Continuous-Time Models

Siqi Liu
CS3750 Advanced Machine Learning

Outline

- Continuous-time time series
- Event time series

Outline

- Continuous-time time series
- Event time series

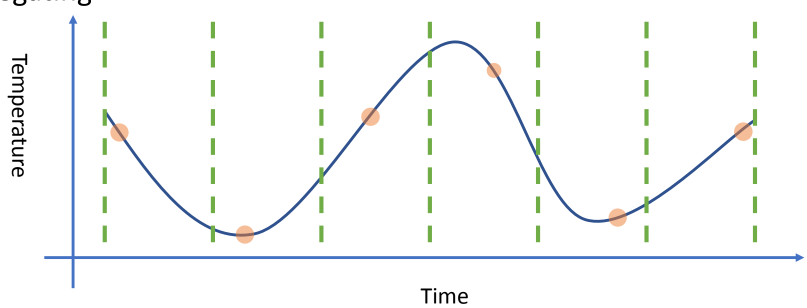
Discrete-time time series

- Time series observed at **regularly spaced** intervals of time
  - E.g., every day or every hour
- Formally represented by $\{y_t : t = 1,2, \dots\}$
- Essentially, "time" is discrete

Time	Temperature (C)
8:00 AM	5
9:00 AM	7
10:00 AM	10
…	..

Discrete-time time series

- However, the underlying data source is always in continuous time
- We get discrete-time time series by
  - Sampling regularly
  - Binning
  - Aggregating



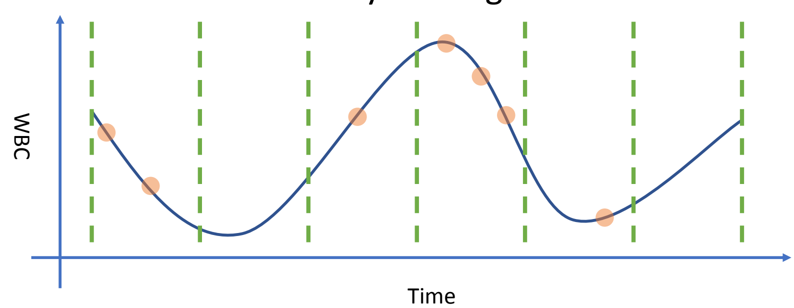
Continuous-time time series

- Time series observed at **irregularly spaced** intervals of time
- Formally represented by $\{y(t): t \in \mathbb{R}\}$

Time	Blood pressure (diastolic)
5/10/2018 8:33 AM	75
5/17/2018 3:10 PM	88
8/10/2018 10:00 AM	85
…	..

Continuous-time time series

- In some domains, regularly sampling time series is NOT **feasible** or **desired**
  - blood pressure
  - white blood cell (WBC) count
- We can still discretize the time by binning



Models for discrete-time time series

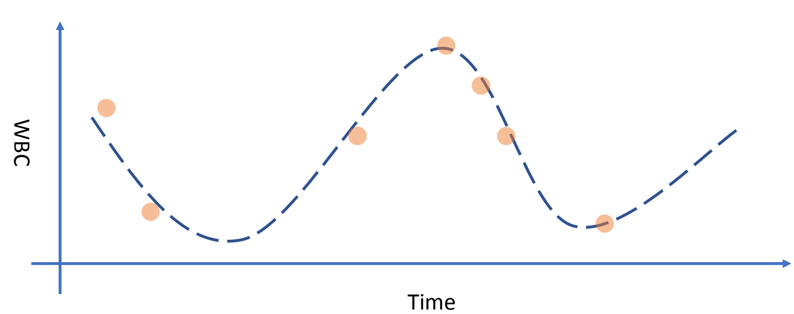
- We have a set of well-studied models for discrete-time time series
- Regression models
  - AR, MA, ARIMA
- State-space models
  - Linear dynamical systems
- Do we have models **directly applicable** to continuous-time time series?

