Multivariate sparse dynamic process modeling and inference

Niels Richard Hansen
Department of Mathematical Sciences
Spike tracks from turtle motor neurons

Turtle motor neurons are used as models for neuronal activity.
Spike tracks from turtle motor neurons

Turtle motor neurons are used as models for neuronal activity.

The (dead) turtle is stimulated by scratching and recordings from one or more electrodes register the spikes.
Data
Data from stimulation period
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 \(\theta \in \Theta\).
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

\[
l_t(\theta) = \int_0^t \lambda_s(\theta) ds - \int_0^t \log \lambda_s(\theta) N(ds) \]

We will study penalized maximum-likelihood estimation of the parameter \(\theta\).
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

\[
l_t(\theta) = \int_0^t \lambda_s(\theta)ds - \int_0^t \log \lambda_s(\theta)N(ds)
\]

We will study penalized maximum-likelihood estimation of the parameter \(\theta\).
What are the goals?

\[ \lambda_s(\theta) =? \]
What are the goals?

\[ \lambda_s(\theta) = ? \]

- Use ensembles of spike patterns for **decoding** of external stimuli signals\(^1\)

---

What are the goals?

\[ \lambda_s(\theta) =? \]

- Use ensembles of spike patterns for **decoding** of external stimuli signals\(^1\)
- Learn the **connectivity graph** for an ensemble of neurons from data\(^2\).

---


What are the goals?

\[ \lambda_s(\theta) = ? \]

- Use ensembles of spike patterns for \textit{decoding} of external stimuli signals\textsuperscript{1}
- Learn the \textit{connectivity graph} for an ensemble of neurons from data\textsuperscript{2}.
- Learn from data how signals propagate at the spike level among connected neurons (the functional forms).

\textsuperscript{1}Pillow et al. \textit{Spatio-temporal correlations and visual signalling in a complete neuronal population}. Nature, 454. 2008

What are the goals?

\[ \lambda_s(\theta) = ? \]

- Use ensembles of spike patterns for **decoding** of external stimuli signals\(^1\).
- Learn the **connectivity graph** for an ensemble of neurons from data\(^2\).
- Learn from data how signals propagate at the spike level among connected neurons (the functional forms).
- Learn if the network and/or the functional forms are adaptive to stimuli.

---


A self-exciting model

With \((N_t)_{t \geq 0}\) the counting process for the spike times and \(\tau_1, \ldots, \tau_{N_t}\) the jumps consider the model

\[
\lambda_t(g) = \phi \left( \sum_{j: \tau_j < t} g(t - \tau_j) \right)
\]

For a multivariate counting process, \((N_{it})_{t \geq 0}, i = 1, \ldots, K\), the jumps consider the model

\[
\lambda_{it}(g) = \phi \left( \sum_{i=1}^K \sum_{j: \tau_{ij} < t} g(t - \tau_{ij}) \right)
\]

which is the non-linear Hawkes process.

With \(\phi(x) = x + d\) we get the linear Hawkes process.

A self-exciting model

With \((N_t)_{t \geq 0}\) the counting process for the spike times and \(\tau_1, \ldots, \tau_{N_t}\) the jumps consider the model

\[
\lambda_t(g) = \phi \left( \sum_{j: \tau_j < t} g(t - \tau_j) \right)
\]

For a multivariate counting process, \((N^i_t)_{t \geq 0, i=1,\ldots,K}\),

\[
\lambda_t^k(g) = \phi \left( \sum_{i=1}^{K} \sum_{j: \tau^i_j < t} g^{ik}(t - \tau^i_j) \right),
\]

which is the non-linear Hawkes process\(^3\).

A self-exciting model

With \((N_t)_{t \geq 0}\) the counting process for the spike times and \(\tau_1, \ldots, \tau_{N_t}\) the jumps consider the model

\[
\lambda_t(g) = \phi \left( \sum_{j: \tau_j < t} g(t - \tau_j) \right)
\]

For a multivariate counting process, \((N^i_t)_{t \geq 0, i=1,\ldots,K}\),

\[
\lambda^k_t(g) = \phi \left( \sum_{i=1}^{K} \sum_{j: \tau^i_j < t} g^{i^k}(t - \tau^i_j) \right),
\]

which is the non-linear Hawkes process\(^3\). With \(\phi(x) = x + d\) we get the linear Hawkes process.

