An introduction to Gaussian Processes

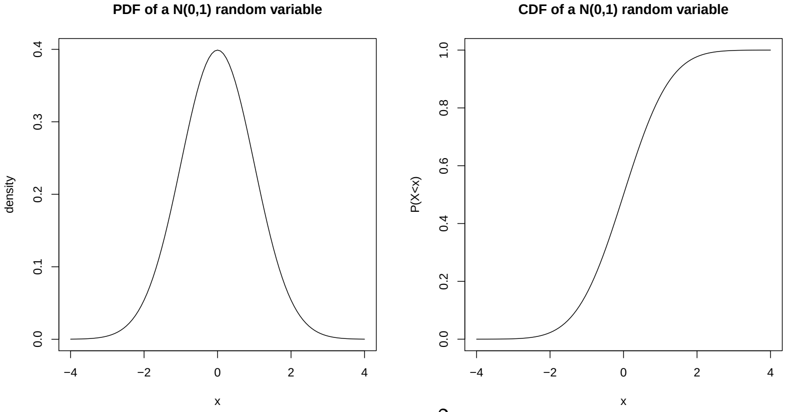
Richard Wilkinson

School of Maths and Statistics
University of Sheffield

GP summer school
September 2019

Introduction

Univariate Gaussian distributions

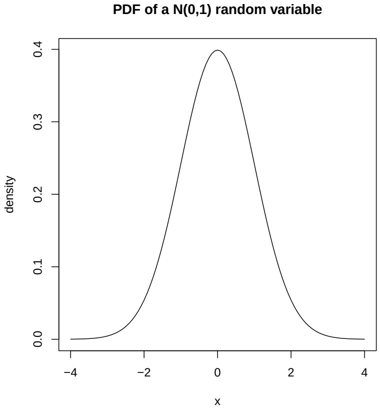


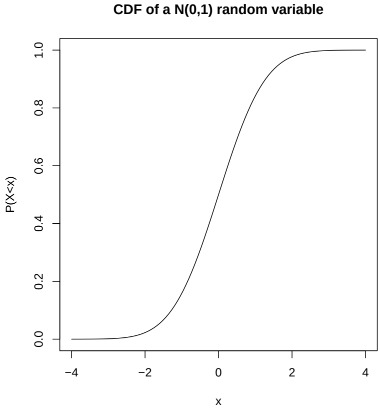
$$X \sim N(\mu, \sigma^2)$$

PDF: $$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

CDF: $F_X(x) = \mathbb{P}(X \leq x)$ not known in closed form

Univariate Gaussian distributions





$$X \sim N(\mu, \sigma^2)$$

PDF: $$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

CDF: $F_X(x) = \mathbb{P}(X \leq x)$ not known in closed form

If $Z \sim N(0,1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Family of normal distributions is closed under linear operations (more later).

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Family of normal distributions is closed under linear operations (more later).
- Central limit theorem

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Family of normal distributions is closed under linear operations (more later).
- Central limit theorem
- Maximum entropy: $N(\mu, \sigma^2)$ has maximum entropy of any distribution with mean μ and variance σ^2 (max. ent. principle: the distribution with the largest entropy should be used as a least-informative default)

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Family of normal distributions is closed under linear operations (more later).
- Central limit theorem
- Maximum entropy: $N(\mu, \sigma^2)$ has maximum entropy of any distribution with mean μ and variance σ^2 (max. ent. principle: the distribution with the largest entropy should be used as a least-informative default)
- Infinite divisibility

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Family of normal distributions is closed under linear operations (more later).
- Central limit theorem
- Maximum entropy: $N(\mu, \sigma^2)$ has maximum entropy of any distribution with mean μ and variance σ^2 (max. ent. principle: the distribution with the largest entropy should be used as a least-informative default)
- Infinite divisibility
- If normally distributed rvs X and Y are uncorrelated, then they are independent

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Family of normal distributions is closed under linear operations (more later).
- Central limit theorem
- Maximum entropy: $N(\mu, \sigma^2)$ has maximum entropy of any distribution with mean μ and variance σ^2 (max. ent. principle: the distribution with the largest entropy should be used as a least-informative default)
- Infinite divisibility
- If normally distributed rvs X and Y are uncorrelated, then they are independent
- Square-loss functions lead to procedures that have a Gaussian probabilistic interpretation
  eg Fit model $f_\beta(x)$ to data y by mimizing $\sum(y_i - f_\beta(x_i))^2$ is equivalent to maximum likelihood estimation under the assumption that $y = f_\beta(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.