Outline

- Continuous-time time series
  - Regression model
  - State-space model
- Event time series

Curve fitting for continuous-time time series

- Observe data $\{y(t_n)\}_{n=1}^{N}$ at irregularly spaced time points
- Assume $y(t_n) = f(t_n) + \eta_n$, where η_n is additive noise
- Our goal is to find $f(t)$ given the data



GP for curve fitting

- Gaussian processes (GP) provide an elegant solution to curve fitting (probabilistically)
- Recall that $GP(m, k)$ is a stochastic process defined by
  - Mean function $m(x)$
  - Covariance function $k(x, x')$
  - x, x' are inputs of the GP
- For a set of inputs $\boldsymbol{x} = \{x_1, x_2, \dots, x_N\}$ the outputs have the multivariate Gaussian distribution $N\big(\boldsymbol{m}(\boldsymbol{x}), \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\big)$

GP for curve fitting

- Given observed time series $\boldsymbol{y} = \{y(t_n) \in \mathbb{R}\}_{n=1}^{N}$ at $\boldsymbol{t} = (t_1, t_2, \dots, t_N)$
- Assuming $y(t_n) = f(t_n) + \eta_n$
- To find $f(t)$ or $y(t)$
- We can assume $y(t) \sim GP(m, k)$ with t being the input to the GP
- Then compute the posterior distribution $p(y(t)|\boldsymbol{y})$

GP prediction

- To make predictions $\boldsymbol{y}_* = \boldsymbol{y}(\boldsymbol{t}_*)$ at new time points $\boldsymbol{t}_*$
- We invoke the standard results for GP

$$p(\boldsymbol{y}_*) = N(\boldsymbol{m}_*, \boldsymbol{C}_*)$$

- where
  - $\boldsymbol{m}_* = \boldsymbol{m}(\boldsymbol{t}_*) + \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t})\boldsymbol{K}(\boldsymbol{t}, \boldsymbol{t})^{-1}\big(\boldsymbol{y}(\boldsymbol{t}) - \boldsymbol{m}(\boldsymbol{t})\big)$
  - $\boldsymbol{C}_* = \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) - \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t})\boldsymbol{K}(\boldsymbol{t}, \boldsymbol{t})^{-1}\boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t})^T$

Covariance function

- Different types of kernels can be used as the covariance function
  - White noise $k(x, x') = \sigma^2\delta(x - x')$
  - Squared exponential $k(x, x') = h^2 \exp\left[-\left(\frac{x-x'}{\lambda}\right)^2\right]$
  - Periodic squared exponential $k(x, x') = h^2 \exp\left[-\frac{1}{2w^2}\sin^2\left(\pi\left|\frac{x-x'}{T}\right|\right)\right]$
- They can be combined together by summation and multiplication

Mean function

- If we have clear domain knowledge, we can put it in
  - E.g., if we know there is a linear trend, then $m(t) = \beta_1 t + \beta_0$ would be a good choice
- Most of the time, we are not certain about it, so we put a vague flat mean $m(t) = \beta_0$ or even $m(t) = 0$

Multivariate time series

- So far we assumed the time series is univariate (one dimensional)
- What if the time series is multivariate (multi-dimensional)
- For example, for a patient, we simultaneously collect over time:
  - blood pressures
  - heart beat rates
  - white blood cell counts
- Can we still use GP?

GP for multivariate time series

- We can put a label $l = 1,2, \dots, D$ on each dimension
- The data can be represented as $\{(y_n, t_n, l_n)\}_{n=1}^{N}$
- Or equivalently $\{y(t_n, l_n)\}_{n=1}^{N}$
- The second representation shows that we can just treat the label as another input in addition to the time
- Define