A self-exciting model

With \((N_t)_{t \geq 0}\) the counting process for the spike times and \(\tau_1, \ldots, \tau_{N_t}\) the jumps consider the model

\[
\lambda_t(g) = \phi \left( \sum_{j: \tau_j < t} g(t - \tau_j) \right) = \phi \left( \int_0^{t-} g(t - s) N(ds) \right)
\]

For a multivariate counting process, \((N^i_t)_{t \geq 0, i = 1, \ldots, K}\),

\[
\lambda^k_t(g) = \phi \left( \sum_{i=1}^{K} \sum_{j: \tau^i_j < t} g^{ik}(t - \tau^i_j) \right),
\]

which is the \textbf{non-linear Hawkes process}\(^3\). With \(\phi(x) = x + d\) we get the \textit{linear} Hawkes process.

Intensities

Example of estimated linear filters and log intensity with $\phi = \exp$. 
Local independence

We say that $N^k_t$ is locally independent of $N^i_t$ if the $\mathcal{F}_t$-intensity is $\sigma((N^{-i}_s)_{s \in [0,t)})$-adapted.
Local independence

We say that $N_t^k$ is **locally independent** of $N_t^i$ if the $\mathcal{F}_t$-intensity is $\sigma((N_s^{i-})_{s \in [0,t)})$-adapted. We write $i \not\rightarrow k$ and this defines the oriented local independence graph $G = (\{1, \ldots, K\}, E)$. 
Local independence

We say that $N^k_t$ is locally independent of $N^i_t$ if the $\mathcal{F}_t$-intensity is $\sigma((N^{s-i}_s)_{s \in [0,t]})$-adapted. We write $i \not\rightarrow k$ and this defines the oriented local independence graph $G = ([1, \ldots, K], E)$.

For the non-linear Hawkes process $(i, k) \in E$ if and only if $g_{ik} \neq 0$. 
Local independence

We say that $N^k_t$ is locally independent of $N^i_t$ if the $\mathcal{F}_t$-intensity is $\sigma((N^-_s)^{i\in[0,t]})$-adapted. We write $i \not\rightarrow k$ and this defines the oriented local independence graph $G = (\{1, \ldots, K\}, E)$.

For the non-linear Hawkes process $(i, k) \in E$ if and only if $g^{ik} \neq 0$.

The concept was introduced by Tore Schweder\textsuperscript{4} for finite state space Markov processes and extended to general point processes by Vanessa Didelez\textsuperscript{5}.

\textsuperscript{4} Composable Markov Processes, J. Appl. Prob. (1970), 7(2)
\textsuperscript{5} Graphical models for marked point processes based on local independence, J. R. Statist. Soc. B (2008) 70(1)
Local independence

We say that $N_t^k$ is locally independent of $N_t^i$ if the $\mathcal{F}_t$-intensity is $\sigma((N_s^-)^i_{s \in [0,t)})$-adapted. We write $i \not\rightarrow k$ and this defines the oriented local independence graph $G = (\{1, \ldots, K\}, E)$.

For the non-linear Hawkes process $(i, k) \in E$ if and only if $g^{ik} \neq 0$.

The concept was introduced by Tore Schweder$^4$ for finite state space Markov processes and extended to general point processes by Vanessa Didelez$^5$.

The local independence graph has a causal connotation related to Granger Causality, and it finds applications in the literature on causal inference.

---

$^4$ Composable Markov Processes, J. Appl. Prob. (1970), 7(2)

$^5$ Graphical models for marked point processes based on local independence, J. R. Statist. Soc. B (2008) 70(1)
Joint likelihood

If $\lambda^i(\theta_i)$ is parametrized by $\theta_i \in \Theta_i$ the joint minus-log-likelihood is

$$
\sum_{i=1}^{K} \int_{0}^{t} \lambda^i_s(\theta_i) ds - \int_{0}^{t} \log \lambda^i_s(\theta_i) N^i(ds),
$$

$$
l^i_t(\theta_i)
$$
Joint likelihood

If $\lambda^i(\theta_i)$ is parametrized by $\theta_i \in \Theta_i$ the joint minus-log-likelihood is

$$\sum_{i=1}^{K} \int_{0}^{t} \lambda^i_s(\theta_i) ds - \int_{0}^{t} \log \lambda^i_s(\theta_i) N^i(ds),$$

hence if we have variation independence of $\theta_1, \ldots, \theta_K$ we need to minimize each $l^i_t(\theta_i)$ separately.
Joint likelihood