Multivariate Gaussian distributions

'Multivariate' $=$ two or more random variables

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

Suppose $X \in \mathbb{R}^d$ has a multivariate Gaussian distribution with

- **mean vector** $\mu \in \mathbb{R}^d$
- **covariance matrix** $\Sigma \in \mathbb{R}^{d \times d}$.

Write

$$X \sim N_d(\mu, \Sigma)$$

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

Suppose $X \in \mathbb{R}^d$ has a multivariate Gaussian distribution with

- **mean vector** $\mu \in \mathbb{R}^d$
- **covariance matrix** $\Sigma \in \mathbb{R}^{d \times d}$.

Write

$$X \sim N_d(\mu, \Sigma)$$

Bivariate Gaussian: d=2

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

Suppose $X \in \mathbb{R}^d$ has a multivariate Gaussian distribution with

- **mean vector** $\mu \in \mathbb{R}^d$
- **covariance matrix** $\Sigma \in \mathbb{R}^{d \times d}$.

Write

$$X \sim N_d(\mu, \Sigma)$$

Bivariate Gaussian: d=2

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$\mathbb{V}\mathrm{ar}(X_i) = \sigma_i^2 \quad \mathbb{C}\mathrm{ov}(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j \quad \mathrm{Cor}(X_i, X_j) = \rho_{12} \text{ for } i \neq j$$

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

Suppose $X \in \mathbb{R}^d$ has a multivariate Gaussian distribution with

- **mean vector** $\mu \in \mathbb{R}^d$
- **covariance matrix** $\Sigma \in \mathbb{R}^{d \times d}$.

Write

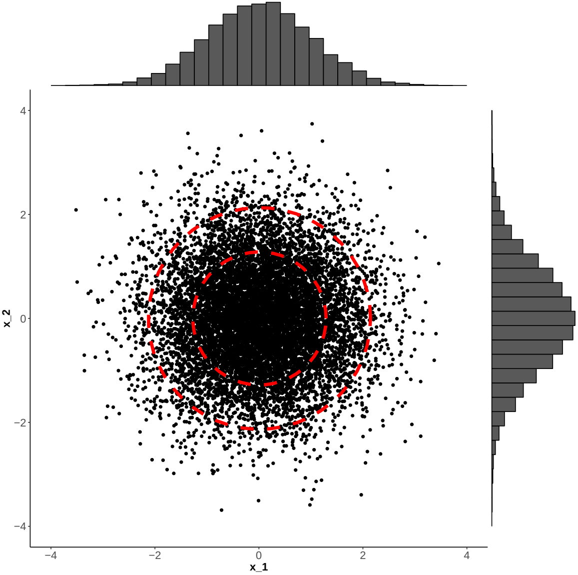
$$X \sim N_d(\mu, \Sigma)$$

Bivariate Gaussian: d=2

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

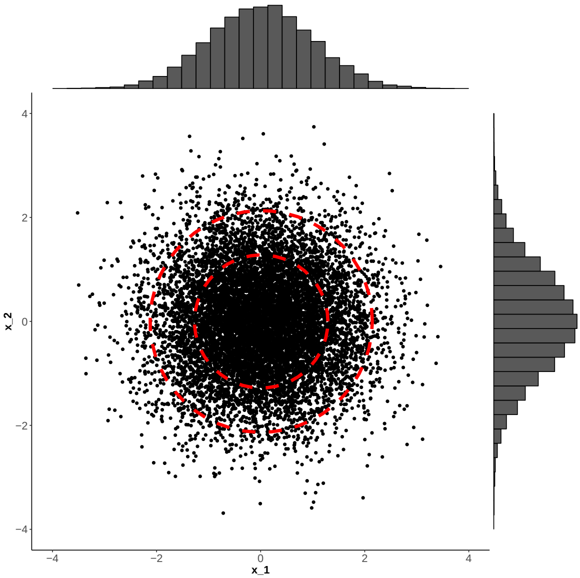
$$\mathbb{V}\text{ar}(X_i) = \sigma_i^2 \quad \mathbb{C}\text{ov}(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j \quad \text{Cor}(X_i, X_j) = \rho_{12} \text{ for } i \neq j$$

pdf: $\quad f(x \mid \mu, \Sigma) = |\Sigma|^{-\frac{1}{2}}(2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

