Probabilistic Models for Audio Signals

an intro via time-frequency analysis

William Wilkinson

December 11, 2018

Queen Mary University of London

Probabilistic Modelling / Bayesian Analysis

In any modelling task, our choice of model structure / architecture should encode our knowledge about the world.

Probabilistic Modelling / Bayesian Analysis

In any modelling task, our choice of model structure / architecture should encode our knowledge about the world.

What does it mean to be "Bayesian"?

Probabilistic Modelling / Bayesian Analysis

In any modelling task, our choice of model structure / architecture should encode our knowledge about the world.

What does it mean to be "Bayesian"?

- Place probability distributions over all model components about which we are uncertain.

Probabilistic Modelling / Bayesian Analysis

In any modelling task, our choice of model structure / architecture should encode our knowledge about the world.

What does it mean to be "Bayesian"?

- Place probability distributions over all model components about which we are uncertain.
    - In practice we're uncertain about most things, including the data.

Example: Time-frequency analysis

We want to uncover the **time-varying spectral content** of a signal.

Typically in signal processing we use the STFT or a **filter bank**:

Example: Time-frequency analysis

We want to uncover the **time-varying spectral content** of a signal.

Typically in signal processing we use the STFT or a **filter bank**:

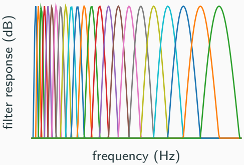
audio signal

Example: Time-frequency analysis

We want to uncover the **time-varying spectral content** of a signal.

Typically in signal processing we use the STFT or a **filter bank**:

audio signal



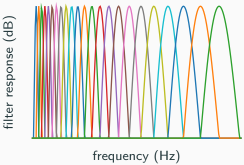
filter bank

Example: Time-frequency analysis

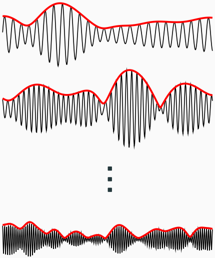
We want to uncover the **time-varying spectral content** of a signal.

Typically in signal processing we use the STFT or a **filter bank**:

audio signal



filter bank



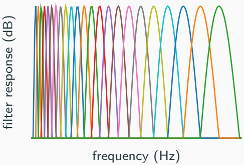
filter outputs

Example: Time-frequency analysis

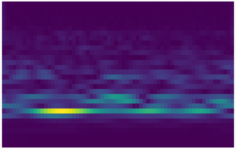
We want to uncover the **time-varying spectral content** of a signal.

Typically in signal processing we use the STFT or a **filter bank**:

audio signal



filter bank



spectrogram

Probabilistic time-frequency analysis

What are we uncertain about in TF analysis?

Probabilistic time-frequency analysis

What are we uncertain about in TF analysis?

There are actually (infinitely) many ways that a given signal can be decomposed into a sum of periodic components.

Probabilistic time-frequency analysis

What are we uncertain about in TF analysis?

There are actually (infinitely) many ways that a given signal can be decomposed into a sum of periodic components.

- **which is the "right" one?**

Probabilistic time-frequency analysis

What are we uncertain about in TF analysis?

There are actually (infinitely) many ways that a given signal can be decomposed into a sum of periodic components.

- **which is the “right” one?**
- **which is the “right” one for your specific task?**

Probabilistic time-frequency analysis

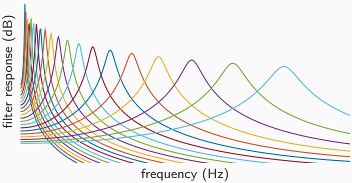
How should we choose the filter bank parameters?

- centre-frequency,
- bandwidth,
- scale

Probabilistic time-frequency analysis

How should we choose the filter bank parameters?

- centre-frequency,
- bandwidth,
- scale

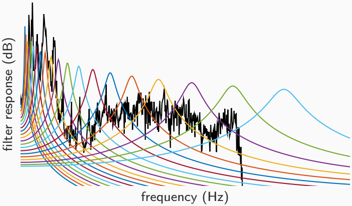


logarithmically spaced

Probabilistic time-frequency analysis

How should we choose the filter bank parameters?

- centre-frequency,
- bandwidth,
- scale

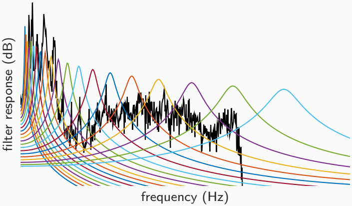


logarithmically spaced

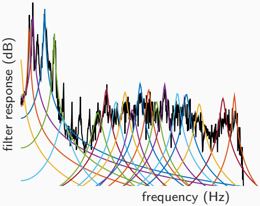
Probabilistic time-frequency analysis

How should we choose the filter bank parameters?

- centre-frequency,
- bandwidth,
- scale



logarithmically spaced

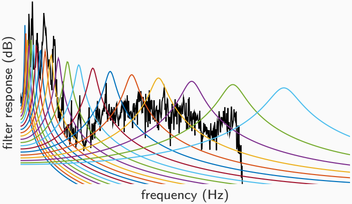


fit to the signal

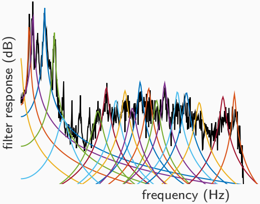
Probabilistic time-frequency analysis

How should we choose the filter bank parameters?

- centre-frequency,
- bandwidth,
- scale



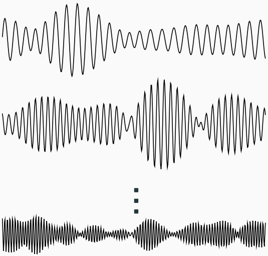
logarithmically spaced



fit to the signal

Benefits include: uncertainty quantification, can adapt to specific tasks, generative (amplitude and phase correlations)

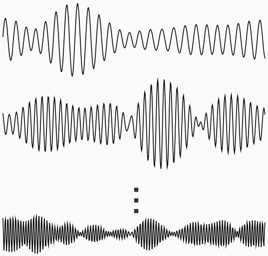
Probabilistic time-frequency analysis



filter outputs

Place a Gaussian distribution over each frequency component.

Probabilistic time-frequency analysis

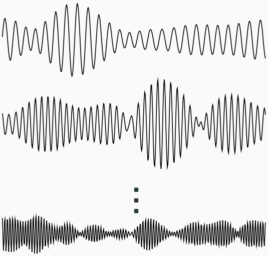


filter outputs

Place a Gaussian distribution over each frequency component.

Integrate over all possible decompositions to find the statistically most likely one given the data.

Probabilistic time-frequency analysis

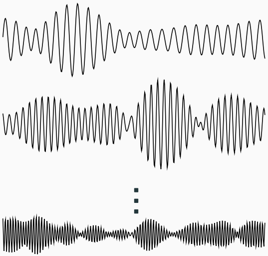


filter outputs

Place a Gaussian distribution over each frequency component. **What does it mean to specify a distribution over temporal data?**

Integrate over all possible decompositions to find the statistically most likely one given the data.