$$m(t, l) = \beta_l, \qquad k\big((t, l), (t', l')\big) = k_t(t, t')k_l(l, l')$$

- Then we can assume $y(t, l) \sim GP(m, k)$

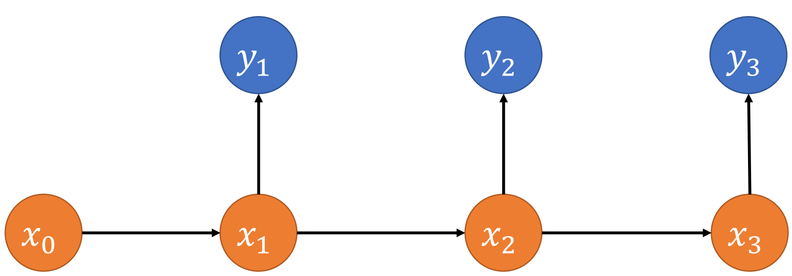
Outline

- Continuous-time time series
  - Regression model
  - State-space model
- Event time series

Linear dynamical system

- Recall for discrete-time time series, we can use hidden states x_t to track the underlying dynamics of the time series
  - $p(x_t|x_{t-1}) = N(x_t|Ax_{t-1}, \Gamma)$
  - $p(y_t|x_t) = N(y_t|Cx_t, \Sigma)$
- Although conditionally independent
  - $p(y_t|x_t, y_1, y_2 \dots, y_{t-1}) = p(y_t|x_t)$
- Marginally y_t could depend on all the past observations $y_1, y_2, \dots y_{t-1}$

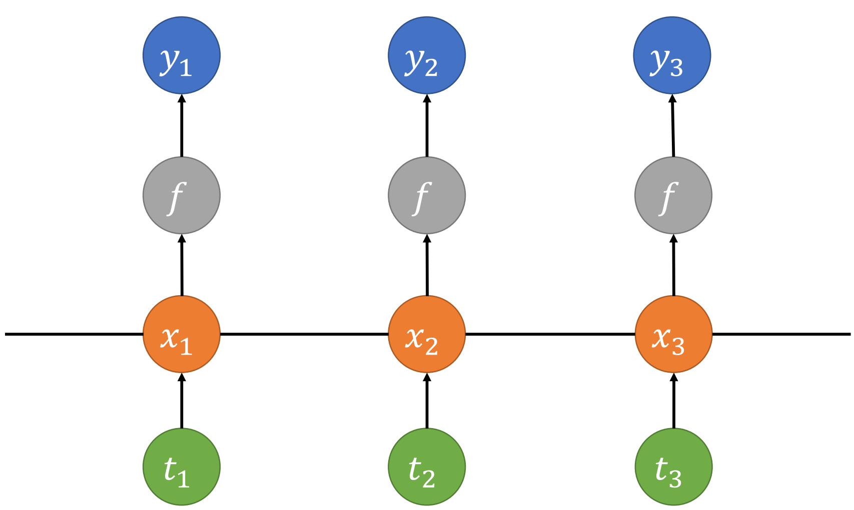
Linear dynamical system



GP dynamical system

- Assume we observe $\{(\boldsymbol{y}_n, t_n)\}_{n=1}^N$, where $\boldsymbol{y}_n \in \mathbb{R}^D$
- Let a set of GPs define the hidden states
  - $x_q(t) \sim GP\big(0, k_x(t,t')\big),\ q = 1,2,\dots,Q$
- Have emission functions take the hidden states to the observations
  - $y_{nd} = f_d(\boldsymbol{x}_n) + \epsilon_{nd},\ \ \epsilon_{nd} \sim N(0, \beta^{-1})$
  - $\boldsymbol{x}_n = \big[x_1(t_n), x_2(t_n), \dots, x_Q(t_n)\big]^T$
- Assume each emission function is drawn from a GP
  - $f_d(\boldsymbol{x}) \sim GP\left(0, k_f(\boldsymbol{x}, \boldsymbol{x}')\right),\ d = 1,2,\dots,D$