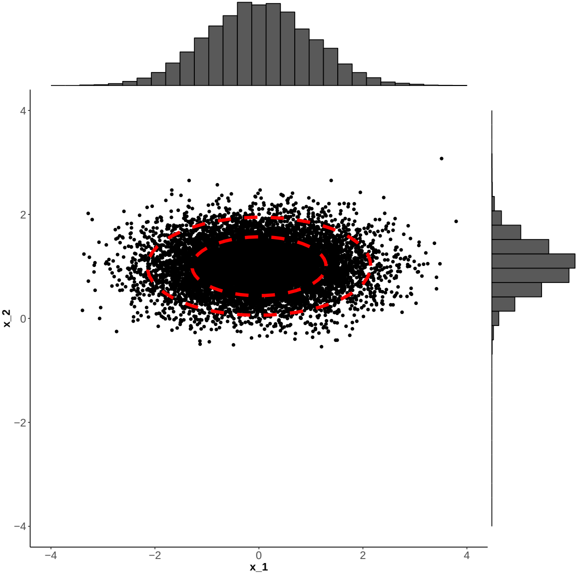
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

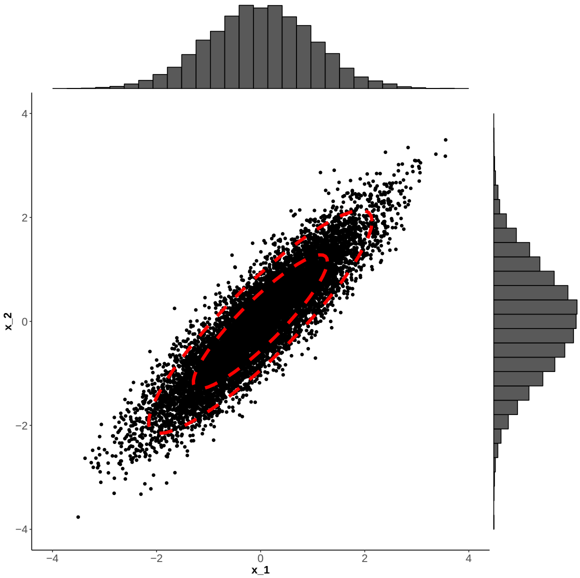
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

So
$Cor(X_1, X_2) = 0$
hence X_1
independent of X_2



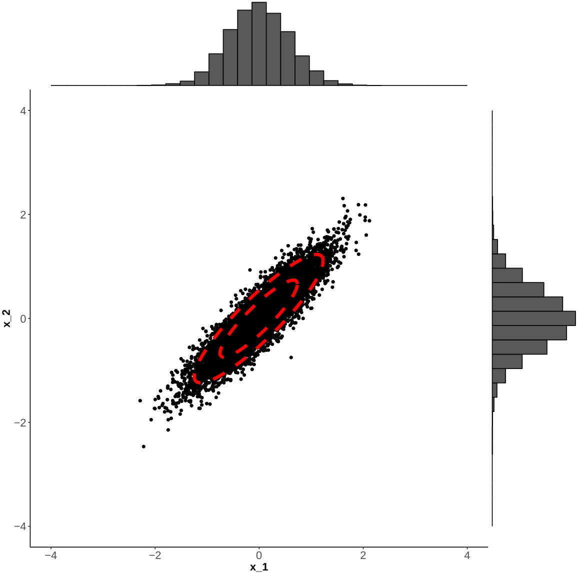
$$\mu = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix}$$



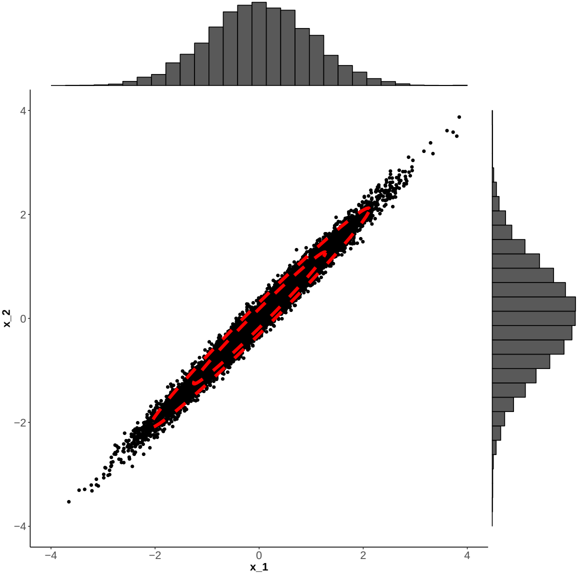
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$



