Multivariate point process models

Niels Richard Hansen
Department of Mathematical Sciences
Ideas and outline

**General aim:** To build and implement a flexible (non-parametric), intensity based framework of multivariate point process models – aka discretely marked point processes.

**Secondary aim:** Development and implementation of model selection techniques – LASSO, grouped LASSO, penalization based ideas – with the multivariate point processes as a testbed. Joint with Patricia Bouret, Vincent Rivoirard.

**Main application:** Organizational model of active transcription regulatory elements along genomes – joint with Lisbeth Carstensen, Albin Sandelin, Ole Winther.
We focus on the distribution of point-like transcriptional regulatory elements\(^1\) at the meso-genomic scale.

\(^1\)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 the different regulators as marks and with 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.
Obtaining and preprocessing data (ChIP-chip)
The 10 Affymetrix ChIP-chip measurements for the ENCODE pilot project segment ENm001.

Histone modifications are not regarded as point-like, figure gives start positions.
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.\(^2\)

\(^2\)Chen et al, Cell 133, 1106-1117, 2008
The Hawkes process

If \( t_{i1} < \ldots < t_{iN_i} \) denote the observed points for transcription regulator \( i \) we specify the intensity process for the occurrence of regulator \( j \) as

\[
\lambda_j(t) = \varphi(\beta_0 + \sum_i \sum_{k: t_{ik} < t} h^{i,j}(t - t_{ik}))
\]

for a fixed \( \varphi \).

We expand the \( h \)-function in a suitable basis;

\[
h^{i,j}(s) = \sum_{b=1}^{B_{i,j}} \beta_{b}^{i,j} B_b(s)
\]

and the minus-log-likelihood in terms of \( \beta = (\beta_0, \beta_{ij}^{ab}) \) is directly expressible.
ppstat

The current implementation in the R package ppstat offers

- Formula interface to model specification.
- Standard (e.g. spline) basis function expansions of linear filters.
- Inclusion of continuous time covariate effects and additive model specification.
- Currently only with $\varphi = \exp$.
- Soon with a more flexible class of $\varphi$’s including $\varphi(x) = x_+$. Why is $\varphi(x) = x_+$ a problem? The minus-log-likelihood is a priori not twice differentiable, but it turns out to be so under certain regularity assumptions.
Some spline bases
Estimated multiplicative effects - ES cells

<table>
<thead>
<tr>
<th>Gene</th>
<th>250</th>
<th>750</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2f1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zfx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.Myc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n.Myc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nanog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sox2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suz12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAT3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smad1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

position

1
3
5
7
9
Interaction terms

The formula

$$\sum_i \sum_{k: t_{ik} < t} h^{i,j}(t - t_{ik})$$

that entered in the intensity is an additive model of linear, univariate filters.

**Question:** Can we capture interaction effects?

**Suggestion:** Introduce bivariate, linear filters

$$\sum_{k: t_{ik} < t} \sum_{l: t_{rl} < t} h^{(i,r),j}(t - t_{ik}, t - t_{rl})$$

and expand them in a suitable bivariate basis.
Bivariate bases

Many possible choices of 2d-bases:

- Thin plate splines

  \[ h(s, t) = \sqrt{(s - s_0)^2 + (t - t_0)^2 \log((s - s_0)^2 + (t - t_0)^2)}. \]

- Tensor products

  \[ h(s, t) = h_1(s)h_2(t). \]

- For illustration we consider an example with a spline based tensor product construction with 16 basis functions.
Bivariate spline tensor product bases
Additive estimated Oct4-Sox2 effect on Nanog

\[ \hat{h}_{\text{Oct4,Nanog}}(s) + \hat{h}_{\text{Sox2,Nanog}}(t) \]
Estimated Oct4-Sox2 interaction effect on Nanog

\[
\hat{h}_{\text{Oct4}, \text{Nanog}}(s) + \hat{h}_{\text{Sox2}, \text{Nanog}}(t) + \hat{h}(\text{Sox2, Oct4}), \text{Nanog}(s, t)
\]
\( \hat{h}(\text{Sox2, Oct4}), \text{Nanog}(s, t) \)
What do we model?

The point process model is a model of the organizational part of evolution.

It is a model of which loci are active
- for a given cell line/organism/tissue
- under given experimental conditions.

The driving stochastic mechanism is evolution – we do not account for measurement noise or biological variation in an explicit way.

Data are integrated over cells – we do not observe single cell transcription factor bindings.
Which conclusions do the model facilitate?

Working assumption:

Organizational association $\sim$ Functional association

It is a systemic analysis that facilitate phrasing and answering question of a systemic nature.

Example: For two tissues find systemic differences in the organization of active regulatory loci.
Thanks ...

... to Lisbeth Carstensen, Albin Sandelin and Ole Winther ...

... and to all of you for attending this meeting.