GP dynamical system



Likelihood function

- Notations
  - $X \in \mathbb{R}^{N\times Q}$ collect all $x_{nq} = x_q(t_n)$
  - $F \in \mathbb{R}^{N\times D}$ collect all $f_d(\boldsymbol{x}_n)$
  - $Y \in \mathbb{R}^{N\times D}$ collect all y_{nd}
- Joint distribution

$$p(Y, F, X|\boldsymbol{t}) = p(Y|F)p(F|X)p(X|\boldsymbol{t})$$

- Marginal distribution

$$p(Y|\boldsymbol{t}) = \int p(Y|F)p(F|X)p(X|\boldsymbol{t})dXdF$$

- The marginal likelihood is intractable
- Approximated by variational lower bound

Prediction

- Given a set of new time points $\boldsymbol{t}_*$
- Let F_* and Y_* be the values of $f(\cdot)$ and $y(\cdot)$ at those points

$$p(Y_*|Y) = \int p(Y_*, F_*, X_*|Y)dF_*dX_*$$

$$= \int p(Y_*|F_*)p(F_*|X_*, Y)p(X_*|Y)dF_*dX_*$$

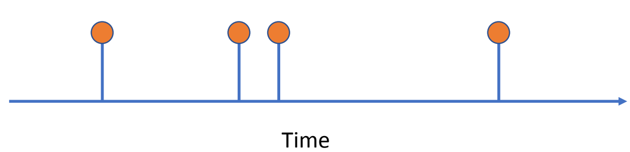
- Using variational approximation for $p(F_*|X_*, Y)$ and $p(X_*|Y)$
- We can find analytically the mean and covariance of Y_*

Outline

- Continuous-time time series
- Event time series

Event time series

- **Discrete** events in continuous time
  - Earthquakes
  - Accidents
- Different from continuous-time time series
- Represented as points on a time line



Distribution of events

- A sequence of events can be represented by their times $\boldsymbol{t} = \{t_n\}_{n=1}^{N}$
  - $0 < t_1 < t_2 < \cdots < t_N < \infty$
  - Time in $[0, \infty)$
  - No coincidence
- A temporal point process is a probability distribution of points over the time line
- It defines the density $f(\boldsymbol{t})$ for any $\boldsymbol{t}$

Temporal point process

- Let H_t denote the history of the events at time t including t

$$H_t = \{t_n : t_n \leq t\}$$

- Let H_{t-} denote the history of events at time t excluding t

$$H_{t-} = \{t_n : t_n < t\}$$

- Let $t_0 = 0$ and $H_0 = \emptyset$
- The joint density function for the events is

$$f(\boldsymbol{t}) = \prod_{n=1}^{N} f\left(t_n \middle| H_{t_{n-1}}\right)$$

- We can define a point process by specifying $f(t_n | H_{t_{n-1}})$

Renewal process

- A renewal process is a point process with IID interevent times
$$f\left(t_n \middle| H_{t_{n-1}}\right) = g(t_n - t_{n-1}) = g(\Delta t_n)$$
- g is the density function of a probability distribution on $(0, \infty)$
  - E.g., $g(t) = e^{-t}$, that is $\Delta t_n \sim Exp(1)$

Conditional intensity function

- Let t_n be the last point before t.
- We derive the cumulative distribution function
$$F\left(t \middle| H_{t_n}\right) = \int_{t_n}^{t} f\left(u \middle| H_{t_n}\right) du$$
- Probability of next point in $(t_n, t]$
- Then the conditional intensity function (CIF) is defined by
$$\lambda^*(t) = \frac{f\left(t \middle| H_{t_n}\right)}{1 - F\left(t \middle| H_{t_n}\right)}$$

Conditional intensity function

- CIF defines the rate of events at time t
$$\lambda^*(t)dt = E[N([t, t + dt])|H_{t-}]$$
- $N(A)$ is the number of points in the interval A
- We can define a point process by specifying its CIF

