Multivariate point process models with applications to genomic organization

Niels Richard Hansen
Department of Mathematical Sciences

Motivation and aims

Two motivating applications:

- The modeling of physical binding positions of transcription factors on the genome.
- The modeling of the spike times for single neurons or neuronal nets.

General aims:

- To build and implement a flexible (non-parametric), intensity based framework of multivariate point process models – aka discretely marked point processes.
- To develop an organizational model of active transcription regulatory elements along genomes – joint with Lisbeth Carstensen, Albin Sandelin, Ole Winther.

Point process modeling via intensities

We consider a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ and a parametrized family $(\lambda_t(\theta))_{t \geq 0}$ of positive, predictable processes for $\theta \in \Theta$.
The minus-log-likelihood is

$$l_t(\theta) = \int_0^t \lambda_s(\theta) \mathrm{d}s - \int_0^t \log \lambda_s(\theta) N(\mathrm{d}s)$$

We will study penalized maximum-likelihood estimation of the parameter θ.

Simple example form neuron modeling

A neuron cell emits electro-chemical signals known as spikes. The cells membrane potential is influenced by signals from other neurons and determines the signaling of the cell.

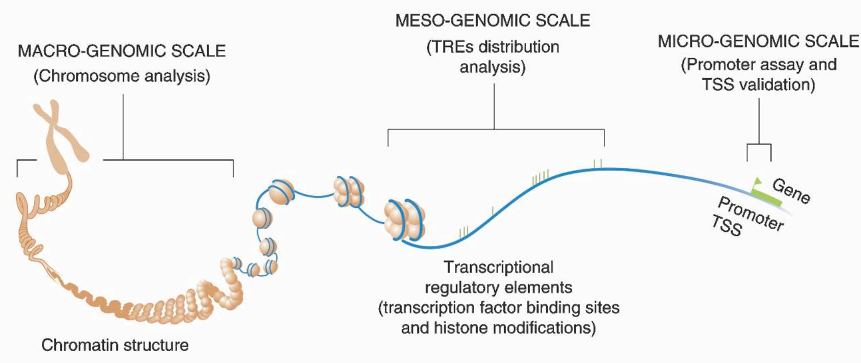
Take $(X_t)_{t \geq 0}$ to be the membrane potential and $\lambda_t(a, b) = \varphi(aX_t + b)$ for a fixed φ to generalize commonly used threshold models:

- $\varphi(x) = c1_{\{x>1\}}$. Threshold model, a, b determines the threshold, c how rapid the neuron fires.
- $\varphi(x) = x_+$, "convex version", b determines the threshold and a how rapid the neuron fires.

More complicated examples are found in the recent developments of multivariate modeling of spikes in neuronal nets[1].

[1] Pillow et al. *Spatio-temporal correlations and visual signalling in a complete neuronal population*. Nature, 454. 2008

Genomic scales


MESO-GENOMIC SCALE
(TREs distribution analysis)
MACRO-GENOMIC SCALE
(Chromosome analysis)
MICRO-GENOMIC SCALE
(Promoter assay and TSS validation)
Gene
Promoter
TSS
Transcriptional regulatory elements (transcription factor binding sites and histone modifications)
Chromatin structure

We focus on the distribution of point-like transcriptional regulatory elements[2] at the meso-genomic scale.

[2] Figure from Zhang et al., Genome Res. 17. 787-797. 2007

Transcription regulator binding loci

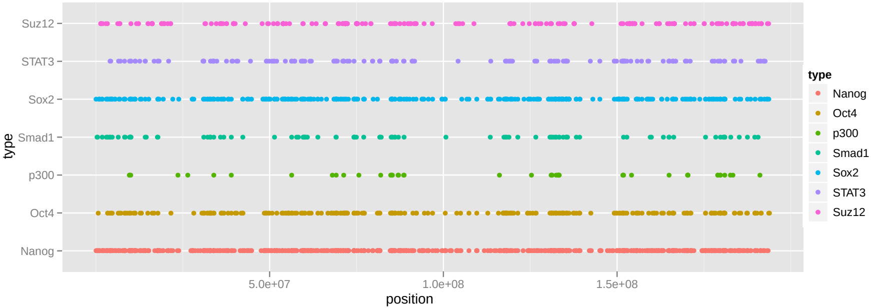
Attempt of a broad definition: Transcription regulators are proteins that modify, interact with or bind to the DNA, chromatin or other transcription regulators to either activate or repress the transcription of DNA.

An active transcription regulatory loci is a loci on the genome where we observe the presence of a transcription regulator.

Fact: Transcription regulators cluster – in promoter regions and intergenic regions. Why? Is there a combined effect? Do they recruit each other ... ?

With regulators as marks and measurements of the active loci as points on the meso-genomic scale we use a multivariate point process model of the organization of active loci.