If $\lambda^i(\theta_i)$ is parametrized by $\theta_i \in \Theta_i$ the joint minus-log-likelihood is

$$
\sum_{i=1}^{K} \left[ \int_0^t \lambda^i_s(\theta_i) \, ds - \int_0^t \log \lambda^i_s(\theta_i) N^i(\, ds) \right] = l^i_t(\theta_i)
$$

hence if we have variation independence of $\theta_1, \ldots, \theta_K$ we need to minimize each $l^i_t(\theta_i)$ separately.

We can think of the $i$'th term as a likelihood for a (conditional) model of $i$'th counting process and we consider only such models.
Joint likelihood

If $\lambda^i(\theta_i)$ is parametrized by $\theta_i \in \Theta_i$ the joint minus-log-likelihood is

$$\sum_{i=1}^{K} \left[ \int_0^t \lambda^i_s(\theta_i) ds - \int_0^t \log \lambda^i_s(\theta_i) N^i(ds), \right]$$

hence if we have variation independence of $\theta_1, \ldots, \theta_K$ we need to minimize each $l^i_t(\theta_i)$ separately.

We can think of the $i$'th term as a likelihood for a (conditional) model of $i$'th counting process and we consider only such models.

Which class of models to consider?
Generalized linear point process models

\((X_t)_{t \geq 0}\) is a predictable, càdlàg process process with values in \(V^*\) – the dual of the vector space \(V\), and

\[ \Theta(D) = \{ \beta \in V \mid X_s - \beta \in D \text{ for all } s \in [0, t] \text{ } P\text{-a.s.} \}. \]
Generalized linear point process models

$(X_t)_{t \geq 0}$ is a predictable, càdlàg process process with values in $V^*$ -- the dual of the vector space $V$, and

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

$\phi : D \to [0, \infty)$ and $(Y_t)_{t \geq 0}$ is a predictable, càdlàg process with values in $[0, \infty)$. 
Generalized linear point process models

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

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

$\phi : D \rightarrow [0, \infty)$ and $(Y_t)_{t \geq 0}$ is a predictable, càdlàg process with values in $[0, \infty)$.

**Definition**

A generalized linear point process model on $[0, t]$ is the statistical model with parameter space $\Theta(D)$ such that for $\beta \in \Theta(D)$ the point process on $[0, t]$ has intensity

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

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

$$X_tg = \int_0^t g(t - s) dZ_s$$
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

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

The function

$$g \mapsto X_t g$$

is an $\omega$-wise continues linear functional on the Sobolev space $W^{m,2}([0, t])$ for $m \geq 1$. 
Stochastic integrals as linear functionals

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

\[
X_t g = \int_0^t g(t - s) \, dZ_s
\]

The function

\[
g \mapsto X_t g
\]

is an \( \omega \)-wise continues linear functional on the Sobolev space \( W^{m,2}([0, t]) \) for \( m \geq 1 \).

Proof by integration by parts

\[
\int_0^t h(s) \, dZ_s = h(t)Z_t - h(0)Z_0 - \int_0^t Z_s \, h'(s) \, ds
\]
Penalized maximum likelihood estimation

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

$$l_t(g) = \int_0^t Y_s \phi \left( \int_0^{s-} g(s - u)dZ_u \right) ds - \int_0^t \log(Y_s \phi \left( \int_0^{s-} g(s - u)dZ_u \right)) N(ds)$$
Penalized maximum likelihood estimation

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

$$l_t(g) = \int_0^t Y_s \phi \left( \int_0^{s^-} g(s - u)dZ_u \right) ds - \int_0^t \log(Y_s \phi \left( \int_0^{s^-} g(s - u)dZ_u \right))N(ds)$$