Poisson process

- A homogeneous Poisson process is defined by
$$\lambda^*(t) = \mu$$
  - The numbers of points in disjoint sets are independent
  - The interevent times are IID and exponentially distributed
  - A special case of renewal processes
- A inhomogeneous Poisson process is defined by
$$\lambda^*(t) = \mu(t)$$
  - The numbers of points in disjoint sets are independent
  - Not necessarily a renewal process

Hawkes process

- A Hawkes process is defined by

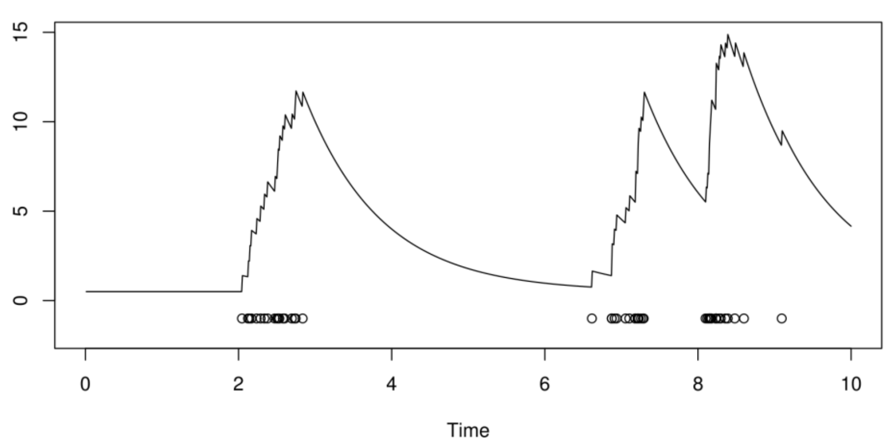
$$\lambda^*(t) = \mu + \alpha \sum_{t_n<t} \exp(-(t - t_n))$$

  - Has a baseline rate of μ
  - A new point increases the rate temporarily by α, which gradually decays
  - Self-exciting or clustering effects
- A Hawkes process can be generalized to

$$\lambda^*(t) = \mu(t) + \alpha \sum_{t_n<t} \gamma(t - t_n; \beta)$$

  - $\gamma(t; \beta)$ is a density on $(0, \infty)$

Hawkes process



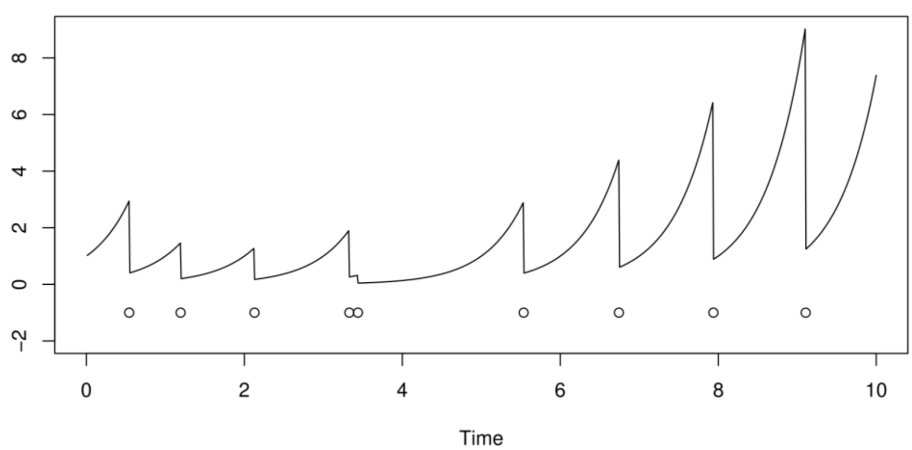
Self-correcting process

- A self-correcting process is defined by

$$\lambda^*(t) = \exp\left(\mu t - \sum_{t_n < t} \alpha\right)$$

  - Baseline intensity keeps increasing over time
  - A new point decreases the rate by a ratio of $\exp(-\alpha)$
  - Point patterns tend to be regular, not clustered as in Hawkes processes

