(X_{s-}g) - \varphi'(X_{s-}g')\big) \int_0^{s-} R^1(s-u, \cdot)\mathrm{d}Z_u \mathrm{d}s$$

and as above, by the isometric isomorphism that identifies $W^{m,2}[0, t]$ with its dual, we get by Lemma A.4 that if also $g, g' \in B(0, L)$ then

$$\begin{aligned} ||f_g - f_{g'}|| &\leq \int_0^t |Y_s|\, |\varphi'(X_{s-}g) - \varphi'(X_{s-}g')|\, ||X_{s-}P|| \mathrm{d}s \\ &\leq K \int_0^t |Y_s|\, ||X_{s-}||^2\, ||g - g'|| \mathrm{d}s \\ &\leq \underbrace{K \int_0^t |Y_s|\, ||X_{s-}||^2 \mathrm{d}s}_{C_1}\, ||g - g'||. \end{aligned}$$

Since φ is strictly positive – and twice continuously differentiable – $x \mapsto \varphi'(x)/\varphi(x)$ is Lipschitz continuous on $[-M, M]$ with Lipschitz constant K', say. Then for $g, g' \in B(0, L)$

$$\begin{aligned} \left|\left|\sum_{i=1}^{N_t} \frac{\varphi'\left(X_{\tau_i-}g\right)}{\varphi\left(X_{\tau_i-}g\right)}\eta_i - \sum_{i=1}^{N_t} \frac{\varphi'\left(X_{\tau_i-}g'\right)}{\varphi\left(X_{\tau_i-}g'\right)}\eta_i\right|\right| &\leq \sum_{i=1}^{N_t} \left|\frac{\varphi'\left(X_{\tau_i-}g\right)}{\varphi\left(X_{\tau_i-}g\right)} - \frac{\varphi'\left(X_{\tau_i-}g'\right)}{\varphi\left(X_{\tau_i-}g'\right)}\right| ||\eta_i|| \\ &\leq K' \sum_{i=1}^{N_t} ||X_{\tau_i-}||\, ||g - g'||\, ||\eta_i|| \\ &\leq \underbrace{K' \sum_{i=1}^{N_t} ||X_{\tau_i-}||\, ||\eta_i||}_{C_2}\, ||g - g'||. \end{aligned}$$

By Proposition 3.5 we have showed that the gradient ∇l_t is Lipschitz continuous on the bounded set $B(0, L)$ with Lipschitz constant $C = C_1 + C_2$. Since $\nabla\Lambda = \nabla l_t + 2\lambda P$ and $2\lambda P$ is linear this proves that $\nabla\Lambda$ is Lipschitz continuous on bounded sets. □

Proof: (Theorem 3.7) We prove first by induction that it is possible to iteratively choose $\hat{g}_h$ as prescribed in Algorithm 3.6. The induction start is given by assumption.

Assume that $\hat{g}_h$ is chosen as in Algorithm 3.6. Since $\Lambda : W^{m,2}([0, t]) \to \mathbb{R}$ is continuous and

$$\mathcal{S}_h := \{g \in W^{m,2}([0, t]) \mid \Lambda(g) \leq \Lambda(\hat{g}_h)\} \subseteq \mathcal{S}$$

is bounded by assumption we find that Λ is bounded below along the ray $\hat{g}_h - \alpha\nabla\Lambda(\hat{g}_h)$ for $\alpha > 0$. If $\nabla\Lambda(\hat{g}_h) \neq 0$ we can proceed exactly as in the proof of Lemma 3.1 in Nocedal & Wright (2006), and there exists $\alpha > 0$ such that

$$\tilde{g}_{h+1} = \hat{g}_h - \alpha\nabla\Lambda(\hat{g}_h) \in \mathcal{S}_h$$

fulfills the two Wolfe conditions;

$$\begin{aligned} \Lambda(\tilde{g}_{h+1}) &\leq \Lambda(\hat{g}_h) - c_1\alpha||\nabla\Lambda(\hat{g}_h)||^2 \\ < \nabla\Lambda(\tilde{g}_{h+1}), \nabla\Lambda(\hat{g}_h) > &\leq c_2||\nabla\Lambda(\hat{g}_h)||^2. \end{aligned}$$