Probabilistic time-frequency analysis



filter outputs

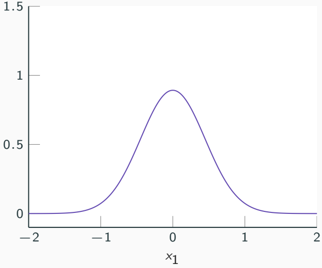
Place a Gaussian distribution over each frequency component. **What does it mean to specify a distribution over temporal data?**

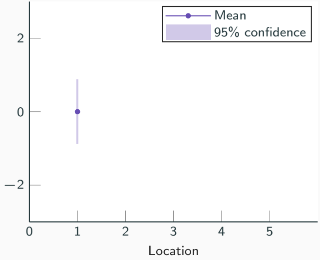
Integrate over all possible decompositions to find the statistically most likely one given the data. **Bayesian analysis provides a principled way to do this without testing every scenario.**

Specifying a distribution over temporal data

1D Gaussian: $x_1 \sim N(\mu, \sigma^2)$

$\mu = 0, \sigma^2 = 0.2$

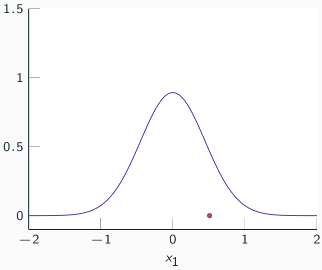


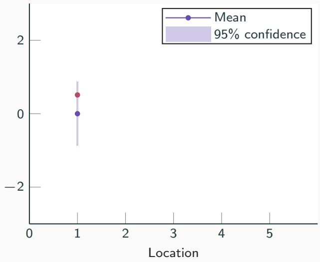


Specifying a distribution over temporal data

1D Gaussian: $x_1 \sim N(\mu, \sigma^2)$

$\mu = 0, \sigma^2 = 0.2$

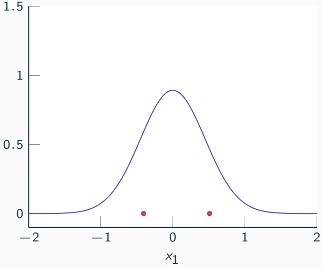


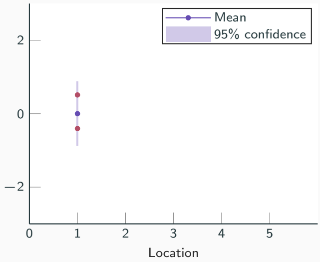


Specifying a distribution over temporal data

1D Gaussian: $x_1 \sim N(\mu, \sigma^2)$

$\mu = 0, \sigma^2 = 0.2$

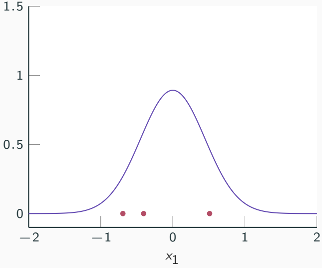


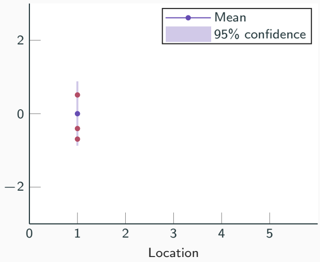


Specifying a distribution over temporal data

1D Gaussian: $x_1 \sim N(\mu, \sigma^2)$

$\mu = 0, \sigma^2 = 0.2$

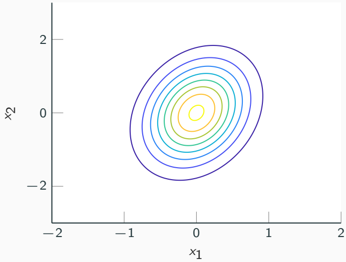


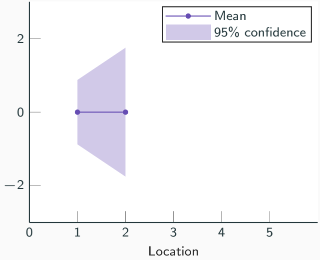


Specifying a distribution over temporal data

2D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.8 \end{pmatrix}$$

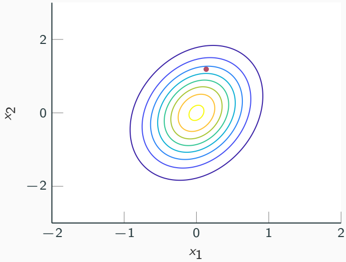


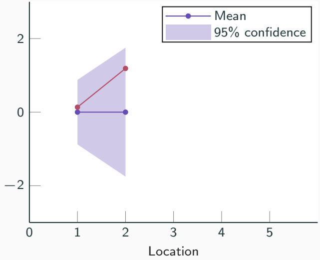


Specifying a distribution over temporal data

2D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.8 \end{pmatrix}$$

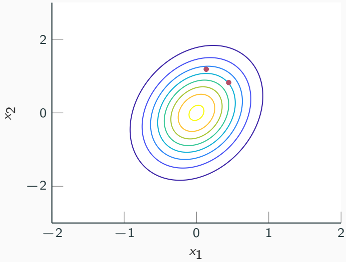


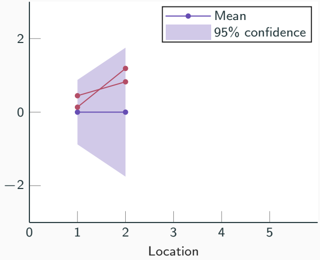


Specifying a distribution over temporal data

2D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.8 \end{pmatrix}$$

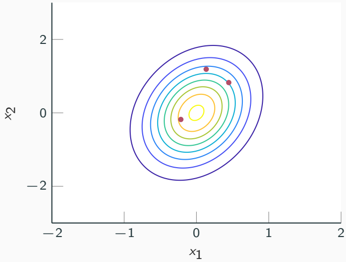


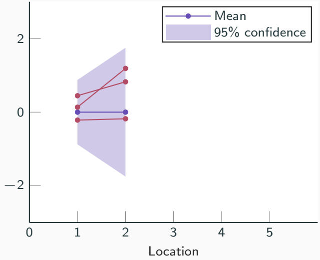


Specifying a distribution over temporal data

2D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.8 \end{pmatrix}$$

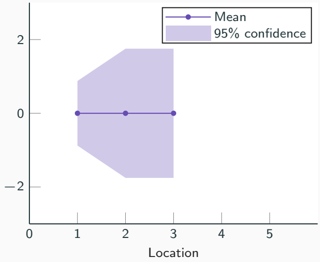




Specifying a distribution over temporal data

3D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

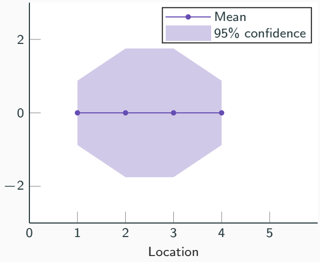
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.4 \\ 0.1 & 0.4 & 0.8 \end{pmatrix}$$



Specifying a distribution over temporal data

4D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

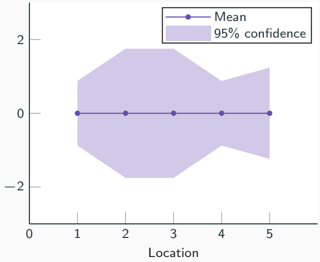
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 & 0.1 & 0.0 \\ 0.1 & 0.8 & 0.4 & 0.1 \\ 0.1 & 0.4 & 0.8 & 0.2 \\ 0.0 & 0.1 & 0.2 & 0.4 \end{pmatrix}$$