Self-correcting process



From CIF to distribution functions

- Let t_n be the last point before t. Recall

$$\lambda^*(t) = \frac{f(t|H_{t_n})}{1 - F(t|H_{t_n})}$$

- Then

$$F(t|H_{t_n}) = 1 - \exp\left(-\int_{t_n}^{t} \lambda^*(u)du\right)$$

$$f(t|H_{t_n}) = \lambda^*(t)\exp\left(-\int_{t_n}^{t} \lambda^*(u)du\right)$$

Terminating point process

- Typically, we assume next point will eventually come

$$\lim_{t\to\infty} F(t|H_{t_n}) = 1$$

- But we can relax this assumption
- Allow the process to terminate with no more points after some point

$$\lim_{t\to\infty} F(t|H_{t_n}) < 1$$

Terminating point process

- Define a unit-rate point process terminating after $t = 1$

$$\lambda^*(t) = \mathbb{I}(t \in [0,1])$$

- Then

$$F(t|H_{t_n}) = 1 - \exp(-(\min\{t, 1\} - t_n))$$

Terminating point process

- Define a unit-rate point process terminating after getting m points

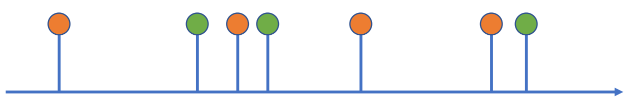
$$\lambda^*(t) = \mathbb{I}(N([0,t)) < m)$$

- Then

$$F(t|H_{t_n}) = (1 - \exp(-(t - t_n)))\mathbb{I}(n < m)$$

Marked event time series

- Sometimes, our data contain not only events t_n
- But also values v_n associated with events
- Examples
  - Earthquakes: time + magnitude
  - Accidents: time + type of injury
- Call these values marks



Marked point process

- Treat the values as marks $v_n \in \mathbb{M}$, where $\mathbb{M} \subseteq \mathbb{R}$ or $\mathbb{M} \subseteq \mathbb{N}$
- Extend the original CIF

$$\lambda^*(t) = \frac{f\left(t \middle| H_{t_n}\right)}{1 - F\left(t \middle| H_{t_n}\right)}$$

- to

$$\lambda^*(t, v) = \lambda^*(t) f^*(v|t) = \frac{f\left(t, v \middle| H_{t_n}\right)}{1 - F\left(t \middle| H_{t_n}\right)}$$

- $f^*(v|t) = f\left(v \middle| t, H_{t_n}\right)$ is the conditional density of the mark
- $f\left(t, v \middle| H_{t_n}\right) = f\left(t \middle| H_{t_n}\right) f^*(v|t)$ is the joint density of time and mark

Marked point process

- If the marks are discrete
$$\lambda^*(t, v)dt = E[N(dt \times v)|H_t]$$
- $N(dt \times v)$ is the number of events in the small time interval dt with the mark v
- If the marks are continuous
$$\lambda^*(t, v)dtdv = E[N(dt \times dv)|H_t]$$
- $N(dt \times dv)$ is the number of events in the small time interval dt with marks in the small interval dv

Marked Hawkes process

- For modeling earthquakes with times and magnitudes
- Assume the magnitudes are in $[0, \infty)$
- Define a marked Hawkes process
$$\lambda^*(t, v) = \left(\mu + \alpha \sum_{t_n < t} e^{\beta v_n} e^{-\gamma(t - t_n)}\right) \delta e^{-\delta v}$$
- New points increase the intensity by $\alpha e^{\beta v_n}$
  - Large earthquakes increase intensity more than small ones

Likelihood function

- Given events $\boldsymbol{t} = (t_1, t_2, \dots, t_N)$ observed in a time interval $[0, T)$

$$p(\boldsymbol{t}) = \left(\prod_{n=1}^{N} f\left(t_n \middle| H_{t_{n-1}}\right)\right)\left(1 - F\left(T \middle| H_{t_N}\right)\right)$$
$$= \left(\prod_{n=1}^{N} \lambda^*(t_n)\right) \exp\left(-\int_0^T \lambda^*(u)\, du\right)$$