We are optimizing the penalized minus-log-likelihood

$$l_t(g) + \lambda \int_0^t D^m g(s)^2 ds$$

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

Dimensions: 28 x 28

Column
Row
5 10 15 20 25
5 10 15 20 25
Estimates
Estimates

![Graphs showing time lag estimates](image-url)
A theorem

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

Theorem

If $\phi(x) = x + d$ 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 $\phi_1, \ldots, \phi_m$, the functions $h_i(r) = \int_0^{\tau_i^-} R^1(\tau_i - u, r) dZ_u$ for $i = 1, \ldots, N_t$ together with the function

$$f(r) = \int_0^t Y_s \int_0^s R^1(s - u, r) \, dZ_u \, ds = \int_0^t \int_0^t Y_s R^1(s - u, r) \, ds \, dZ_u.$$ 

$$R_m^1(s, r) = \int_0^{s \wedge r} \frac{(s - u)^{m-1} (r - u)^{m-1}}{((m-1)!)^2} \, du, \quad \phi_k(t) = t^{k-1}/(k - 1)!$$
Counting process integrals

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

\[
\{ \tau_i - \sigma_j \mid i = 1, \ldots, N_t, j : \sigma_j < \tau_i \}.
\]
Another theorem

**Theorem**

If $\phi$ is continuously differentiable,

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

and

$$
f_g(r) = \int_0^t \int_0^{s-} Y_s \phi' \left( \int_0^{s-} g(s - u) dZ_u \right) R^1(s - u, r) ds dZ_u.
$$

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

$$
\nabla l_t(g) = f_g - \sum_{i=1}^{N_t} \phi' \left( \int_0^{\tau_i -} g(\tau_i - u) dZ_u \right) \phi \left( \int_0^{\tau_i -} g(\tau_i - u) dZ_u \right) \eta_i.
$$
The problem with explosion

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

\[
\mathcal{L}_t = \exp \left( t - \int_0^t \lambda_s \, ds + \int_0^t \log \lambda_s \, N(ds) \right)
\]

may not be a martingale.
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

\[ L_t = \exp \left( t - \int_0^t \lambda_s \, ds + \int_0^t \log \lambda_s \, N(ds) \right) \]

may not be a martingale.

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

\[
\mathcal{L}_t = \exp \left( t - \int_0^t \lambda_s ds + \int_0^t \log \lambda_s N(ds) \right)
\]

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 only have a dominated statistical model if we restrict our attention to combinations of \(\phi\) and processes \((X_t)_{t \geq 0}\) such that the likelihood process is a martingale.
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

\[
\mathcal{L}_t = \exp \left( t - \int_0^t \lambda_s ds + \int_0^t \log \lambda_s N(ds) \right)
\]

may not be a martingale.

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

We only have a dominated statistical model if we restrict our attention to combinations of \(\phi\) and processes \((X_t)_{t \geq 0}\) such that the likelihood process is a martingale.

For non-exploding data \(\mathcal{L}_t(\theta)\) is, however, always sensible for relative comparisons of models.
Generalization of local independence

There are abstract generalizations of local independence to semi-martingales, of particular interest are the solutions to the SDE

\[ dX_t = G(X_t)dt + DdB_t \]

where \((B_t)_{t \geq 0}\) is \(p\)-dimensional Brownian motion, \(D\) is diagonal and \(G : \mathbb{R}^p \rightarrow \mathbb{R}^p\).
Generalization of local independence

There are abstract generalizations of local independence to semi-martingales, of particular interest are the solutions to the SDE

\[ \text{d}X_t = G(X_t)\text{d}t + D\text{d}B_t \]

where \((B_t)_{t \geq 0}\) is \(p\)-dimensional Brownian motion, \(D\) is diagonal and \(G : \mathbb{R}^p \to \mathbb{R}^p\).