Since $\hat{g}_h \in \text{span}\{h_1, \ldots, h_{N_t}, f_0, \ldots, f_{h-1}\}$ and $\nabla\Lambda(\hat{g}_h) \in \text{span}\{h_1, \ldots, h_{N_t}, f_0, \ldots, f_h\}$ and since $\tilde{g}_{h+1} - \hat{g}_h = -\alpha\nabla\Lambda(\hat{g}_h)$ we find that

$$\tilde{g}_{h+1} \in W(\hat{g}_h) \cap \text{span}\{h_1, \ldots, h_{N_t}, f_0, \ldots, f_h\}$$

and the set on the right hand side is in particular non-empty. This proves that it is possible to iteratively choose $\hat{g}_h$ as in Algorithm 3.6.

For the entire sequence $(\hat{g}_h)_{h\geq 0}$ we get from the second Wolfe condition together with the Cauchy-Schwarz inequality and Lipschitz continuity of $\nabla\Lambda$ on $\mathcal{S}$ that

$$\begin{aligned} (c_2 - 1) < \nabla\Lambda(\hat{g}_h), \hat{g}_{h+1} - \hat{g}_h > &\leq < \nabla\Lambda(\hat{g}_{h+1}) - \nabla\Lambda(\hat{g}_h), \hat{g}_{h+1} - \hat{g}_h > \\ &\leq C||\hat{g}_{h+1} - \hat{g}_h||^2, \end{aligned}$$

which implies that

$$||\hat{g}_{h+1} - \hat{g}_h|| \geq \frac{(c_2 - 1)}{C} \frac{< \nabla\Lambda(\hat{g}_h), \hat{g}_{h+1} - \hat{g}_h >}{||\hat{g}_{h+1} - \hat{g}_h||}.$$

Combining the angle condition with the first Wolfe condition gives that

$$\begin{aligned} \Lambda(\hat{g}_{h+1}) &\leq \Lambda(\hat{g}_h) + c_1||\hat{g}_{h+1} - \hat{g}_h|| \frac{< \nabla\Lambda(\hat{g}_h), \hat{g}_{h+1} - \hat{g}_h >}{||\hat{g}_{h+1} - \hat{g}_h||} \\ &\leq \Lambda(\hat{g}_h) - \frac{c_1(1 - c_2)}{C} \frac{< \nabla\Lambda(\hat{g}_h), \hat{g}_{h+1} - \hat{g}_h >^2}{||\nabla\Lambda(\hat{g}_h)||^2||\hat{g}_{h+1} - \hat{g}_h||^2} ||\nabla\Lambda(\hat{g}_h)||^2 \\ &\leq \Lambda(\hat{g}_h) - \frac{c_1(1 - c_2)\delta}{C} ||\nabla\Lambda(\hat{g}_h)||^2. \end{aligned}$$

By induction

$$\Lambda(\hat{g}_{h+1}) \leq \Lambda(\hat{g}_0) - \frac{c_1(1 - c_2)\delta}{C} \sum_{k=0}^{h} ||\nabla\Lambda(\hat{g}_k)||^2.$$

To finish the proof we need to show that Λ is bounded below on $\mathcal{S}$, because then the inequality above implies that

$$||\nabla\Lambda(\hat{g}_h)|| \to 0$$

for $h \to \infty$. To show that Λ is bounded below we observe that

$$\begin{aligned}\Lambda(g) &\geq -\int_0^t \log(Y_s\varphi\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right))N(\mathrm{d}s) \\ &= -\sum_{i=1}^{N_t} \log(Y_{\tau_i}\varphi\left(\int_0^{s-} g(\tau_i-u)\mathrm{d}Z_u\right)) \\ &= -\sum_{i=1}^{N_t} \log(Y_{\tau_i}\varphi(<\eta_i, g>)).\end{aligned}$$

Since this lower bound as a function of g is weakly continuous and since a bounded set is weakly compact by reflexivity of a Hilbert space and Banach-Alaoglu's Theorem we have proved that Λ is bounded below on the bounded set $\mathcal{S}$. □

For the proof of Corollary 3.8 we need the following lemma.