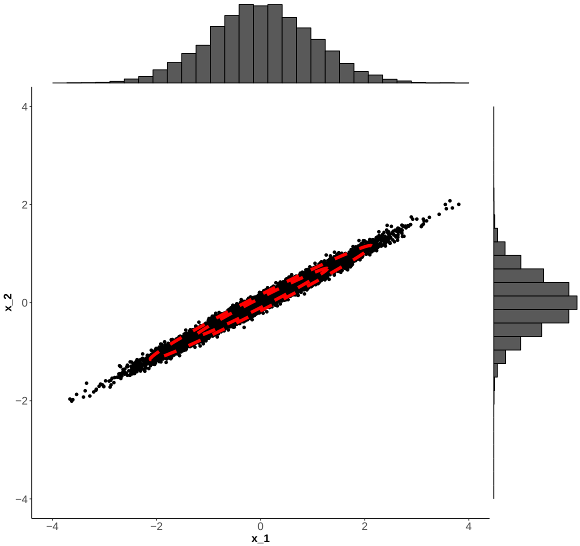
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \frac{1}{3} \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$$



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.54 \\ 0.54 & 0.3 \end{pmatrix}$$

$Cor(X_1, X_2) =$
$0.54/\sqrt{(0.3)} =$
0.99

More pictures

Consider $d = 5$ with

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.8 & 0.7 & 0.6 \\ 0.9 & 1 & 0.9 & 0.8 & 0.7 \\ 0.8 & 0.9 & 1 & 0.9 & 0.8 \\ 0.7 & 0.8 & 0.9 & 1 & 0.9 \\ 0.6 & 0.7 & 0.8 & 0.9 & 1 \end{pmatrix}$$

It's hard to visualise in dimensions > 3, so let's stack points next to each other.

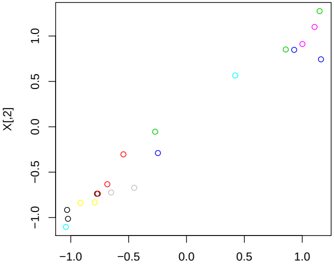
More pictures

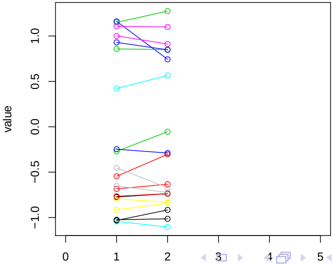
Consider $d = 5$ with

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.8 & 0.7 & 0.6 \\ 0.9 & 1 & 0.9 & 0.8 & 0.7 \\ 0.8 & 0.9 & 1 & 0.9 & 0.8 \\ 0.7 & 0.8 & 0.9 & 1 & 0.9 \\ 0.6 & 0.7 & 0.8 & 0.9 & 1 \end{pmatrix}$$

It's hard to visualise in dimensions > 3, so let's stack points next to each other.

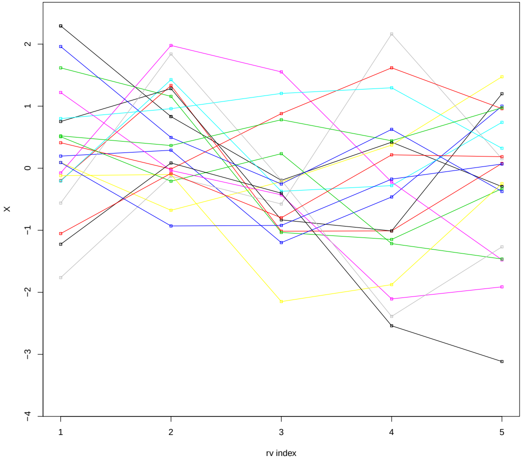
So for 2d instead of we have





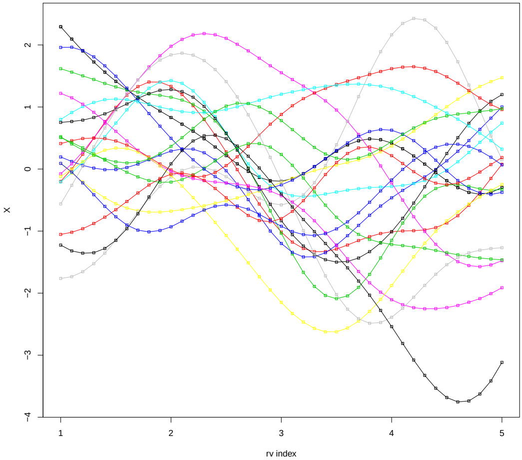
d=5

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.8 & 0.7 & 0.6 \\ 0.9 & 1 & 0.9 & 0.8 & 0.7 \\ 0.8 & 0.9 & 1 & 0.9 & 0.8 \\ 0.7 & 0.8 & 0.9 & 1 & 0.9 \\ 0.6 & 0.7 & 0.8 & 0.9 & 1 \end{pmatrix}$$