Embryonic mouse stem cell data (ChIP-seq)



Mouse, chromosome I: 7 of 15 active transcription factor binding loci measured by ChIP-seq for embryonic stem cells.[3]

[3] *Chen et al*, Cell 133, 1106-1117, 2008

A self-exciting model

If $(N_t)_{t\geq 0}$ is the counting process for the occurrences of a TR binding loci along the genome. With $\tau_1, \ldots \tau_{N_t}$ the jumps we consider the model

$$\lambda_t(g) = \varphi\left(\sum_{j:\tau_j<t} g(t-\tau_j)\right) = \varphi\left(\int_0^{t-} g(t-s)N(\mathrm{d}s)\right)$$

For a multivariate counting process, $(N_t^i)_{t\geq 0, i=1,\ldots,K}$, we consider

$$\lambda_t^k(g) = \varphi\left(\sum_{i=1}^{K}\sum_{j:\tau_j^i<t} g^{ik}(t-\tau_j^i)\right),$$

which is the non-linear Hawkes process[4]. With $\varphi(x) = x + d$ we get the *linear* Hawkes process.

[4] Bremaud, P. and Massoulie, L. *Stability of nonlinear Hawkes processes*. Ann. Probab. 24(3), 1996.

Generalized linear point process models

Assume that $(X_t)_{t\geq 0}$ is an adapted, càdlàg process with values in V^* – the dual of the vector space V and define

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

Assume that $\varphi : D \to [0, \infty)$ and assume in addition that $(Y_t)_{t\geq 0}$ 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 $s \in [0, t]$.

The problem with explosion

Given a predictable (candidate) intensity process $(\lambda_t)_{t\geq 0}$ does it define a point process? Yes but the likelihood process

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

may not be a martingale.

$E_P(\mathcal{L}_t) = 1$ if and only if the intensity defines a point process that does not explode in $[0, t]$.

We must restrict our attention to combinations of φ and processes $(X_t)_{t\geq 0}$ such that the likelihood process is a martingale. Otherwise we don't have a dominated statistical model and we don't have a likelihood function.

Banach space parameters

We can take $V = \mathbb{R}^d$ and

$$X_t = (X_{t,1}, \ldots, X_{t,d})$$

a d-dimensional (row) vector with $(X_{t,i})_{t \geq 0}$ an adapted, real valued, càdlàg process.

The definition of a glppm applies whenever V is a separable Banach space, the process $(X_t)_{t \geq 0}$ takes values in the dual space V^* of continuous linear functionals, is adapted and norm-càdlàg.

Under these circumstances $(X_t \beta)_{t \geq 0}$ is adapted and càdlàg, $(X_{t-} \beta)_{t \geq 0}$ is predictable and $(||X_t||)_{t \geq 0}$ and $(||X_{t-}||)_{t \geq 0}$ are bounded on bounded intervals.

Stochastic integrals as linear functionals

If $g : [0, \infty) \to \mathbb{R}$ is a measurable, locally bounded function and $(Z_t)_{t \geq 0}$ a semi-martingale we can define the linear filter

$$X_t g = \int_0^t g(t-s)\mathrm{d}Z_s$$

Question: Is the function

$$g \mapsto X_t g$$

an ω-wise linear functional on some Banach space V? Is it continuous?

The solution is integration by parts

$$\int_0^t h(s)\mathrm{d}Z_s = h(t)Z_t - h(0)Z_0 - \int_0^t Z_{s-}h'(s)\mathrm{d}s$$

We have to choose V to be a space of functions where the derivative makes sense.

Sobolev spaces

Definition: $W^{m,2}([0,t])$ is the space of functions that are m times weakly differentiable with the m'th derivative in $L_2([0,t])$.

Several equivalent norms, we choose one that makes $W^{m,2}([0,t])$ a reproducing kernel Hilbert space with kernel given as $R_m = R_m^0 + R_m^1$ where

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

and

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

with $\varphi_k(t) = t^{k-1}/(k-1)!$ for $k = 1, \ldots, m$.

Example: $m = 1$, $R_m(s,r) = s \wedge r + 1$ and

$$< f, g >= f(0)g(0) + \int_0^t f'(s)g'(s)\mathrm{d}t.$$

Penalized maximum likelihood estimation

As a function of $g \in W^{m,2}([0,t])$ the minus-log-likelihood function reads

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

We are aiming at optimizing the penalized minus-log-likelihood

$$l_t(g) + \lambda \int_0^t D^m g(s)^2 \mathrm{d}s$$

over $W^{m,2}([0,t])$.

Main Theorem

Let $\tau_1, \ldots \tau_{N_t}$ denote the jump times for N.