Specifying a distribution over temporal data

5D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

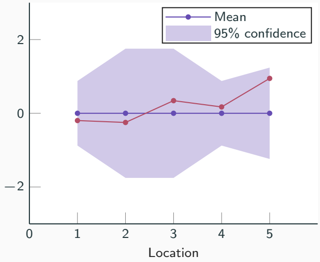
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 & 0.1 & 0.0 & 0.0 \\ 0.1 & 0.8 & 0.4 & 0.1 & 0.0 \\ 0.1 & 0.4 & 0.8 & 0.1 & 0.1 \\ 0.0 & 0.1 & 0.1 & 0.2 & 0.1 \\ 0.0 & 0.0 & 0.1 & 0.1 & 0.4 \end{pmatrix}$$



Specifying a distribution over temporal data

5D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

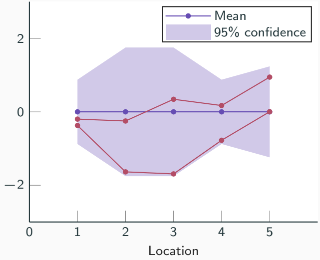
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 & 0.1 & 0.0 & 0.0 \\ 0.1 & 0.8 & 0.4 & 0.1 & 0.0 \\ 0.1 & 0.4 & 0.8 & 0.1 & 0.1 \\ 0.0 & 0.1 & 0.1 & 0.2 & 0.1 \\ 0.0 & 0.0 & 0.1 & 0.1 & 0.4 \end{pmatrix}$$



Specifying a distribution over temporal data

5D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

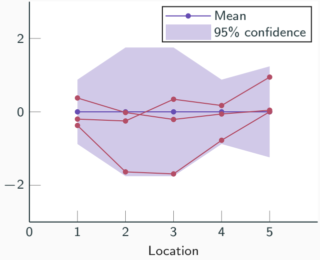
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 & 0.1 & 0.0 & 0.0 \\ 0.1 & 0.8 & 0.4 & 0.1 & 0.0 \\ 0.1 & 0.4 & 0.8 & 0.1 & 0.1 \\ 0.0 & 0.1 & 0.1 & 0.2 & 0.1 \\ 0.0 & 0.0 & 0.1 & 0.1 & 0.4 \end{pmatrix}$$



Specifying a distribution over temporal data

5D Gaussian: $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0.1 & 0.1 & 0.0 & 0.0 \\ 0.1 & 0.8 & 0.4 & 0.1 & 0.0 \\ 0.1 & 0.4 & 0.8 & 0.1 & 0.1 \\ 0.0 & 0.1 & 0.1 & 0.2 & 0.1 \\ 0.0 & 0.0 & 0.1 & 0.1 & 0.4 \end{pmatrix}$$



Distributions over temporal data - Gaussian processes

∞D Gaussian

Any finite set of locations we care to consider will be distributed as:

$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Distributions over temporal data - Gaussian processes

∞D Gaussian
Any finite set of locations we care to consider will be distributed as:
$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

This is called a **Gaussian process**.

Distributions over temporal data - Gaussian processes

∞D Gaussian
Any finite set of locations we care to consider will be distributed as:
$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

This is called a **Gaussian process**.

But how do we choose $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, since we don't know a priori which locations we will be considering?

Distributions over temporal data - Gaussian processes

∞D Gaussian

Any finite set of locations we care to consider will be distributed as:
$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

This is called a **Gaussian process**.

We must define a *mean function* $\boldsymbol{\mu}(t)$ and *covariance function* $\boldsymbol{\Sigma}(t, t')$.

Distributions over temporal data - Gaussian processes

∞D Gaussian
Any finite set of locations we care to consider will be distributed as:
$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

This is called a **Gaussian process**.

We must define a *mean function* $\boldsymbol{\mu}(t)$ and *covariance function* $\boldsymbol{\Sigma}(t, t')$.

Notation:
$\boldsymbol{x}(t) \sim GP(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t, t'))$

The exponential covariance function

The mean and covariance functions encode our prior knowledge.
One common choice is:

$\boldsymbol{\mu}(t) = \mathbf{0}$

$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|/\ell)$

$\sigma^2 = 1, \ell = 10$

The exponential covariance function

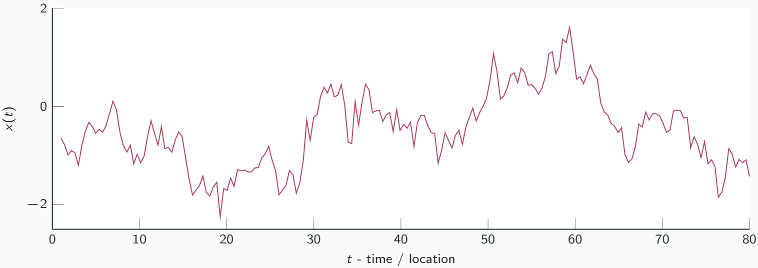
The mean and covariance functions encode our prior knowledge.

One common choice is:

$\boldsymbol{\mu}(t) = \mathbf{0}$

$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|/\ell)$

$\sigma^2 = 1, \ell = 10$



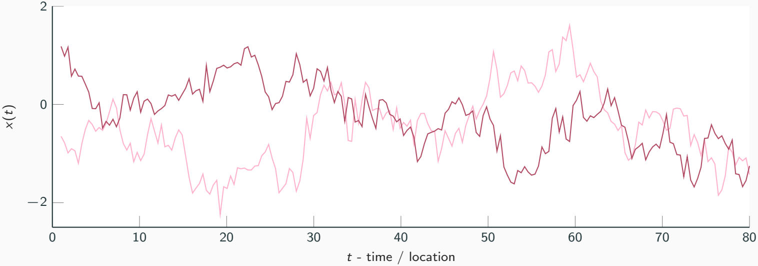
The exponential covariance function

The mean and covariance functions encode our prior knowledge.
One common choice is:

$$\boldsymbol{\mu}(t) = \mathbf{0}$$

$$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|/\ell)$$

$$\sigma^2 = 1, \ell = 10$$



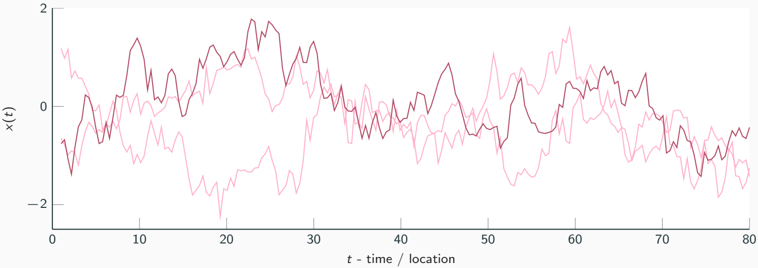
The exponential covariance function

The mean and covariance functions encode our prior knowledge.
One common choice is:

$$\boldsymbol{\mu}(t) = \mathbf{0}$$

$$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|/\ell)$$

$$\sigma^2 = 1, \ell = 10$$



Demo

The “squared exponential” covariance function

Another common choice:

$\boldsymbol{\mu}(t) = \mathbf{0}$

$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|^2/2\ell^2)$

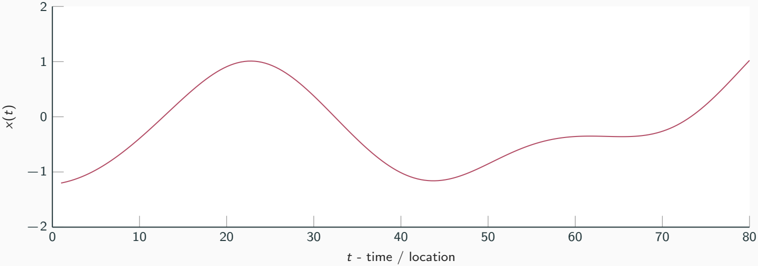
The "squared exponential" covariance function

Another common choice:

$\boldsymbol{\mu}(t) = \mathbf{0}$

$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|^2 / 2\ell^2)$

$\sigma^2 = 1, \ell = 10$



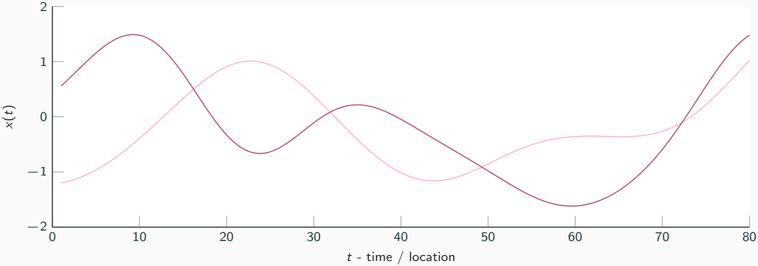
The “squared exponential” covariance function

Another common choice:

$\boldsymbol{\mu}(t) = \mathbf{0}$

$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|^2/2\ell^2)$

$\sigma^2 = 1, \ell = 10$



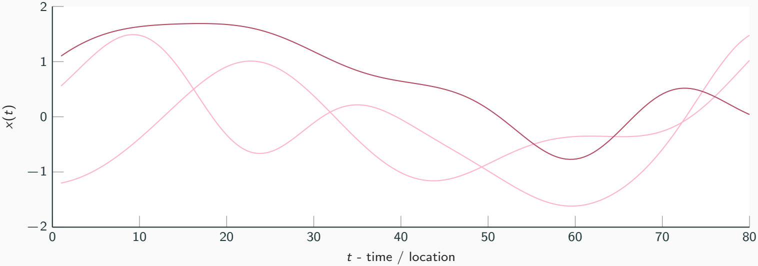
The "squared exponential" covariance function

Another common choice:

$\boldsymbol{\mu}(t) = \mathbf{0}$

$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|^2 / 2\ell^2)$

$\sigma^2 = 1, \ell = 10$

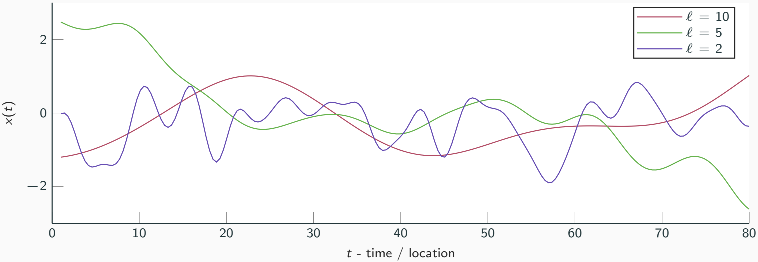


The "squared exponential" covariance function

Another common choice:

$$\boldsymbol{\mu}(t) = \mathbf{0}$$

$$\boldsymbol{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|^2 / 2\ell^2)$$



The quasi-periodic covariance function

$$\boldsymbol{\mu}(t) = \mathbf{0}$$

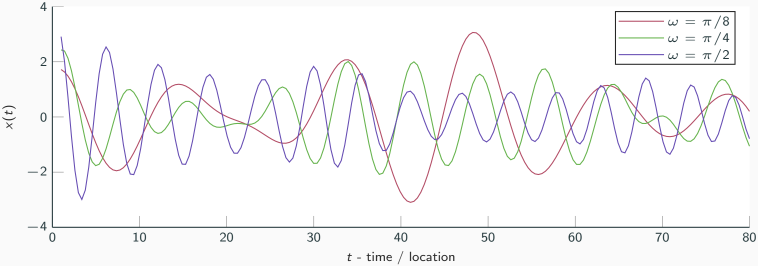
$$\boldsymbol{\Sigma}(t, t') = \sigma^2 \cos(\omega(t - t')) \exp(-|t - t'|^2 / 2\ell)$$

The quasi-periodic covariance function

$$\boldsymbol{\mu}(t) = \mathbf{0}$$

$$\boldsymbol{\Sigma}(t, t') = \sigma^2 \cos(\omega(t - t')) \exp(-|t - t'|^2/2\ell)$$

$$\sigma^2 = 1, \ell = 10$$



Stochastic Differential Equations

GPs have a strong connection to **stochastic differential equations (SDEs)**.

Stochastic Differential Equations

GPs have a strong connection to **stochastic differential equations (SDEs)**.

Assume $\mathbf{\Sigma}(t, t') = \sigma^2 \exp(-|t - t'|/\ell)$.

It can be shown that the SDE with this covariance is:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{-1}{\ell} x + \frac{\mathrm{d}\beta}{\mathrm{d}t}$$

where β is a Brownian motion with spectral density $2\sigma^2/\ell$.

Stochastic Differential Equations

More generally, we can write (almost) any

$$x(t) \sim GP(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t, t'))$$

as

$$\begin{aligned} \frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} &= \mathbf{F}\mathbf{z}(t) + \mathbf{L}\,\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t}, \\ x(t_k) &= \mathbf{H}\mathbf{z}(t_k) \end{aligned}$$

Discrete-time SDEs

The discrete-time representation of these SDEs is of the general form

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{A}\mathbf{z}_k + \mathbf{q}_k, \quad \mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}), \\ x_k &= \mathbf{H}\mathbf{z}_k \end{aligned}$$

Discrete-time SDEs

The discrete-time representation of these SDEs is of the general form

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{A}\mathbf{z}_k + \mathbf{q}_k, \quad \mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}), \\ x_k &= \mathbf{H}\mathbf{z}_k \end{aligned}$$

This gives us a new interpretation of sampling from a Gaussian process.

For exponential covariance:

$\mathbf{A} = \exp(-\Delta_t/\ell)$, $\mathbf{Q} = 2\sigma^2/\ell$, $\mathbf{H} = 1$

Discrete-time SDEs

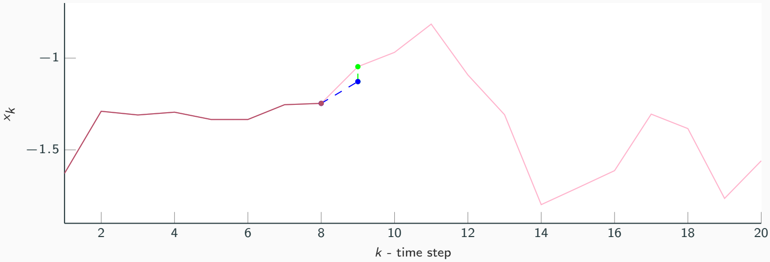
The discrete-time representation of these SDEs is of the general form

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{A}\mathbf{z}_k + \mathbf{q}_k, \quad \mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}), \\ x_k &= \mathbf{H}\mathbf{z}_k \end{aligned}$$

This gives us a new interpretation of sampling from a Gaussian process.

