Machine Learning with Signal Processing

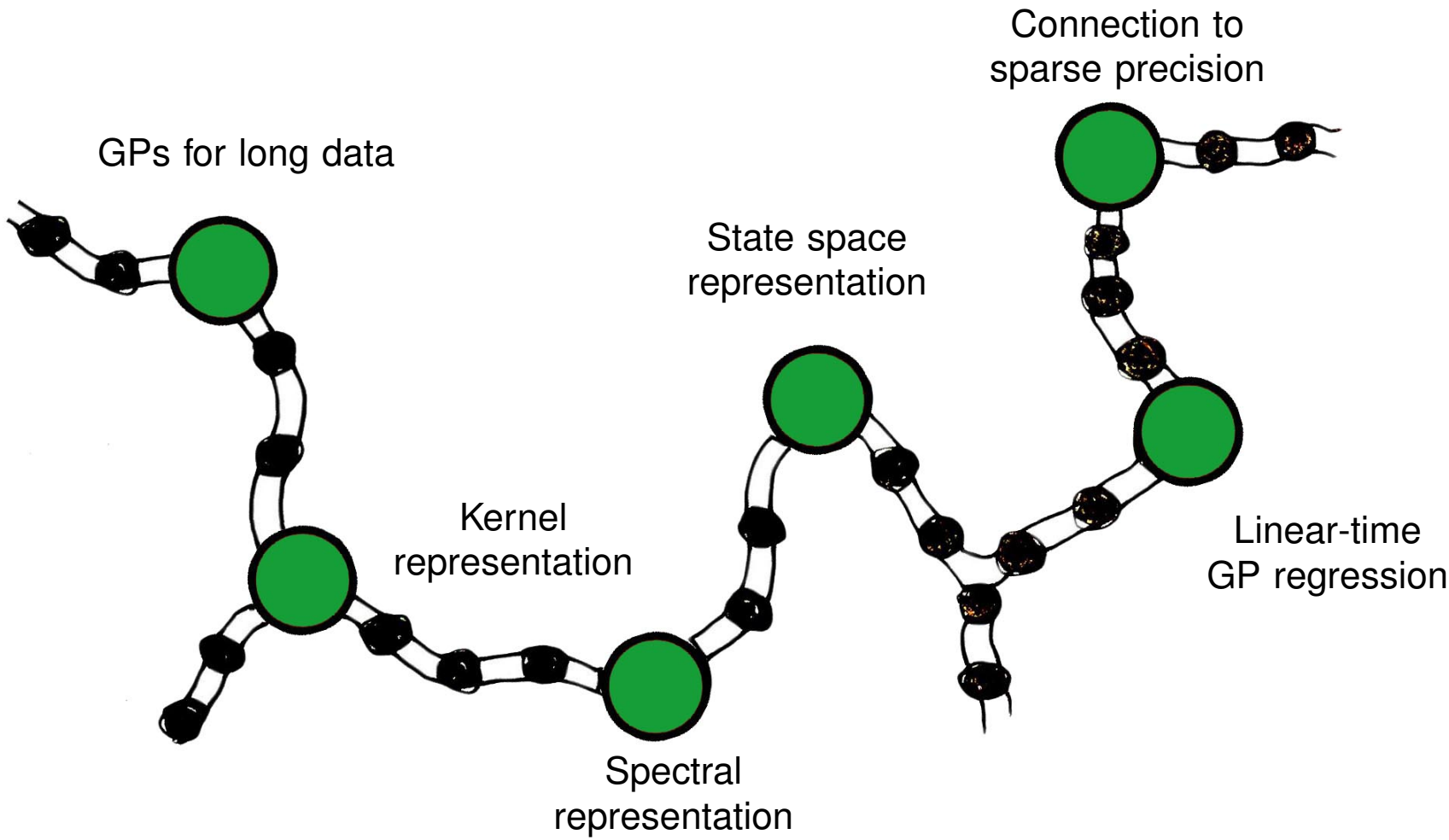
Part III: Three Views into Gaussian Processes

Arno Solin
Assistant Professor in Machine Learning
Department of Computer Science
Aalto University