Theorem

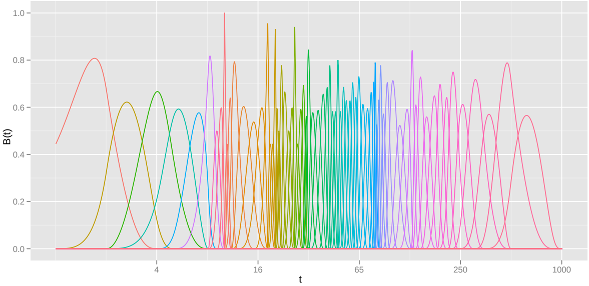
If $\varphi(x) = x + d$ with domain $(-d, \infty)$ then a minimizer of the penalized minus-log-likelihood function over $\Theta((-d, \infty))$ belongs to the finite dimensional subspace of $W^{m,2}([0, t])$ spanned by the functions $\varphi_1, \ldots, \varphi_m$, the functions

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

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

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

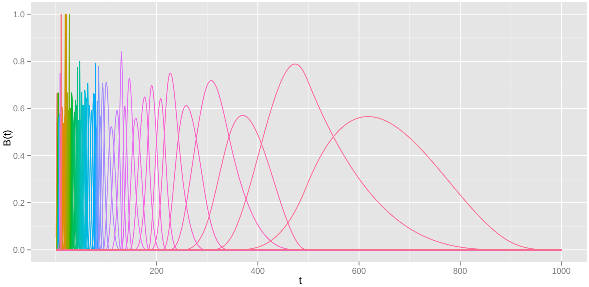
Some spline bases

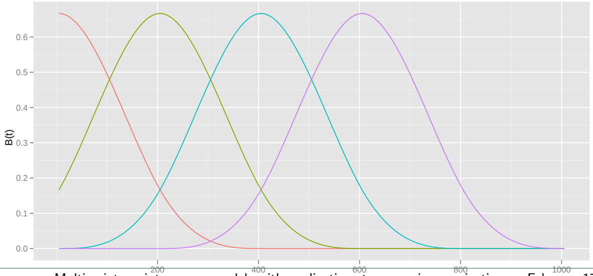


If $(Z_s)_{0 \leq s \leq t}$ is a counting process with jumps $\sigma_1, \ldots, \sigma_{Z_t}$ the h_i basis functions are order $2m$ splines with knots in

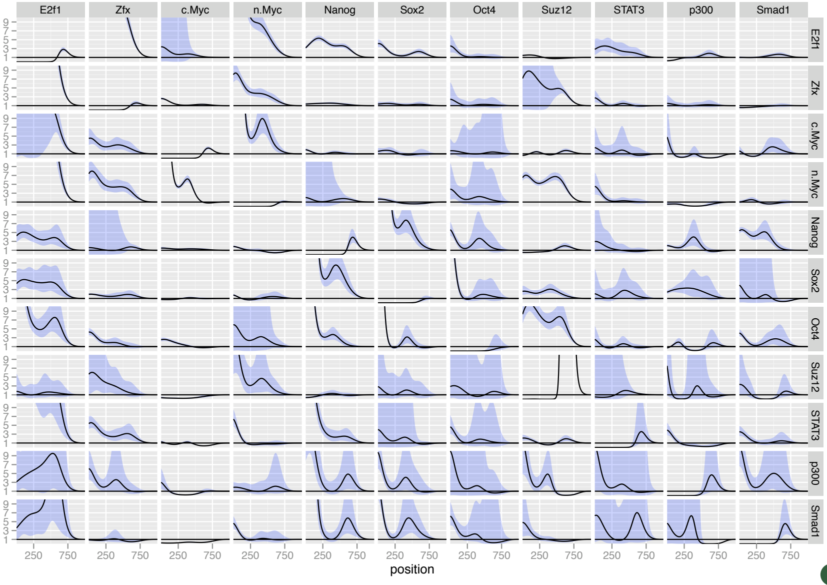
$$\{\tau_i - \sigma_j \mid i = 1, \ldots, N_t, \; j : \sigma_j < \tau_i\}.$$

Some spline bases





Estimated multiplicative effects - ES cells



Further results

Theorem

If φ is continuously differentiable,

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

and

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

Then the gradient of l_t at $g \in \Theta(D)^\circ$ is

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

ppstat

The current implementation in the R-package `ppstat` offers

- a formula interface to model specification, standard summary tables etc.
- standard (e.g. spline) basis function expansions of linear filters.
- inclusion of continuous time covariate effects and additive model specification.

We are currently implementing

- simulations; in particular MCMC simulation of conditional distributions.
- infinite dimensional descent algorithm.
- grouped lasso estimation and interaction term support.

Acknowledgments

Thanks to Lisbeth Carstensen, Albin Sandelin and Ole Winther, Bioinformatics Centre, University of Copenhagen.

Thanks to you for listening.