We say that \(X_t^k\) is **locally independent** of \(X_t^i\) if \(G_k\) does not depend upon the \(i\)'th coordinate.
A Gaussian process

With $G(x) = B(x - A)$ the solution to the SDE

$$dX_t = B(X_t - A)dt + dB_t$$

becomes a Gaussian process with normal discrete time transitions

$$X_t \sim N(\xi(x_0, t), \Sigma(t))$$

where

$$\xi(x, t) = A + e^{tB}(x - A)$$  \hfill (1)$$

$$\Sigma(t) = \int_0^t e^{sB} D^2 e^{sB^T} ds$$  \hfill (2)$$
A Gaussian process

With $G(x) = B(x - A)$ the solution to the SDE

$$dX_t = B(X_t - A)dt + DdB_t$$

becomes a Gaussian process with normal discrete time transitions

$$X_t \sim N(\xi(x_0, t), \Sigma(t))$$

where

$$\xi(x, t) = A + e^{tB}(x - A) \quad (1)$$
$$\Sigma(t) = \int_0^t e^{sB}D^2e^{sB^T}ds \quad (2)$$

The minus-log-likelihood for discrete observations

$$l_x(A, B, D) = \sum_{i=2}^{n} \left[ (X_{t_i} - \xi(x_{t_{i-1}}, \Delta_i))^T \Sigma(\Delta_i)^{-1}(X_{t_i} - \xi(x_{t_{i-1}}, \Delta_i)) ight]$$
$$+ \log \det \Sigma(\Delta_i) \right].$$
A Gaussian process

With \( G(x) = B(x - A) \) the solution to the SDE

\[
dX_t = B(X_t - A)dt + DdB_t
\]

becomes a Gaussian process with normal discrete time transitions

\[
X_t \sim N(\xi(x_0, t), \Sigma(t))
\]

where

\[
\xi(x, t) = A + e^{tB}(x - A) \quad (1)
\]

\[
\Sigma(t) = \int_0^t e^{sB} D^2 e^{sB^T} ds \quad (2)
\]

The pseudo minus-log-likelihood for discrete observations

\[
l_x(A, B) = \sum_{i=2}^{n} (x_{t_i} - \xi(x_{t_{i-1}}, \Delta_i))^T (x_{t_i} - \xi(x_{t_{i-1}}, \Delta_i))
\]

\[
= \sum_{i=2}^{n} ||x_{t_i} - A - e^{\Delta_i B}(x_{t_{i-1}} - A)||^2.
\]
A Gaussian process

With $G(x) = B(x - A)$ the solution to the SDE

$$dX_t = B(X_t - A)dt + DdB_t$$

becomes a Gaussian process with normal discrete time transitions

$$X_t \sim N (\xi(x_0, t), \Sigma(t))$$

where

$$\xi(x, t) = A + e^{tB}(x - A) \quad (1)$$

$$\Sigma(t) = \int_0^t e^{sB}D^2 e^{sB^T}ds \quad (2)$$

The $\ell_1$-penalized pseudo minus-log-likelihood for discrete observations

$$\tilde{l}_x(A, B) = \sum_{i=2}^{n} \left\| x_{t_i} - A - e^{\Delta_i B}(x_{t_{i-1}} - A) \right\|^2 + \lambda \sum_{i,j} |B_{ij}|$$
The non-linear penalized least squares problem

Minimization of

$$\tilde{l}_x(B) = \sum_{i=2}^{n} ||x_{t_i} - e^{\Delta_i B} x_{t_{i-1}}||^2 + \lambda \sum_{i,j} |B_{ij}|$$

is a non-linear least squares problem.
The non-linear penalized least squares problem

Minimization of

\[ \tilde{l}_x(B) = \sum_{i=2}^{n} \left\| x_{t_i} - e^{\Delta_i B} x_{t_i-1} \right\|^2 + \lambda \sum_{i,j} |B_{ij}| \]

is a non-linear least squares problem.

A coordinate wise Gauss-Newton method iteratively optimize one coordinate at a time for the linear least squares problem

\[ \tilde{l}_x(B_0 + B) \approx \sum_{i=2}^{n} \left\| r_i(B_0) - D_B e^{\Delta_i B_0(B)} x_{t_i-1} \right\|^2 + \lambda \sum_{i,j} |B_{ij}| \]

\[ r_i(B) = x_{t_i} - e^{\Delta_i B} x_{t_i-1}. \]
The non-linear penalized least squares problem

Minimization of

\[ \tilde{l}_x(B) = \sum_{i=2}^{n} \|x_{t_{i}} - e^{\Delta_{i}B} x_{t_{i-1}}\|^2 + \lambda \sum_{i,j} |B_{ij}| \]

is a non-linear least squares problem.