Lemma A.7. *If φ is strictly positive and continuously differentiable the map $g \mapsto \nabla\Lambda(g)$ is weak-weak continuous.*

Proof: By defition of the weak topology we need to show that

$$g \mapsto <\nabla\Lambda(g), h> = <\nabla l_t(g), h> + 2\lambda < Pg, h>$$

is weakly continuous for all $h \in W^{m,2}([0,t])$. Clearly $g \mapsto < Pg, h> = < g, Ph>$ is weakly continuous so we can restrict our attention to $g \mapsto <\nabla l_t(g), h>$. We will use Theorem 3.5, and to do so we observe that

$$g \mapsto \int_0^{s-} g(s-u)\mathrm{d}Z_u$$

for fixed s is weakly continuous by the definition of the weak topology and the fact that we have already shown the map above to be a continuous linear functional. We conclude directly from this that

$$g \mapsto \sum_{i=1}^{N_t} \frac{\varphi'\left(\int_0^{\tau_i-} g(\tau_i-u)\mathrm{d}Z_u\right)}{\varphi\left(\int_0^{\tau_i-} g(\tau_i-u)\mathrm{d}Z_u\right)} < \eta_i, h>$$

is weakly continuous as φ is assumed strictly positive and continuously differentiable. We finish the proof by showing that $g \mapsto < f_g, h>$ is weakly continuous with f_g as in Theorem 3.5. Let $g_n \xrightarrow{w} g$ for $n \to \infty$ in which case

$$\int_0^{s-} g_n(s-u)\mathrm{d}Z_u \to \int_0^{s-} g(s-u)\mathrm{d}Z_u$$

for all $s \in [0,t]$. Since the stochastic integral as a function of s is bounded on $[0,t]$ and φ' is continuous, the pointwise convergence of

$$Y_s\varphi'\left(\int_0^{s-} g_n(s-u)\mathrm{d}Z_u\right)X_sh \to Y_s\varphi'\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right)X_sh$$

for $s \in [0,t]$ is dominated by a constant, which is integrable over $[0,t]$. Hence

$$\begin{aligned} < f_{g_n}, h > &= \int_0^t Y_s\varphi'\left(\int_0^{s-} g_n(s-u)\mathrm{d}Z_u\right)X_sh\mathrm{d}s \\ &\to \int_0^t Y_s\varphi'\left(\int_0^{s-} g(s-u)\mathrm{d}Z_u\right)X_sh\mathrm{d}s = < f_g, h > . \end{aligned}$$

□

Proof: (Corollary 3.8) By assumption, $\hat{g}$ is the unique solution to $\nabla\Lambda(g) = 0$. The bounded set $\mathcal{S}$ is weakly compact as argued above and the weak topology is, moreover, metrizable on $\mathcal{S}$ since $W^{m,2}([0,t])$ is separable. Therefore any subsequence of $(\hat{g}_h)_{h\geq 0}$ has a subsequence that converges weakly in $\mathcal{S}$, necessarily towards a limit with vanishing gradient by Lemma A.7. Uniqueness of $\hat{g}$ implies that $(\hat{g}_h)_{h\geq 0}$ itself is weakly convergent with limit $\hat{g}$. The proof is completed by noting that weak convergence in a reproducing kernel Hilbert space implies pointwise convergence. □

References

Aitkin, M., Francis, B. & Hinde, J. (2005), *Statistical modelling in GLIM 4*, Vol. 32 of *Oxford Statistical Science Series*, second edn, Oxford University Press, Oxford.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993), *Statistical models based on counting processes*, Springer Series in Statistics, Springer-Verlag, New York.

Asmussen, S. (2003), *Applied probability and queues*, Vol. 51 of *Applications of Mathematics*, second edn, Springer-Verlag, New York. Stochastic Modelling and Applied Probability.

Berlinet, A. & Thomas-Agnan, C. (2004), *Reproducing kernel Hilbert spaces in probability and statistics*, Kluwer Academic Publishers, Boston, MA. With a preface by Persi Diaconis.

Brémaud, P. (1981), *Point processes and queues*, Springer-Verlag, New York. Martingale dynamics, Springer Series in Statistics.

Brémaud, P. & Massoulié, L. (1996), 'Stability of nonlinear Hawkes processes', *Ann. Probab.* **24**(3), 1563–1588.