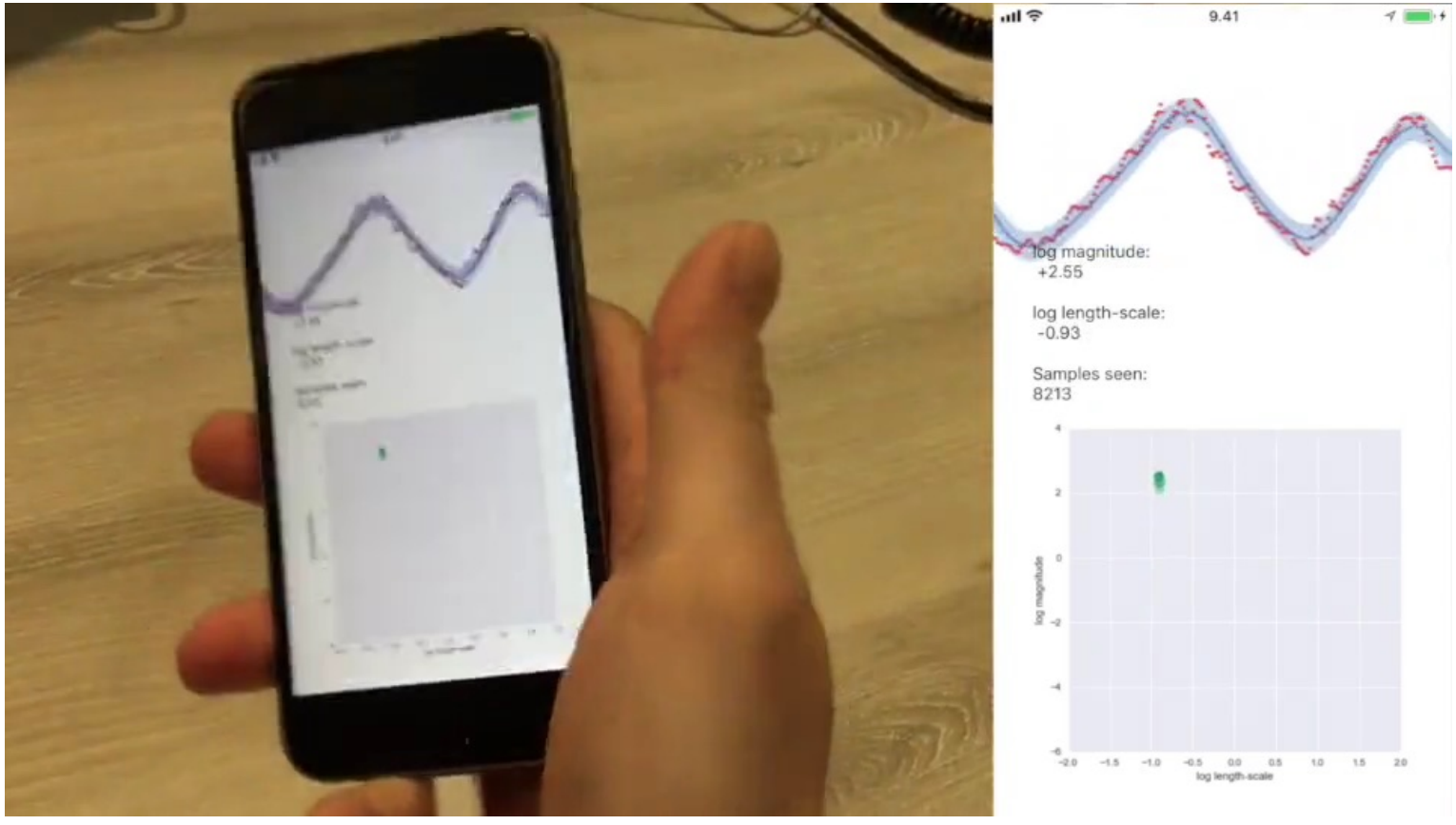
ICML 2020 TUTORIAL

@arnosolin arno.solin.fi

Outline

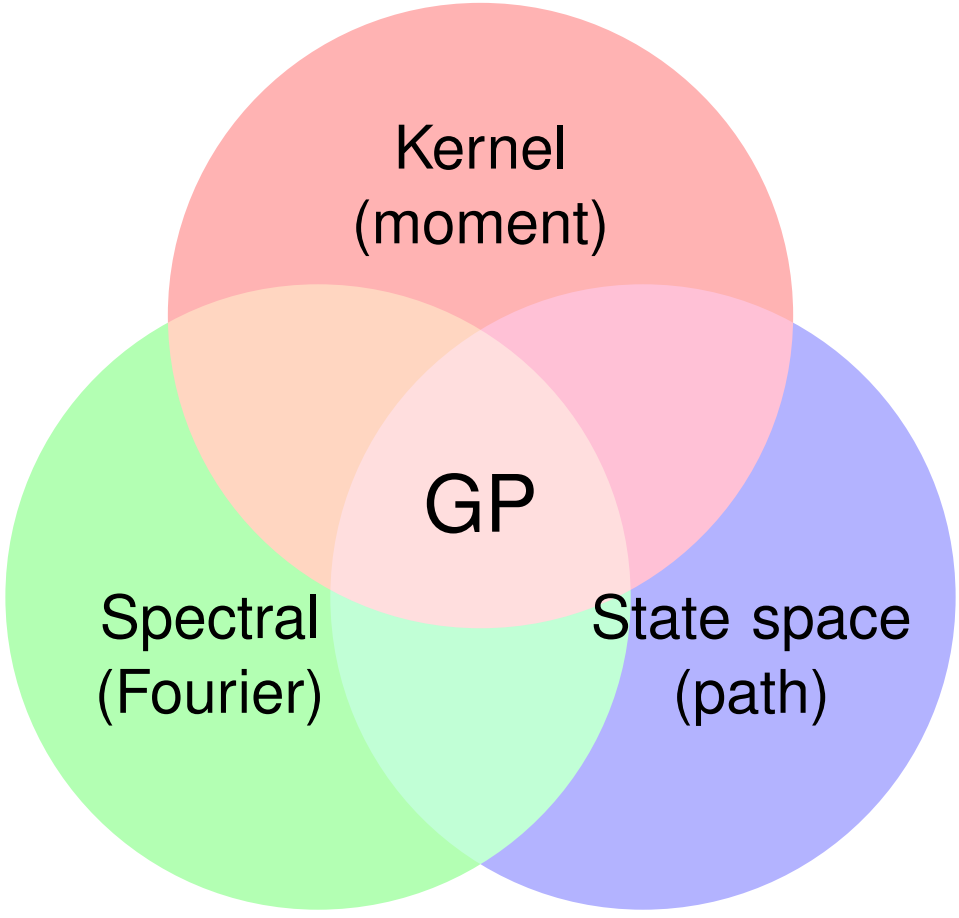

Connection to
sparse precision
GPs for long data
State space
representation
Kernel
representation
Linear-time
GP regression
Spectral
representation

GPs for long (unbounded) data



see Solin *et al.* NeurIPS 2019. Video: `https://youtu.be/myCvUT3XGPc`

Three views into GPs



Kernel (moment) representation

$$f(t) \sim \mathrm{GP}(\mu(t), \kappa(t, t')) \qquad \textit{GP prior}$$
$$\mathbf{y} \mid \mathbf{f} \sim \prod_i p(y_i \mid f(t_i)) \qquad \textit{likelihood}$$

- Let's focus on the GP prior only.
- A temporal Gaussian process (GP) is a random function $f(t)$, such that joint distribution of $f(t_1), \ldots, f(t_n)$ is always Gaussian.
- Mean and covariance functions have the form:

$$\mu(t) = \mathbb{E}[f(t)],$$
$$\kappa(t, t') = \mathbb{E}[(f(t) - \mu(t))(f(t') - \mu(t'))^{\mathsf{T}}].$$

- Convenient for model specification, but expanding the kernel to a covariance matrix can be problematic (the notorious $\mathcal{O}(n^3)$ scaling).