Likelihood function

- Given events $\boldsymbol{t} = (t_1, t_2, \dots, t_N)$ observed in a time interval $[0, T)$
- If we have marks $\boldsymbol{v} = (v_1, v_2, \dots, v_N)$ associated with $\boldsymbol{t}$

$$p(\boldsymbol{t}, \boldsymbol{v}) = \left(\prod_{n=1}^{N} f\left(t_n, v_n \middle| H_{t_{n-1}}\right)\right)\left(1 - F\left(T \middle| H_{t_N}\right)\right)$$
$$= \left(\prod_{n=1}^{N} \lambda^*(t_n, v_n)\right) \exp\left(-\int_0^T \lambda^*(u)\, du\right)$$

Maximum likelihood estimate (MLE)

- For a homogeneous Poisson process $\lambda^*(t) = \mu$
- MLE can be found analytically

$$\hat{\mu} = \frac{N}{T}$$

- In general, we can use numerical methods to find MLE

Time-rescaling theorem

- Let $0 < t_1 < t_2 < \cdots$ be a point process with an integrable CIF $\lambda^*(t)$
- Define $\Lambda^*(t) = \int_0^t \lambda^*(u)du$
- Then $\Lambda^*(t_1), \Lambda^*(t_2)$, … form a unit-rate Poisson process

Model checking

- Given data $\{t_n\}_{n=1}^{N}$
- To check whether a point process with a CIF $\lambda^*(t)$ fits the data
- We check whether $\{\Lambda^*(t_n) - \Lambda^*(t_{n-1})\}_{n=1}^{N}$ can be fit by $Exp(1)$

Sampling from a point process

- Inverse method
- Ogata's modified thinning algorithm

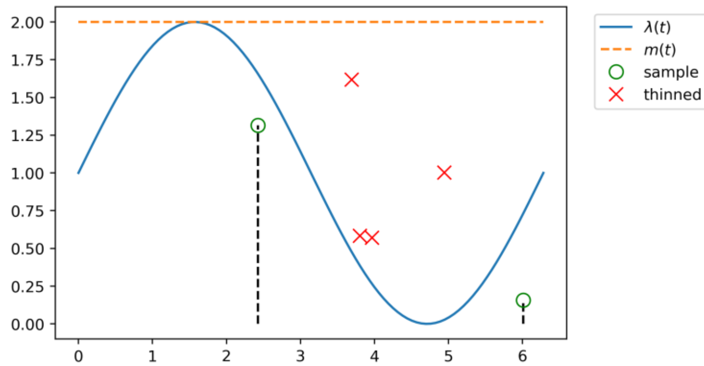
Inverse method

- Define $\Lambda^*(t) = \int_0^t \lambda^*(u)du$
- Set $n = 1, s_0 = 0$
- Repeat
  - Sample $u_n \sim Exp(1)$
  - Set $s_n = s_{n-1} + u_n$
  - Calculate $t_n = \Lambda^{*-1}(s_n)$
  - Set $n = n + 1$

Ogata's modified thinning algorithm

- Define $m(t) \geq \sup_{t<u<\infty} \lambda^*(u)$
- Set $n = 0, t = 0$
- Repeat
  - Sample $s \sim Exp\big(m(t)\big), u \sim Unif([0,1])$
  - If $u \leq \frac{\lambda^*(t+s)}{m(t)}$, set $n = n + 1, t_n = t + s$
  - Set $t = t + s$

Example: thinning



Thank you

Q & A

References

- Rasmussen and Williams, *Gaussian Processes for Machine Learning*.
- Roberts et al., “Gaussian Processes for Time-Series Modelling.”
- Damianou, Titsias, and Lawrence, “Variational Gaussian Process Dynamical Systems.”
- Rasmussen, “Temporal Point Processes the Conditional Intensity Function.”