For exponential covariance:

$\mathbf{A} = \exp(-\Delta_t/\ell)$, $\mathbf{Q} = 2\sigma^2/\ell$, $\mathbf{H} = 1$



Demo

Bayesian Analysis

Now we can specify our **prior** knowledge and sample hypothetical signals.

But we're missing a crucial component

Bayesian Analysis

Now we can specify our **prior** knowledge and sample hypothetical signals.

But we're missing a crucial component - **the data**.

Bayesian Analysis

In Bayesian analysis, a complete **model** is specified by:

The **prior**

The **likelihood**

Bayesian Analysis

In Bayesian analysis, a complete **model** is specified by:

The **prior** - our assumptions / the data generating process

$$p(x)$$

The **likelihood**

Bayesian Analysis

In Bayesian analysis, a complete **model** is specified by:

The **prior** - our assumptions / the data generating process

$$p(x)$$

The **likelihood** - how we observe the data y given our prior

$$p(y|x)$$

Bayesian Analysis

In our previous examples we could choose the following:

$$\begin{aligned} &\textit{Prior} && p(x) && = GP(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t, t')) \\ &\textit{Likelihood} && p(y|x) && = N(x, \sigma_y^2 \mathbf{I}) \end{aligned}$$

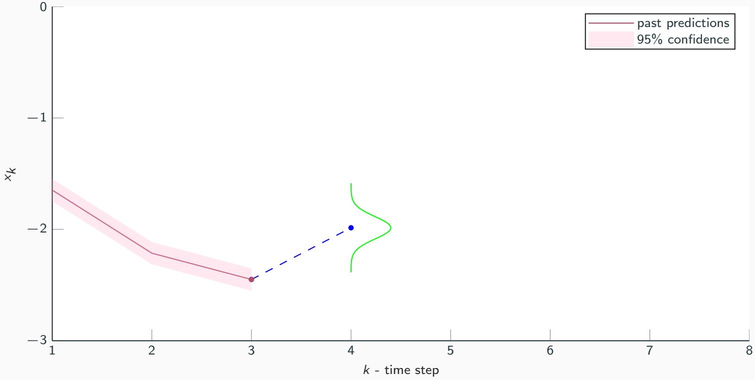
and

$$\begin{aligned} &\textit{Prior} && \frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} && = \mathbf{F}\mathbf{z}(t) + \mathbf{L}\,\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t}, \\ & && x(t_k) && = \mathbf{H}\mathbf{z}(t_k) \\ &\textit{Likelihood} && y_k && = x(t_k) + \sigma_y \varepsilon_k \end{aligned}$$

where $\varepsilon_k \sim N(0, 1)$ is Gaussian noise.

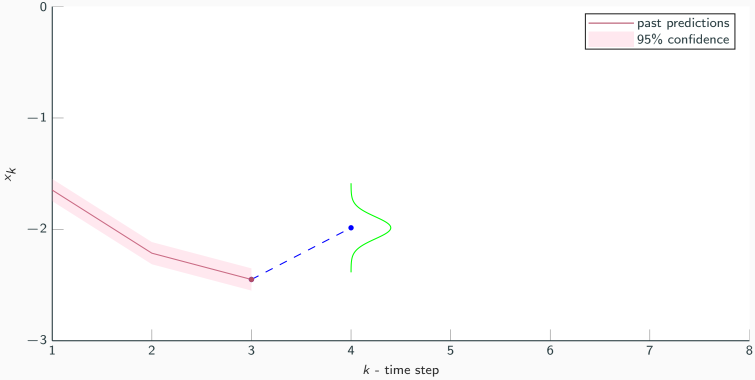
Bayesian Analysis

prior $p(x)$



Bayesian Analysis

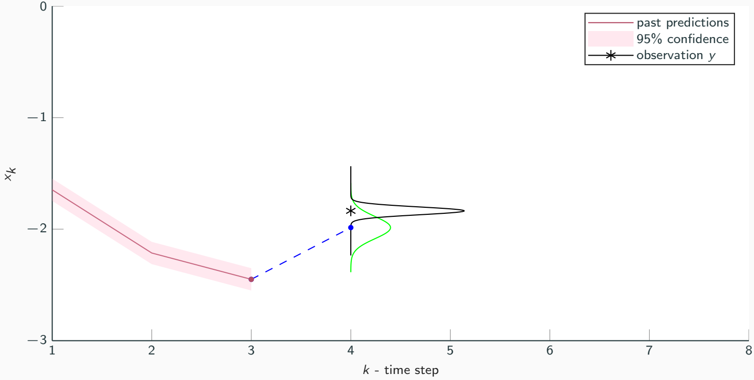
prior $p(x_4|x_{1:3})$



Bayesian Analysis

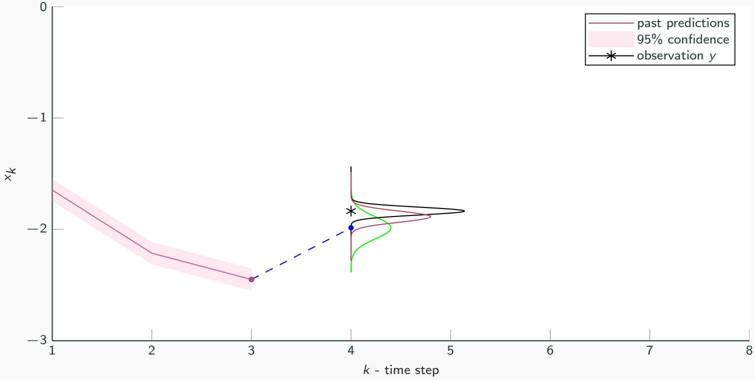
prior $p(x_4|x_{1:3})$

likelihood $p(y_4|x_4)$



Bayesian Analysis

prior	$p(x_4 \mid x_{1:3})$
likelihood	$p(y_4 \mid x_4)$
posterior	$p(x_4 \mid y_4)$

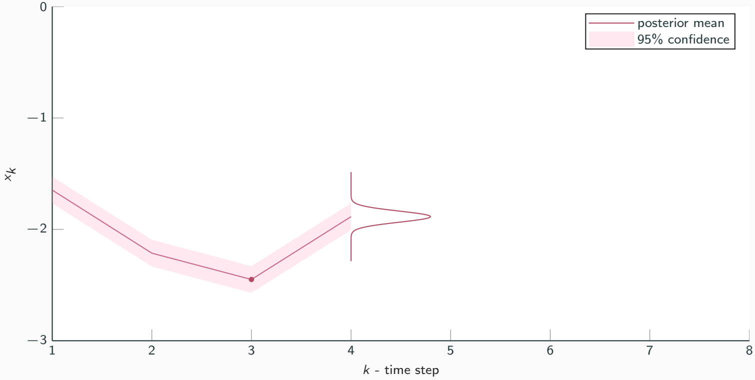


Bayesian Analysis

prior $p(x_4|x_{1:3})$

likelihood $p(y_4|x_4)$

posterior $p(x_4|y_4)$

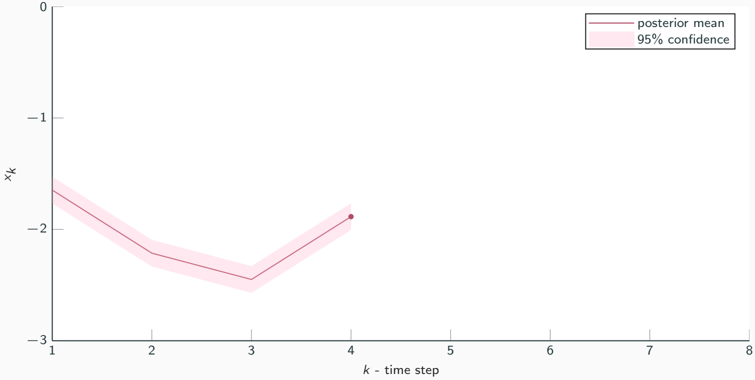


Bayesian Analysis

prior $p(x_4|x_{1:3})$

likelihood $p(y_4|x_4)$

posterior $p(x_4|y_4)$



Bayesian Analysis

prior	$p(x)$
likelihood	$p(y\|x)$
posterior	$p(x\|y)$

How do we combine the prior and the likelihood to get the posterior?