Spectral (Fourier) representation

- The Fourier transform of a function $f(t) : \mathbb{R} \to \mathbb{R}$ is

$$\mathcal{F}[f](\mathrm{i}\,\omega) = \int_{\mathbb{R}} f(t)\, \exp(-\mathrm{i}\,\omega\, t)\, \mathrm{d}t$$

- For a stationary GP, the covariance function can be written in terms of the difference between two inputs:

$$\kappa(t, t') \triangleq \kappa(t - t')$$

- Wiener–Khinchin: If $f(t)$ is a stationary Gaussian process with covariance function $\kappa(t)$, then its spectral density is $S(\omega) = \mathcal{F}[\kappa]$.
- Spectral representation of a GP in terms of spectral density function

$$S(\omega) = \mathbb{E}[\tilde{f}(\mathrm{i}\,\omega)\, \tilde{f}^{\mathsf{T}}(-\mathrm{i}\,\omega)]$$

State space (path) representation [1/3]

- Path or state space representation as solution to a linear time-invariant (LTI) stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{f} = \mathbf{F}\,\mathbf{f}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta},$$

  where $\mathbf{f} = (f, \mathrm{d}f/\mathrm{d}t, \ldots)$ and $\boldsymbol{\beta}(t)$ is a vector of Wiener processes.

- Equivalently, but more informally

$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{f}(t) + \mathbf{L}\,\mathbf{w}(t),$$

  where $\mathbf{w}(t)$ is white noise.

- The model now consists of a drift matrix $\mathbf{F} \in \mathbb{R}^{m \times m}$, a diffusion matrix $\mathbf{L} \in \mathbb{R}^{m \times s}$, and the spectral density matrix of the white noise process $\mathbf{Q}_c \in \mathbb{R}^{s \times s}$.

- The scalar-valued GP can be recovered by $f(t) = \mathbf{h}^\mathsf{T}\,\mathbf{f}(t)$.

State space (path) representation [2/3]

- The initial state is given by a stationary state $\mathbf{f}(0) \sim \mathrm{N}(\mathbf{0}, \mathbf{P}_\infty)$ which fulfils

$$\mathbf{F}\,\mathbf{P}_\infty + \mathbf{P}_\infty\,\mathbf{F}^\mathsf{T} + \mathbf{L}\,\mathbf{Q}_\mathrm{c}\,\mathbf{L}^\mathsf{T} = \mathbf{0}$$

- The covariance function at the stationary state can be recovered by

$$\kappa(t, t') = \begin{cases} \mathbf{h}^\mathsf{T}\mathbf{P}_\infty\ \exp((t'-t)\mathbf{F})^\mathsf{T}\mathbf{h}, & t' \geq t \\ \mathbf{h}^\mathsf{T}\exp((t'-t)\mathbf{F})\,\mathbf{P}_\infty\,\mathbf{h}, & t' < t \end{cases}$$

  where $\exp(\cdot)$ denotes the matrix exponential function.

- The spectral density function at the stationary state can be recovered by

$$S(\omega) = \mathbf{h}^\mathsf{T}(\mathbf{F} + \mathrm{i}\,\omega\,\mathbf{I})^{-1}\,\mathbf{L}\,\mathbf{Q}_\mathrm{c}\,\mathbf{L}^\mathsf{T}\,(\mathbf{F} - \mathrm{i}\,\omega\,\mathbf{I})^{-\mathsf{T}}\mathbf{h}$$

State space (path) representation [3/3]

- Similarly as the kernel has to be evaluated into a covariance matrix for computations, the SDE can be solved for discrete time points $\{t_i\}_{i=1}^n$.
- The resulting model is a discrete state space model:

$$\mathbf{f}_i = \mathbf{A}_{i-1}\,\mathbf{f}_{i-1} + \mathbf{q}_{i-1}, \quad \mathbf{q}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_i),$$

  where $\mathbf{f}_i = \mathbf{f}(t_i)$.
- The discrete-time model matrices are given by:

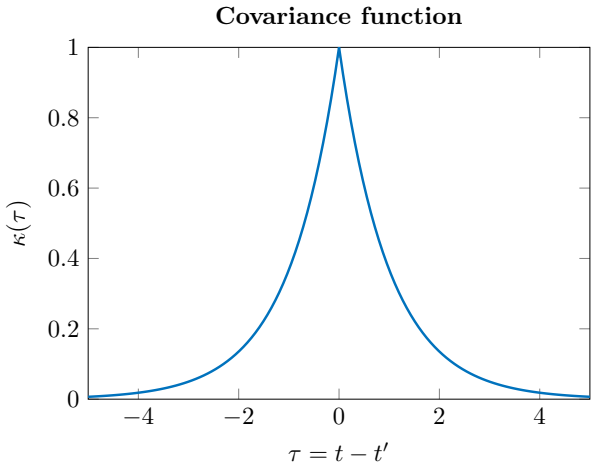
$$\mathbf{A}_i = \exp(\mathbf{F}\,\Delta t_i),$$

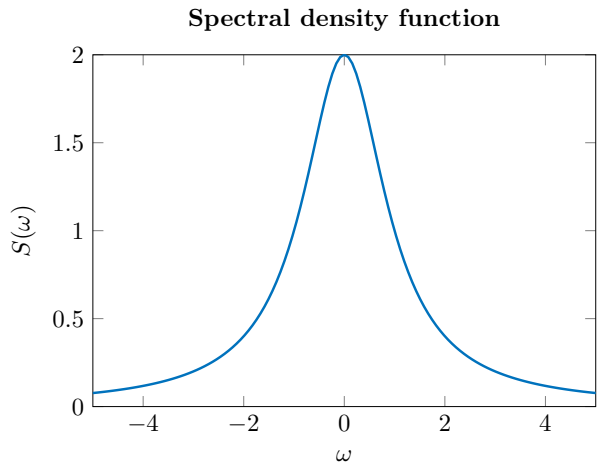
$$\mathbf{Q}_i = \int_0^{\Delta t_i} \exp(\mathbf{F}\,(\Delta t_i - \tau))\,\mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^{\mathsf{T}}\,\exp(\mathbf{F}\,(\Delta t_i - \tau))^{\mathsf{T}}\,\mathrm{d}\tau,$$