A coordinate wise Gauss-Newton method iteratively optimize one coordinate at a time for the linear least squares problem

\[ \tilde{l}_x(B_0 + B) \simeq \sum_{i=2}^{n} \|r_i(B_0) - D_B e^{\Delta_{i}B_0}(B)x_{t_{i-1}}\|^2 + \lambda \sum_{i,j} |B_{ij}| \]

\[ r_i(B) = x_{t_{i}} - e^{\Delta_{i}B} x_{t_{i-1}}. \] Computations of \(e^{\Delta_{i}B}\) and \(D_B e^{\Delta_{i}B}\) dominates.
The non-linear penalized least squares problem

Minimization of

\[ \tilde{l}_x(B) = \sum_{i=2}^{n} ||x_{t_i} - e^{\Delta^i B}x_{t_i-1}||^2 + \lambda \sum_{i,j} |B_{ij}| \]

is a non-linear least squares problem.

A coordinate wise Gauss-Newton method iteratively optimize one coordinate at a time for the linear least squares problem

\[ \tilde{l}_x(B_0 + B) \simeq \sum_{i=2}^{n} ||r_i(B_0) - D_B e^{\Delta^i B_0(B)}x_{t_i-1}||^2 + \lambda \sum_{i,j} |B_{ij}| \]

\[ r_i(B) = x_{t_i} - e^{\Delta^i B}x_{t_i-1}. \] Computations of \(e^{\Delta^i B}\) and \(D_B e^{\Delta^i B}\) dominates. Problem does not “decouple” into regression problems for each coordinate.
Toy example: $p = 20$, $A = 0$, $D = I$
Toy example: $p = 20$, $A = 0$, $D = I$
Toy example: $p = 20$, $A = 0$, $D = I$
Toy example: $p = 20, A = 0, D = I$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Toy example, Lasso estimation with $N = 100$
Is the discrete time AR-process not enough?

For equidistant observations the process is a vector AR(1)-process (with correlated noise)

\[ X_i - X_{i-1} = (e^{\Delta B} - I)X_{i-1} + \epsilon_i \]
Is the discrete time AR-process not enough?

For equidistant observations the process is a vector AR(1)-process (with correlated noise)

\[ X_i - X_{i-1} = (e^{\Delta B} - I)X_{i-1} + \epsilon_i \]

Is sparseness of \( B \) not related to sparseness of

\[ e^{\Delta B} - I = \Delta B + O(\Delta^2). \]
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
Is the discrete time AR-process not enough?
What about the (asymptotic) variance?

When the spectrum of $B$ has no positive real parts

$$\Sigma(t) \to \Gamma = \int_0^\infty e^{sB} D^2 e^{sB^T} ds$$

for $t \to \infty$. 
What about the (asymptotic) variance?

When the spectrum of $B$ has no positive real parts

$$\Sigma(t) \to \Gamma = \int_0^\infty e^{sB} D^2 e^{sB^T} ds$$

for $t \to \infty$.

The variance $\Gamma$ for the invariant distribution solves

$$(B \otimes I + I \otimes B) \Gamma = -D^2.$$
What about the (asymptotic) variance?

When the spectrum of $B$ has no positive real parts

$$\Sigma(t) \to \Gamma = \int_0^\infty e^{sB}D^2e^{sB^T}ds$$

for $t \to \infty$.

The variance $\Gamma$ for the invariant distribution solves

$$(B \otimes I + I \otimes B)\Gamma = -D^2.$$ 

If $BB^T = B^TB$ and $BD = DB$ then $B\Gamma = \Gamma B$ and

$$\Gamma = -(B + B^T)^{-1}D^2.$$
Toy example asymptotic variance
Acknowledgements

- Rune Berg and Susanne Ditlevsen, University of Copenhagen, for collaboration on the neuron data analysis.
- Lisbeth Carstensen, Albin Sandelin and Ole Winther, Bioinformatics Centre, University of Copenhagen, for collaboration on the use of Hawkes processes for modeling regulatory cites on genomes.
- Alexander Sokol, University of Copenhagen, who currently work with estimation of sparse SDE’s.
- Patricia Bouret, Nice Sophia-Antipolis University, and Vicent Rivoirard, Université Paris Sud - Orsay, with whom I work on oracle inequalities for the multivariate Hawkes process.