Bayesian Analysis

prior	$p(x)$
likelihood	$p(y\|x)$
posterior	$p(x\|y)$

How do we combine the prior and the likelihood to get the posterior?

$$p(x|y) = \frac{1}{Z} p(x, y)$$

Bayesian Analysis

prior	$p(x)$
likelihood	$p(y\|x)$
posterior	$p(x\|y)$

How do we combine the prior and the likelihood to get the posterior?

$$p(x|y) = \frac{1}{Z} p(y|x) p(x)$$

Bayesian Analysis

prior	$p(x)$
likelihood	$p(y\|x)$
posterior	$p(x\|y)$

How do we combine the prior and the likelihood to get the posterior?

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)\,\mathrm{d}x}$$

Bayesian Analysis

prior	$p(x)$
likelihood	$p(y\|x)$
posterior	$p(x\|y)$

How do we combine the prior and the likelihood to get the posterior?

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Bayesian Analysis

prior	$p(x)$
likelihood	$p(y\|x)$
posterior	$p(x\|y)$

How do we combine the prior and the likelihood to get the posterior?

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

This is called **Bayes rule**.

Marginal Likelihood

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

What is going on in the denominator?

$p(y) = \int p(y|x)p(x)\,\mathrm{d}x$

Marginal Likelihood

$$p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)}$$

What is going on in the denominator?

$p(y|\theta) = \int p(y|x,\theta)p(x|\theta)\,\mathrm{d}x$

Marginal Likelihood

$$p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)}$$

What is going on in the denominator?

$p(y|\theta) = \int p(y|x,\theta)p(x|\theta)\,\mathrm{d}x$

We integrate over all possible values of the latent variable x. This gives us the **marginal likelihood**.

Marginal Likelihood

$$p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)}$$

What is going on in the denominator?

$p(y|\theta) = \int p(y|x,\theta)p(x|\theta)\,\mathrm{d}x$

We integrate over all possible values of the latent variable x. This gives us the **marginal likelihood**.

It measures how much the data and the model "agree" with each other.

Marginal Likelihood

$$p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)}$$

What is going on in the denominator?

$p(y|\theta) = \int p(y|x,\theta)p(x|\theta)\,\mathrm{d}x$

This gives us a way to tune the model parameters. We treat it as an optimisation problem: maximising $p(y|\theta)$ with respect to θ.

Posterior calculations

Gaussian assumptions allow for efficient closed form calculations of the posterior process.

Posterior calculations

Standard approach

Posterior process characterised as $N(\mathbf{m}, \mathbf{P})$ where

$$\begin{aligned} \mathbf{m} &= \boldsymbol{\Sigma}_{t_*,t}(\boldsymbol{\Sigma}_{t,t} + \sigma_y^2 I)^{-1}\boldsymbol{y} \\ \mathbf{P} &= \boldsymbol{\Sigma}_{t_*,t_*} - \boldsymbol{\Sigma}_{t_*,t}(\boldsymbol{\Sigma}_{t,t} + \sigma_y^2 I)^{-1}\boldsymbol{\Sigma}_{t,t_*} \end{aligned}$$

$t_* =$ training locations
$t =$ test locations

SDE approach

Kalman filtering and smoothing returns the posterior.

prediction step:

$$\begin{aligned} \mathbf{m}_k &= \mathbf{A}\mathbf{m}_{k-1} \\ \mathbf{P}_k &= \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^\top \end{aligned}$$

update step:

$$\begin{aligned} \mathbf{v}_k &= y_k - \mathbf{H}_k\mathbf{m}_k \\ \mathbf{S}_k &= \mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\top + \sigma_y^2 \\ \mathbf{K}_k &= \mathbf{P}_k\mathbf{H}_k^\top\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k + \mathbf{K}_k\mathbf{v}_k \\ \mathbf{P}_k &= \mathbf{P}_k - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^\top \end{aligned}$$

A probabilistic model for time-frequency analysis

(Matérn) **Spectral Mixture GP**:

$$\text{[Prior]} \qquad x(t) \sim \mathrm{GP}(\mathbf{0}, \sum_{d=1}^{D} \kappa_{\mathrm{sm}}^{(d)}(t, t')),$$

$$\text{[Likelihood]} \qquad y_k = x(t_k) + \sigma_{\mathrm{y}_k} \varepsilon_k,$$

A probabilistic model for time-frequency analysis

(Matérn) **Spectral Mixture GP**:

$$\text{[Prior]} \qquad x(t) \sim \mathrm{GP}(\mathbf{0}, \sum_{d=1}^{D} \kappa_{\mathrm{sm}}^{(d)}(t, t')),$$

$$\text{[Likelihood]} \qquad y_k = x(t_k) + \sigma_{\mathrm{y}_k}\, \varepsilon_k,$$

$$\kappa_{\mathrm{sm}}^{(d)}(t, t') = \sigma_d^2 \cos(\omega_d\,(t - t'))\exp(-|t - t'|/\ell_d)$$

A probabilistic model for time-frequency analysis

(Matérn) **Spectral Mixture GP**:

$$
\begin{aligned}
&[\text{Prior}] \qquad x(t) \sim \mathrm{GP}(\mathbf{0}, \sum_{d=1}^{D} \kappa_{\mathrm{sm}}^{(d)}(t, t')), \\
&[\text{Likelihood}] \quad y_k = x(t_k) + \sigma_{\mathrm{y}_k}\, \varepsilon_k,
\end{aligned}
$$

$\kappa_{\mathrm{sm}}^{(d)}(t, t') = \sigma_d^2 \cos(\omega_d\,(t - t'))\exp(-|t - t'|/\ell_d)$
$d = 1, \ldots, D$ frequency channels / filters

A probabilistic model for time-frequency analysis

(Matérn) **Spectral Mixture GP**:

$$
\begin{aligned}
&[\text{Prior}] \qquad && x(t) \sim \mathrm{GP}(\mathbf{0}, \sum_{d=1}^{D} \kappa_{\mathrm{sm}}^{(d)}(t, t')), \\
&[\text{Likelihood}] \qquad && y_k = x(t_k) + \sigma_{\mathrm{y}_k} \, \varepsilon_k,
\end{aligned}
$$

$\kappa_{\mathrm{sm}}^{(d)}(t, t') = \sigma_d^2 \cos(\omega_d \, (t - t')) \exp(-|t - t'|/\ell_d)$
$d = 1, \dots, D$ frequency channels / filters
ω_d - centre frequency