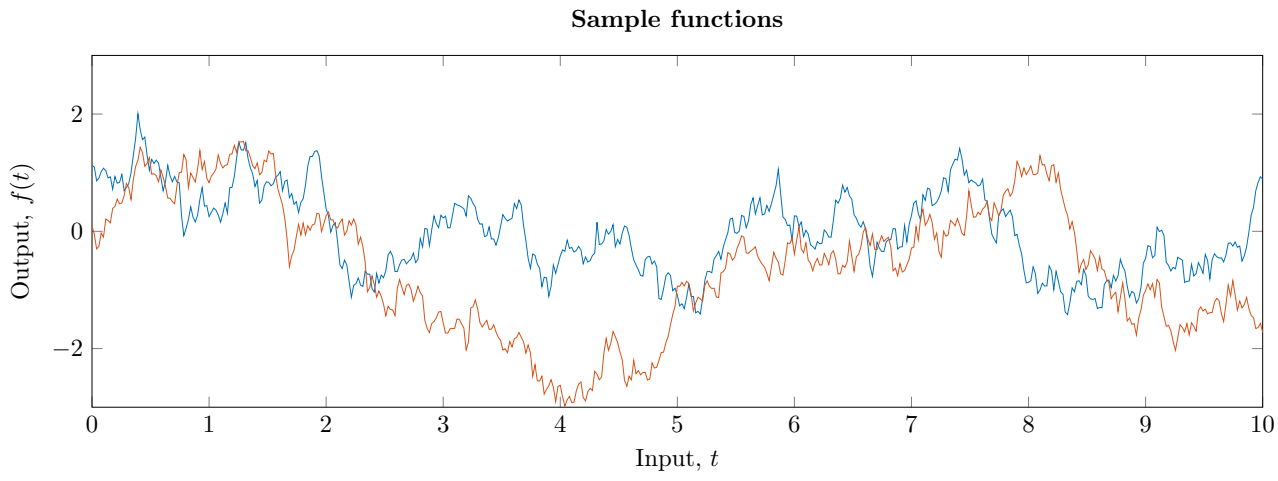
  where $\Delta t_i = t_{i+1} - t_i$
- If the model is stationary, $\mathbf{Q}_i$ is given by

$$\mathbf{Q}_i = \mathbf{P}_\infty - \mathbf{A}_i\,\mathbf{P}_\infty\,\mathbf{A}_i^{\mathsf{T}}$$

Three views into GPs







Example: Exponential covariance function

- Exponential covariance function (Ornstein-Uhlenbeck process):

$$\kappa(t, t') = \exp(-\lambda |t - t'|)$$

- Spectral density function:

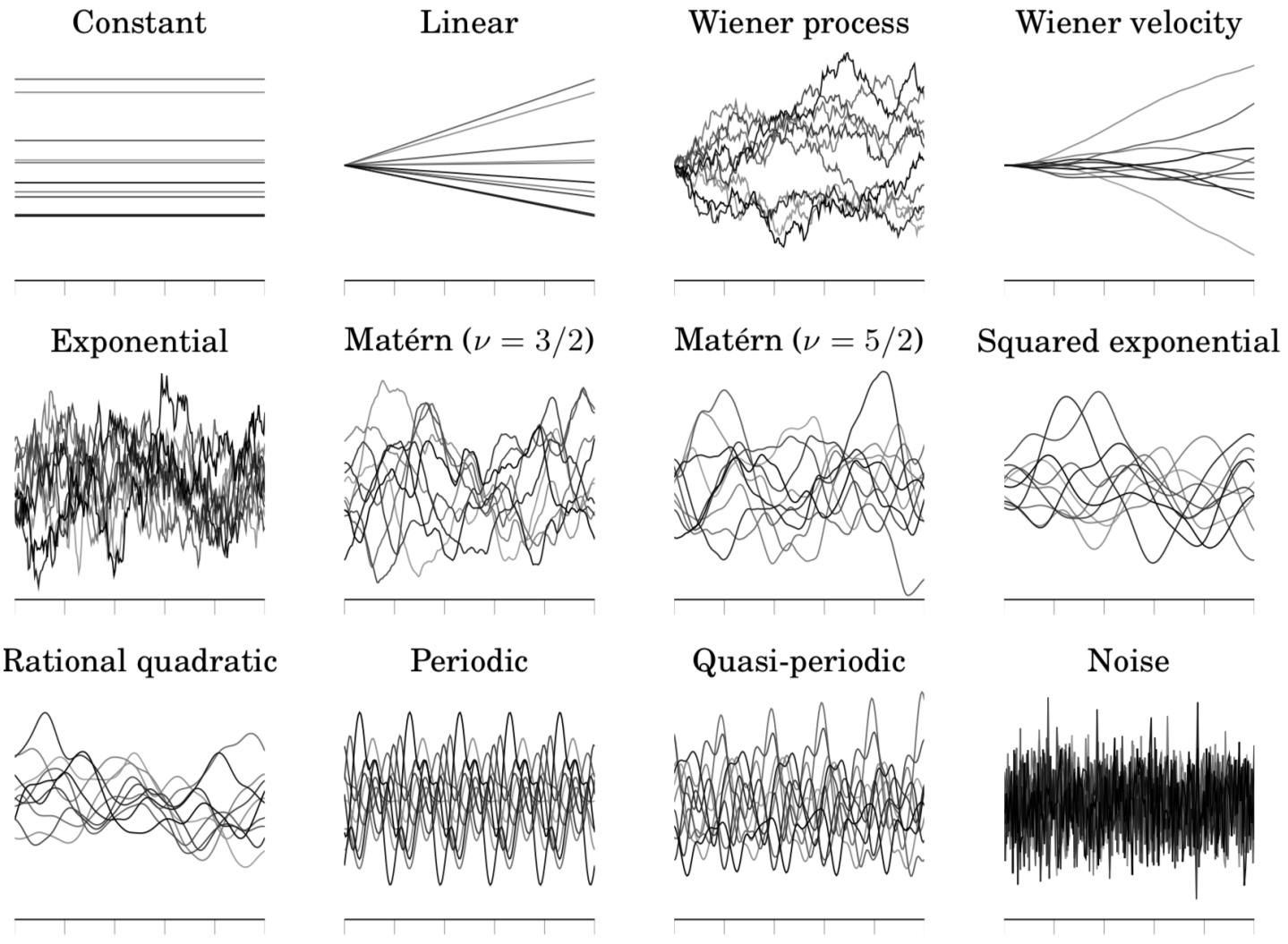
$$S(\omega) = \frac{2}{\lambda + \omega^2/\lambda}$$