Brillinger, D. R. (1992), 'Nerve cell spike train data analysis: A progression of technique', *Journal of the American Statistical Association* **87**(418), 260–271.
URL: *http://www.jstor.org/stable/2290256*

Bühlmann, P. & Hothorn, T. (2007), 'Boosting algorithms: regularization, prediction and model fitting', *Statist. Sci.* **22**(4), 477–505.
URL: *http://dx.doi.org/10.1214/07-STS242*

Carstensen, L., Sandelin, A., Winther, O. & Hansen, N. R. (2010), Multivariate hawkes process models of the occurrence of regulatory elements. Working paper.

Cox, D. D. & O'Sullivan, F. (1990), 'Asymptotic analysis of penalized likelihood and related estimators', *Ann. Statist.* **18**(4), 1676–1695.
URL: *http://dx.doi.org/10.1214/aos/1176347872*

Fleming, T. R. & Harrington, D. P. (1991), *Counting processes and survival analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York.

Friedman, J. H., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.
URL: *http://www.jstatsoft.org/v33/i01*

Green, P. J. & Silverman, B. W. (1994), *Nonparametric regression and generalized linear models*, Vol. 58 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London. A roughness penalty approach.

Gusto, G. & Schbath, S. (2005), 'FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model', *Stat. Appl. Genet. Mol. Biol.* **4**, Art. 24, 28 pp. (electronic).

Hautsch, N. (2004), *Modelling irregularly spaced financial data*, Vol. 539 of *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, Berlin. Theory and practice of dynamic duration models, Dissertation, University of Konstanz, Konstanz, 2003.

Jacobsen, M. (2006), *Point process theory and applications*, Probability and its Applications, Birkhäuser Boston Inc., Boston, MA. Marked point and piecewise deterministic processes.

Jacod, J. (1975), 'Multivariate point processes: predictable projection, Radon-Nikodým derivatives, representation of martingales', *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **31**, 235–253.

Jacod, J. & Shiryaev, A. N. (2003), *Limit theorems for stochastic processes*, Vol. 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, second edn, Springer-Verlag, Berlin.

Karr, A. F. (1991), *Point processes and their statistical inference*, Vol. 7 of *Probability: Pure and Applied*, second edn, Marcel Dekker Inc., New York.

Kiefer, J. & Wolfowitz, J. (1956), 'Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters', *The Annals of Mathematical Statistics* **27**(4), 887–906.
URL: *http://www.jstor.org/stable/2237188*

Maston, G. A., Evans, S. K. & Green, M. R. (2006), 'Transcriptional regulatory elements in the human genome', *ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS* **7**, 29–59.

Mikosch, T. (2004), *Non-life insurance mathematics*, Universitext, Springer-Verlag, Berlin. An introduction with stochastic processes.

Nocedal, J. & Wright, S. J. (2006), *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, second edn, Springer, New York.

Ogata, Y. & Katsura, K. (1986), 'Point-process models with linearly parametrized intensity for application to earthquake data', *J. Appl. Probab.* **Special Vol. 23A**, 291–310. Essays in time series and allied processes.

Ogata, Y., Katsura, K. & Tanemura, M. (2003), 'Modelling heterogeneous space-time occurrences of earthquakes and its residual analysis', *J. Roy. Statist. Soc. Ser. C* **52**(4), 499–509.
URL: *http://dx.doi.org/10.1111/1467-9876.00420*

Paninski, L. (2004), 'Maximum likelihood estimation of cascade point-process neural encoding models', *Network: Computation in Neural Systems* **15**(4), 243–262.

Pedersen, G. K. (1989), *Analysis now*, Vol. 118 of *Graduate Texts in Mathematics*, Springer-Verlag, New York.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J. & Simoncelli, E. P. (2008), 'Spatio-temporal correlations and visual signalling in a complete neuronal population', *Nature* **454**, 995–999.

Protter, P. E. (2005), *Stochastic integration and differential equations*, Vol. 21 of *Stochastic Modelling and Applied Probability*, Springer-Verlag, Berlin. Second edition. Version 2.1, Corrected third printing.

Toyoizumi, T., Rad, K. R. & Paninski, L. (2009), 'Mean-field approximations for coupled populations of generalized linear model spiking neurons with Markov refractoriness', *Neural Comput.* **21**(5), 1203–1243.
URL: *http://dx.doi.org/10.1162/neco.2008.04-08-757*

Wahba, G. (1990), *Spline models for observational data*, Vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Whitehead, J. (1980), 'Fitting cox's regression model to survival data using glim', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **29**(3), 268–275.
URL: *http://www.jstor.org/stable/2346901*