A probabilistic model for time-frequency analysis

(Matérn) **Spectral Mixture GP**:

$$
\begin{aligned}
&\text{[Prior]} \qquad && x(t) \sim \mathrm{GP}(\mathbf{0}, \sum_{d=1}^{D} \kappa_{\mathrm{sm}}^{(d)}(t, t')), \\
&\text{[Likelihood]} \qquad && y_k = x(t_k) + \sigma_{\mathrm{y}_k} \varepsilon_k,
\end{aligned}
$$

$\kappa_{\mathrm{sm}}^{(d)}(t, t') = \sigma_d^2 \cos(\omega_d \, (t - t')) \exp(-|t - t'|/\ell_d)$

$d = 1, \ldots, D$ frequency channels / filters

ω_d - centre frequency

ℓ_d - controls the filter bandwidth

A probabilistic model for time-frequency analysis

$\kappa_{\mathrm{sm}}^{(d)}(t,t') = \sigma_d^2 \cos(\omega_d\,(t-t'))\,\exp(-|t-t'|/\ell_d)$

The SDE with this covariance is:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{F}\mathbf{x}(t) + \mathbf{L}\,\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t}, \\ y(t_k) &= \mathbf{H}\mathbf{x}(t_k) + \sigma_y \varepsilon_k \end{aligned}$$

A probabilistic model for time-frequency analysis

$$\kappa^{(d)}_{\mathrm{sm}}(t, t') = \sigma_d^2 \cos(\omega_d\,(t - t'))\,\exp(-|t - t'|/\ell_d)$$

The SDE with this covariance is:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{F}\mathbf{x}(t) + \mathbf{L}\,\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t}, \\ y(t_k) &= \mathbf{H}\mathbf{x}(t_k) + \sigma_y \varepsilon_k \end{aligned}$$

$\mathbf{F}^{(d)}_{\cos} = \left(\begin{smallmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{smallmatrix}\right)$, and $\mathbf{F}^{(d)}_{\exp} = -1/\ell_d$

A probabilistic model for time-frequency analysis

$$\kappa_{\mathrm{sm}}^{(d)}(t, t') = \sigma_d^2 \cos(\omega_d (t - t')) \exp(-|t - t'|/\ell_d)$$

The SDE with this covariance is:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{F}\mathbf{x}(t) + \mathbf{L}\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t}, \\ y(t_k) &= \mathbf{H}\mathbf{x}(t_k) + \sigma_y \varepsilon_k \end{aligned}$$

$\mathbf{F}_{\cos}^{(d)} = \begin{pmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{pmatrix}$, and $\mathbf{F}_{\exp}^{(d)} = -1/\ell_d$

$\mathbf{F}^{(d)} = \frac{-1}{\ell_d} \begin{pmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{pmatrix}$

A probabilistic model for time-frequency analysis

$$\kappa_{\mathrm{sm}}^{(d)}(t,t') = \sigma_d^2 \cos(\omega_d\,(t-t'))\,\exp(-|t-t'|/\ell_d)$$

The SDE with this covariance is:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{F}\mathbf{x}(t) + \mathbf{L}\,\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t},\\ y(t_k) &= \mathbf{H}\mathbf{x}(t_k) + \sigma_y \varepsilon_k\end{aligned}$$

$\mathbf{F}_{\cos}^{(d)} = \left(\begin{smallmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{smallmatrix}\right)$, and $\mathbf{F}_{\exp}^{(d)} = -1/\ell_d$

$\mathbf{F}^{(d)} = \frac{-1}{\ell_d}\left(\begin{smallmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{smallmatrix}\right)$

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{(1)} & & 0 \\ & \ddots & \\ 0 & & \mathbf{F}^{(D)} \end{pmatrix}$$

A probabilistic model for time-frequency analysis

$$\kappa_{\rm sm}^{(d)}(t,t') = \sigma_d^2 \cos(\omega_d\,(t-t'))\exp(-|t-t'|/\ell_d)$$

The SDE with this covariance is:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{F}\mathbf{x}(t) + \mathbf{L}\,\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t},\\ y(t_k) &= \mathbf{H}\mathbf{x}(t_k) + \sigma_y \varepsilon_k\end{aligned}$$

$\mathbf{F}_{\rm cos}^{(d)} = \left(\begin{smallmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{smallmatrix}\right)$, and $\mathbf{F}_{\rm exp}^{(d)} = -1/\ell_d$

$$\boldsymbol{\beta} \sim N(0, \mathbf{Q})$$

$\mathbf{F}^{(d)} = \frac{-1}{\ell_d}\left(\begin{smallmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{smallmatrix}\right)$

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{(1)} & & 0 \\ & \ddots & \\ 0 & & \mathbf{F}^{(D)} \end{pmatrix} \qquad \mathbf{Q} = \begin{pmatrix} \frac{2\sigma_1^2}{\ell_1}\mathbf{I} & & 0 \\ & \ddots & \\ 0 & & \frac{2\sigma_D^2}{\ell_D}\mathbf{I} \end{pmatrix}$$

A probabilistic model for time-frequency analysis

What is the discrete form of $\mathbf{F}^{(d)} = \frac{-1}{\ell_d} \left(\begin{smallmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{smallmatrix} \right)$?

A probabilistic model for time-frequency analysis

What is the discrete form of $\mathbf{F}^{(d)} = \frac{-1}{\ell_d} \begin{pmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{pmatrix}$?

$\mathbf{A}^{(d)} = \exp(\Delta_t \mathbf{F}^{(d)})$

A probabilistic model for time-frequency analysis

What is the discrete form of $\mathbf{F}^{(d)} = \frac{-1}{\ell_d} \begin{pmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{pmatrix}$?

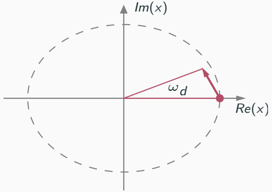
$$\mathbf{A}^{(d)} = \mathrm{e}^{\frac{-1}{\ell_d}} \begin{pmatrix} cos\omega_d & -sin\omega_d \\ sin\omega_d & cos\omega_d \end{pmatrix}$$

A probabilistic model for time-frequency analysis

What is the discrete form of $\mathbf{F}^{(d)} = \frac{-1}{\ell_d} \begin{pmatrix} 0 & -\omega_d \\ \omega_d & 0 \end{pmatrix}$?

$$\mathbf{A}^{(d)} = \mathrm{e}^{\frac{-1}{\ell_d}} \begin{pmatrix} cos\omega_d & -sin\omega_d \\ sin\omega_d & cos\omega_d \end{pmatrix}$$

This describes a rotation with frequency ω_d, i.e. a phasor.