- Path representation: Stochastic differential equation (SDE)

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = -\lambda f(t) + w(t),$$

or using the notation from before:
$F = -\lambda$, $L = 1$, $Q_c = 2$, $h = 1$, and $P_\infty = 1$.

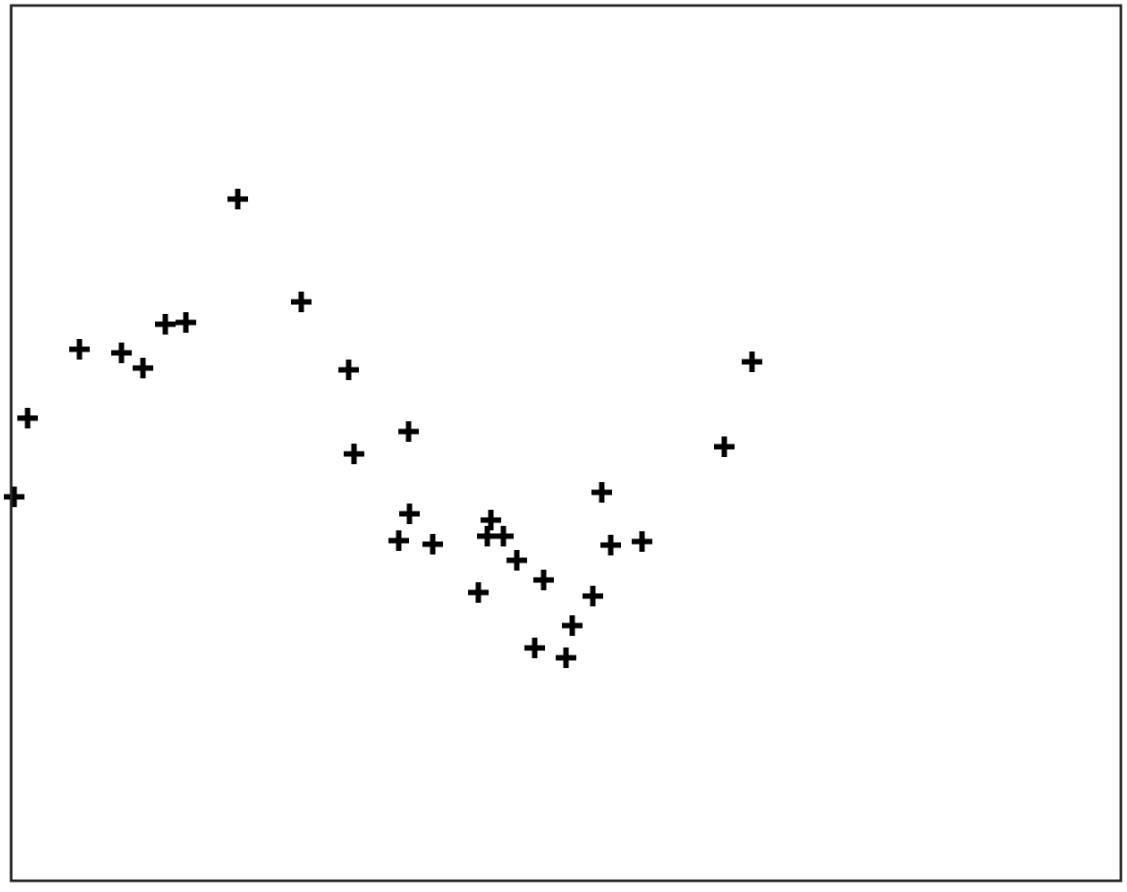
Examples of applicable GP priors



Applicable GP priors

- The covariance function needs to be Markovian (or approximated as such).
- Covers many common stationary and non-stationary models.
- Sums of kernels: $\kappa(t, t') = \kappa_1(t, t') + \kappa_2(t, t')$
  - Stacking of the state spaces
  - State dimension: $m = m_1 + m_2$
- Product of kernels: $\kappa(t, t') = \kappa_1(t, t')\, \kappa_2(t, t')$
  - Kronecker sum of the models
  - State dimension: $m = m_1\, m_2$

Example: GP regression, $\mathcal{O}(n^3)$



Example: GP regression, $\mathcal{O}(n^3)$

- Consider the GP regression problem with input–output training pairs $\{(t_i, y_i)\}_{i=1}^n$:

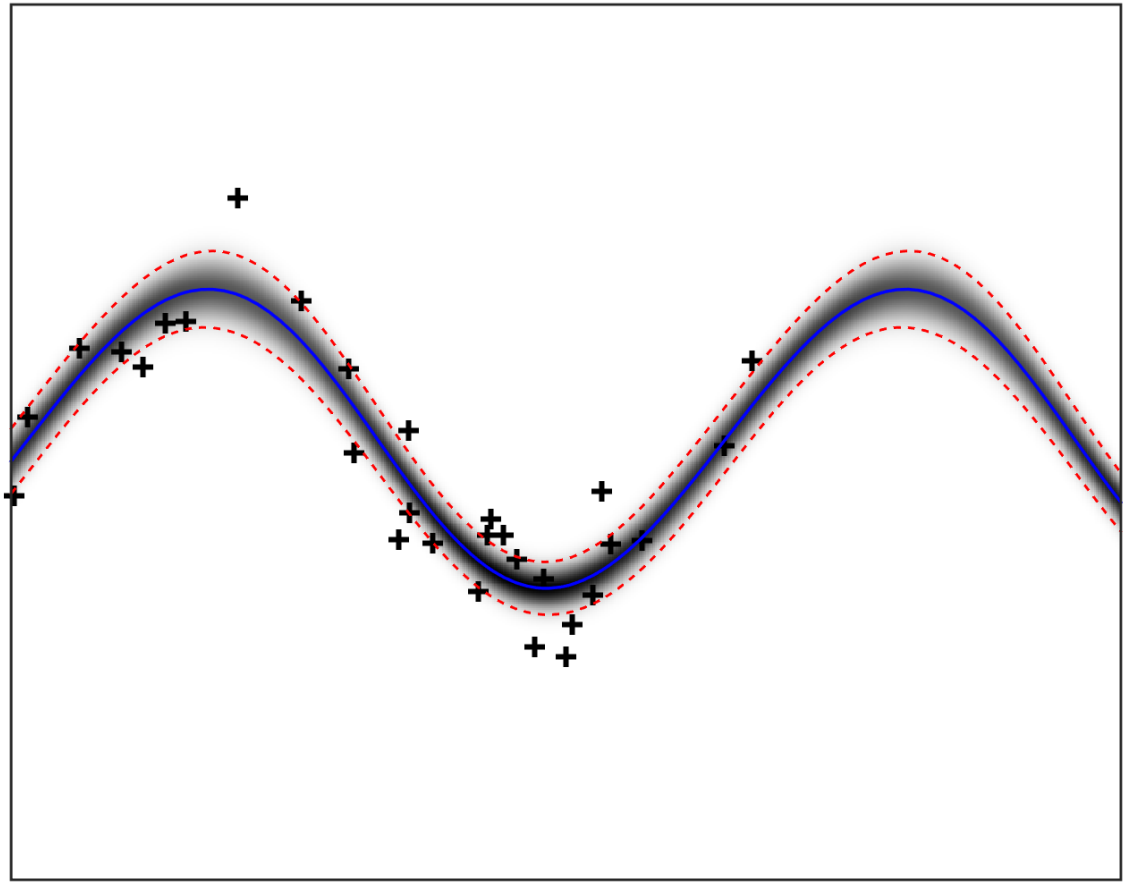
$$f(t) \sim \mathrm{GP}(0, \kappa(t, t')),$$
$$y_i = f(t_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_{\mathrm{n}}^2)$$

- The posterior mean and variance for an unseen test input t_* is given by (see previous lectures):

$$\mathbb{E}[f_*] = \mathbf{k}_* \left(\mathbf{K} + \sigma_{\mathrm{n}}^2 \mathbf{I}\right)^{-1} \mathbf{y},$$
$$\mathbb{V}[f_*] = \mathbf{K}_{**} - \mathbf{k}_* \left(\mathbf{K} + \sigma_{\mathrm{n}}^2 \mathbf{I}\right)^{-1} \mathbf{k}_*^{\mathsf{T}}$$

- Note the inversion of the $n \times n$ matrix.

Example: GP regression, $\mathcal{O}(n^3)$



Example: GP regression, $\mathcal{O}(n)$

- The sequential solution (goes under the name 'Kalman filter') considers one data point at a time, hence the linear time-scaling.
- Start from $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \mathbf{P}_\infty$ and for each data point iterate the following steps.
- Kalman prediction:

$$\mathbf{m}_{i|i-1} = \mathbf{A}_{i-1}\,\mathbf{m}_{i-1|i-1},$$
$$\mathbf{P}_{i|i-1} = \mathbf{A}_{i-1}\,\mathbf{P}_{i-1|i-1}\,\mathbf{A}_{i-1}^{\mathsf{T}} + \mathbf{Q}_{i-1}.$$

- Kalman update:

$$v_i = y_i - \mathbf{h}^{\mathsf{T}}\mathbf{m}_{i|i-1},$$
$$S_i = \mathbf{h}^{\mathsf{T}}\mathbf{P}_{i|i-1}\,\mathbf{h} + \sigma_{\mathrm{n}}^2,$$
$$\mathbf{K}_i = \mathbf{P}_{i|i-1}\,\mathbf{h}\,S_i^{-1},$$
$$\mathbf{m}_{i|i} = \mathbf{m}_{i|i-1} + \mathbf{K}_i\,v_i,$$
$$\mathbf{P}_{i|i} = \mathbf{P}_{i|i-1} - \mathbf{K}_i\,S_i\,\mathbf{K}_i^{\mathsf{T}}.$$

Example: GP regression, $\mathcal{O}(n)$

- To condition all time-marginals on all data, run a backward sweep (Rauch–Tung–Striebel smoother):

$$\begin{aligned}
\mathbf{m}_{i+1|i} &= \mathbf{A}_i \, \mathbf{m}_{i|i}, \\
\mathbf{P}_{i+1|i} &= \mathbf{A}_i \, \mathbf{P}_{i|i} \, \mathbf{A}_i^{\mathsf{T}} + \mathbf{Q}_i, \\
\mathbf{G}_i &= \mathbf{P}_{i|i} \, \mathbf{A}_i^{\mathsf{T}} \, \mathbf{P}_{i+1|i}^{-1}, \\
\mathbf{m}_{i|n} &= \mathbf{m}_{i|i} + \mathbf{G}_i \, (\mathbf{m}_{i+1|n} - \mathbf{m}_{i+1|i}), \\
\mathbf{P}_{i|n} &= \mathbf{P}_{i|i} + \mathbf{G}_i \, (\mathbf{P}_{i+1|n} - \mathbf{P}_{i+1|i}) \, \mathbf{G}_i^{\mathsf{T}},
\end{aligned}$$

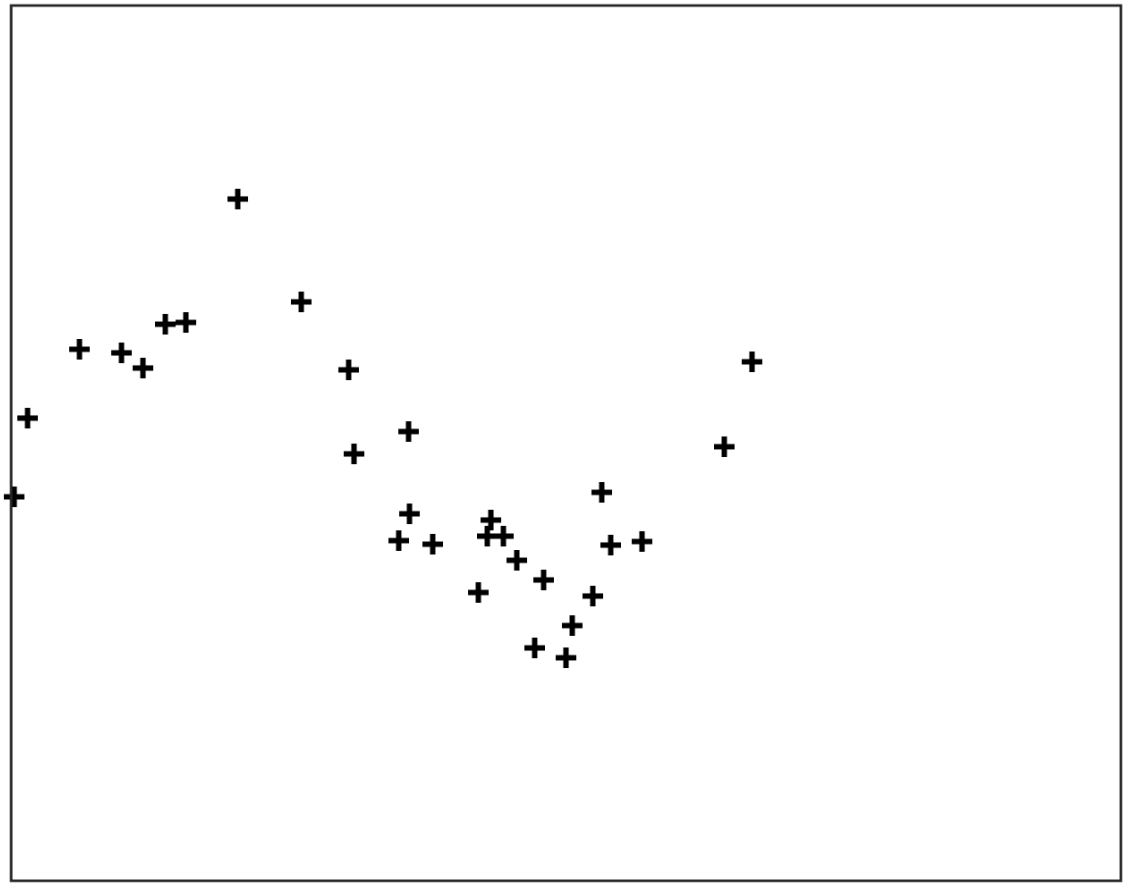
- The marginal mean and variance can be recovered by:

$$\begin{aligned}
\mathbb{E}[f_i] &= \mathbf{h}^{\mathsf{T}} \mathbf{m}_{i|n}, \\
\mathbb{V}[f_i] &= \mathbf{h}^{\mathsf{T}} \mathbf{P}_{i|n} \, \mathbf{h}
\end{aligned}$$

- The log marginal likelihood can be evaluated as a by-product of the Kalman update:

$$\log p(\mathbf{y}) = -\frac{1}{2} \sum_{i=1}^{n} \log |2\pi \, S_i| + v_i^{\mathsf{T}} \, S_i^{-1} \, v_i$$