A probabilistic model for time-frequency analysis

$$
\begin{aligned}
&[\text{Prior}] && \mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{q}_k, \qquad \mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}), \\
&[\text{Likelihood}] && y_k = \mathbf{H}\mathbf{x}_k + \sigma_{\mathrm{y}_k}\varepsilon_k
\end{aligned}
$$

A probabilistic model for time-frequency analysis

$$\begin{aligned}
\mathbf{x}_{k+1}^{(d)} &= \mathrm{e}^{\frac{-1}{\ell_d}} \left(\begin{smallmatrix} cos\omega_d & -sin\omega_d \\ sin\omega_d & cos\omega_d \end{smallmatrix}\right) \mathbf{x}_k^{(d)} + \mathbf{q}_k^{(d)}, \\
y_k &= \begin{pmatrix} 1 & 0 & \dots & 1 & 0 \end{pmatrix} \mathbf{x}_k + \sigma_{\mathrm{y}_k} \varepsilon_k
\end{aligned}$$

A probabilistic model for time-frequency analysis

$$\begin{aligned}\mathbf{x}_{k+1}^{(d)} &= \mathrm{e}^{\frac{-1}{\ell_d}} \begin{pmatrix} cos\omega_d & -sin\omega_d \\ sin\omega_d & cos\omega_d \end{pmatrix} \mathbf{x}_k^{(d)} + \mathbf{q}_k^{(d)}, \\ y_k &= \begin{pmatrix} 1 & 0 & \ldots & 1 & 0 \end{pmatrix} \mathbf{x}_k + \sigma_{\mathrm{y}_k} \varepsilon_k\end{aligned}$$

Consider $\mathbf{x}_k^{(d)} = \begin{pmatrix} \mathrm{Re}(z_k^{(d)}) \\ \mathrm{Im}(z_k^{(d)}) \end{pmatrix}$

A probabilistic model for time-frequency analysis

$$\begin{aligned}
\mathbf{x}_{k+1}^{(d)} &= \mathrm{e}^{\frac{-1}{\ell_d}} \begin{pmatrix} cos\omega_d & -sin\omega_d \\ sin\omega_d & cos\omega_d \end{pmatrix} \mathbf{x}_k^{(d)} + \mathbf{q}_k^{(d)}, \\
y_k &= \begin{pmatrix} 1 & 0 & \dots & 1 & 0 \end{pmatrix} \mathbf{x}_k + \sigma_{\mathrm{y}_k} \varepsilon_k
\end{aligned}$$

Consider $\mathbf{x}_k^{(d)} = \begin{pmatrix} \mathrm{Re}(z_k^{(d)}) \\ \mathrm{Im}(z_k^{(d)}) \end{pmatrix}$

$$\begin{aligned}
z_{k+1}^{(d)} &= \mathrm{e}^{\frac{-1}{\ell_d}} \mathrm{e}^{\mathrm{i}\omega_d} z_k^{(d)} + q_k^{(d)}, \\
y_k &= \sum_{d=1}^{D} \mathrm{Re}(z_k^{(d)}) + \sigma_{\mathrm{y}_k} \varepsilon_k
\end{aligned}$$

A probabilistic model for time-frequency analysis

$$\begin{aligned}\mathbf{x}_{k+1}^{(d)} &= \mathrm{e}^{\frac{-1}{\ell_d}} \begin{pmatrix} cos\omega_d & -sin\omega_d \\ sin\omega_d & cos\omega_d \end{pmatrix} \mathbf{x}_k^{(d)} + \mathbf{q}_k^{(d)}, \\ y_k &= \begin{pmatrix} 1 & 0 & \dots & 1 & 0 \end{pmatrix} \mathbf{x}_k + \sigma_{\mathrm{y}_k} \varepsilon_k\end{aligned}$$

Consider $\mathbf{x}_k^{(d)} = \begin{pmatrix} \mathrm{Re}(z_k^{(d)}) \\ \mathrm{Im}(z_k^{(d)}) \end{pmatrix}$

$$\begin{aligned}z_{k+1}^{(d)} &= \psi_d \mathrm{e}^{\mathrm{i}\omega_d} z_k^{(d)} + q_k^{(d)}, \\ y_k &= \sum_{d=1}^{D} \mathrm{Re}(z_k^{(d)}) + \sigma_{\mathrm{y}_k} \varepsilon_k\end{aligned}$$

A probabilistic model for time-frequency analysis

$$\begin{aligned} z_{k+1}^{(d)} &= \psi_d \mathrm{e}^{\mathrm{i}\omega_d} z_k^{(d)} + q_k^{(d)}, \\ y_k &= \sum_{d=1}^{D} \mathrm{Re}(z_k^{(d)}) + \sigma_{\mathrm{y}_k} \varepsilon_k \end{aligned}$$

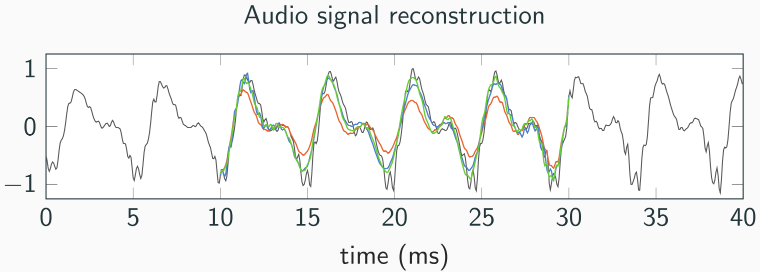
A probabilistic model for time-frequency analysis

$$
\begin{aligned}
z_{k+1}^{(d)} &= \psi_d \mathrm{e}^{\mathrm{i}\omega_d} z_k^{(d)} + q_k^{(d)}, \\
y_k &= \sum_{d=1}^{D} \mathrm{Re}(z_k^{(d)}) + \sigma_{\mathrm{y}_k} \varepsilon_k
\end{aligned}
$$

This is called the **probabilistic phase vocoder**.

Demo

Missing Data Synthesis



Data imputation using a filter bank composed of the following kernels:

Matérn1/2 (exponential) - 1^{st} *order state space form*

Matérn3/2 - 2^{nd} *order state space form*

Matérn5/2 - 3^{rd} *order state space form*

Next steps

Watch this space:

- We're going to make this model really fast - i.e. real time processing.
- We're going to make it accessible.
- We're going to glue on a model for the amplitude (i.e. the spectrogram) which measures correlation across frequency channels.

Summary

Thanks for listening - any questions?

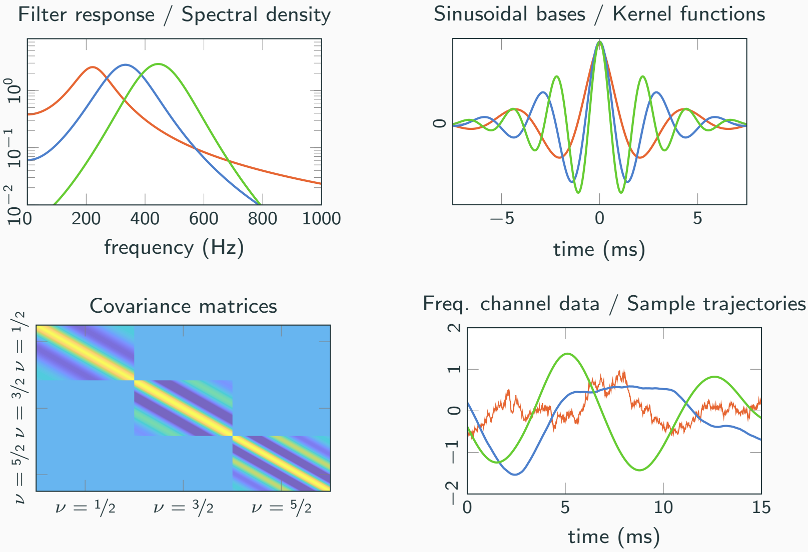
Paper is here:

`https://arxiv.org/abs/1811.02489`

Code is here:

`https://github.com/wil-j-wil/unifying-prob-time-freq`

Appendix - kernel comparison



Matérn$1/2$ (exponential) - 1^{st} *order SS* , Matérn$3/2$ - 2^{nd} *order SS*, Matérn$5/2$ - 3^{rd} *order SS*