Example: GP regression, $\mathcal{O}(n)$



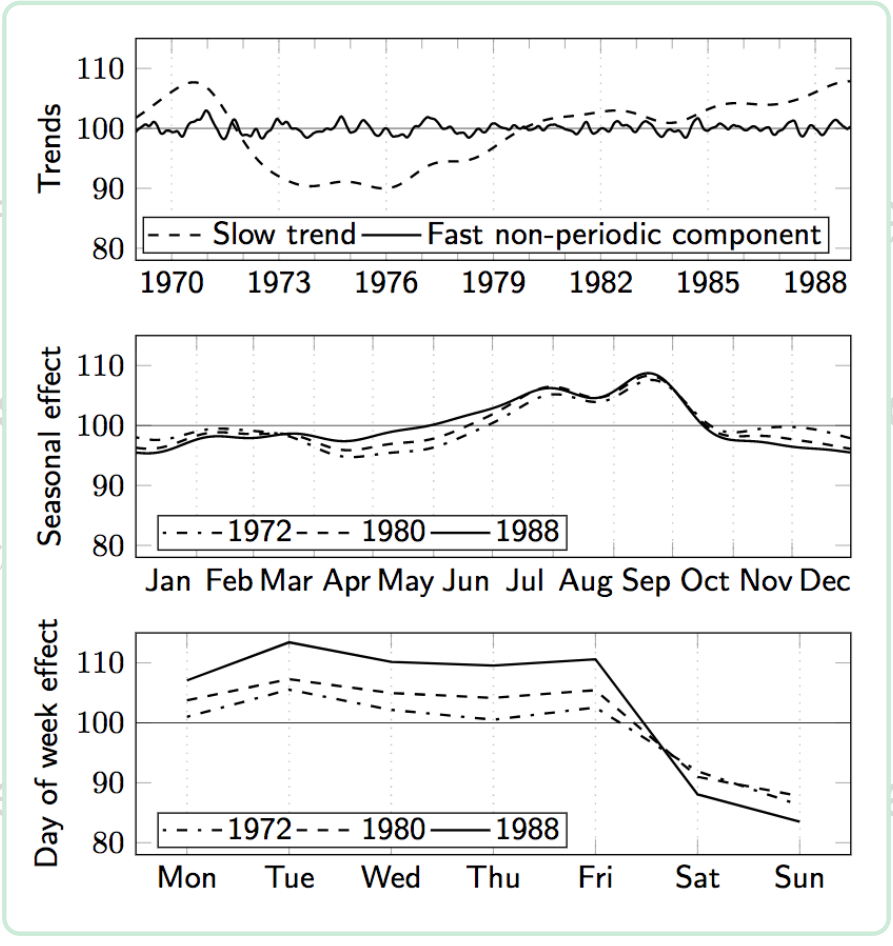
Basic regression example

- Number of births in the US (from BDA3 by Gelman *et al.*)
- Daily data between 1969–1988 ($n = 7305$)
- GP regression with a prior covariance function:

$$\kappa(t,t') = \kappa_{\text{Mat.}}^{\nu=5/2}(t,t') + \kappa_{\text{Mat.}}^{\nu=3/2}(t,t') + \kappa_{\text{Per.}}^{\text{year}}(t,t')\,\kappa_{\text{Mat.}}^{\nu=3/2}(t,t') + \kappa_{\text{Per.}}^{\text{week}}(t,t')\,\kappa_{\text{Mat.}}^{\nu=3/2}(t,t')$$

- Learn hyperparameters by optimizing the marginal likelihood

Basic regression example

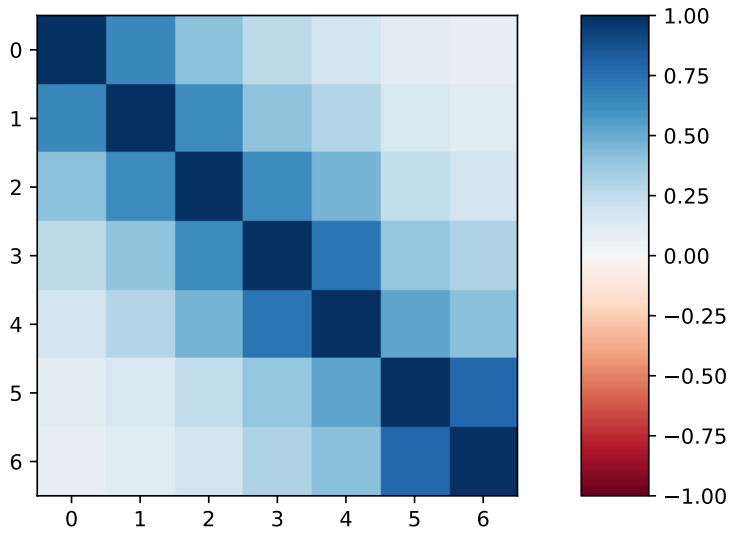


Explaining changes in number of births in the US

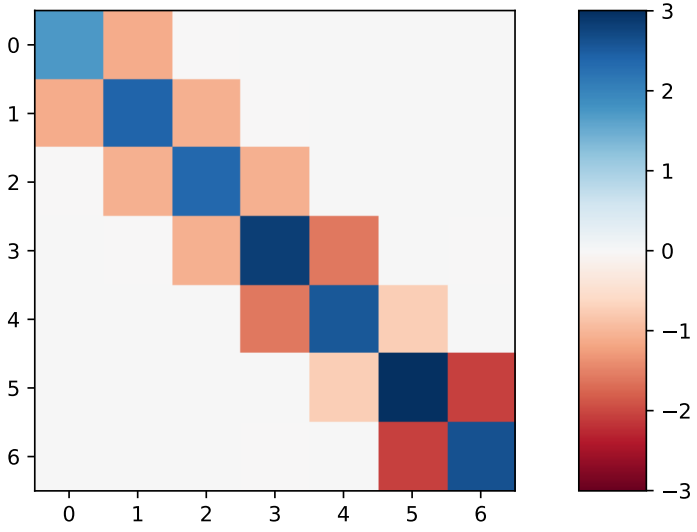
Connection to banded precision matrices

Precision matrices

Covariance (Gram) matrix:
$\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$



Precision matrix:
$\mathbf{K}^{-1}$



For Markovian models the precision is sparse!
(block tri-diagonal)

see Durrande *et al.* AISTATS 2019.

Constructing the precision matrix

- The full precision matrix can be constructed from the state space model matrices:

$$\hat{\mathbf{K}}^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ -\mathbf{A}_1 & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}_2 & \mathbf{I} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & -\mathbf{A}_n & \mathbf{I} \end{pmatrix}^{-\mathsf{T}} \begin{pmatrix} \mathbf{P}_0 & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \mathbf{0} & \mathbf{Q}_2 & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{Q}_n \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ -\mathbf{A}_1 & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}_2 & \mathbf{I} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & -\mathbf{A}_n & \mathbf{I} \end{pmatrix}^{-1}$$

- Discarding the other model states by passing through the measurement model:

$$\mathbf{K}^{-1} = (\mathbf{I}_n \otimes \mathbf{h})\, \hat{\mathbf{K}}^{-1}\, (\mathbf{I}_n \otimes \mathbf{h})^{\mathsf{T}}$$

Summary

- Gaussian processes have different representations:
  • Covariance function • Spectral density • State space
- Temporal (single-input) Gaussian processes
  $\iff$ stochastic differential equations (SDEs)
- Conversions between the representations can make model building easier
- (Exact) inference of the latent functions, can be done in $\mathcal{O}(n)$ time and memory complexity by Kalman filtering

Up next

- Spatio-temporal models
- Non-Gaussian likelihood models
- Hybrid models

Bibliography

These references are sources for finding a more detailed overview on the topics of this part :

- C. E. Rasmussen and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- S. Särkkä and A. Solin (2019). *Applied Stochastic Differential Equations*. Cambridge University Press. Cambridge, UK.