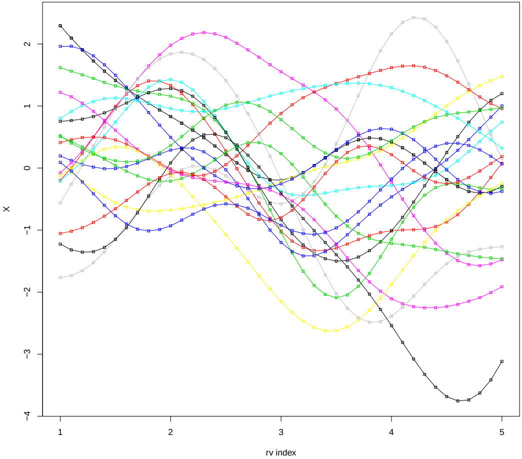
d=50

$$\Sigma = \begin{pmatrix} 1 & 0.99 & 0.98 & 0.97 & 0.96 & \ldots \\ 0.99 & 1 & 0.99 & 0.98 & 0.97 & \ldots \\ 0.98 & 0.99 & 1 & 0.99 & 0.98 & \ldots \\ 0.97 & 0.98 & 0.99 & 1 & 0.99 & \ldots \\ 0.96 & 0.97 & 0.98 & 0.99 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$



d=50

$$\Sigma = \begin{pmatrix} 1 & 0.99 & 0.98 & 0.97 & 0.96 & \dots \\ 0.99 & 1 & 0.99 & 0.98 & 0.97 & \dots \\ 0.98 & 0.99 & 1 & 0.99 & 0.98 & \dots \\ 0.97 & 0.98 & 0.99 & 1 & 0.99 & \dots \\ 0.96 & 0.97 & 0.98 & 0.99 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$



We can think of Gaussian processes as an infinite dimensional distribution over functions.

Gaussian processes

A stochastic process is a collection of random variables indexed by some variable $x \in \mathcal{X}$

$$f = \{f(x) : x \in \mathcal{X}\}$$

Gaussian processes

A stochastic process is a collection of random variables indexed by some variable $x \in \mathcal{X}$

$$f = \{f(x) : x \in \mathcal{X}\}$$

Usually $f(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ i.e. f can be thought of as a function of location x.

Gaussian processes

A stochastic process is a collection of random variables indexed by some variable $x \in \mathcal{X}$

$$f = \{f(x) : x \in \mathcal{X}\}$$

Usually $f(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ i.e. f can be thought of as a function of location x.

If $\mathcal{X} = \mathbb{R}^n$, then f is an infinite dimensional process.

Gaussian processes

A stochastic process is a collection of random variables indexed by some variable $x \in \mathcal{X}$

$$f = \{f(x) : x \in \mathcal{X}\}$$

Usually $f(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ i.e. f can be thought of as a function of location x.
If $\mathcal{X} = \mathbb{R}^n$, then f is an infinite dimensional process.
Thankfully we only need consider the finite dimensional distributions (FDDs), i.e., for all $x_1, \ldots, x_n$ and for all $n \in \mathbb{N}$

$$\mathbb{P}(f(x_1) \leq y_1, \ldots, f(x_n) \leq y_n)$$

as these uniquely determine the law of f.

Gaussian processes

A stochastic process is a collection of random variables indexed by some variable $x \in \mathcal{X}$

$$f = \{f(x) : x \in \mathcal{X}\}$$

Usually $f(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ i.e. f can be thought of as a function of location x.
If $\mathcal{X} = \mathbb{R}^n$, then f is an infinite dimensional process.
Thankfully we only need consider the finite dimensional distributions (FDDs), i.e., for all $x_1, \ldots, x_n$ and for all $n \in \mathbb{N}$

$$\mathbb{P}(f(x_1) \leq y_1, \ldots, f(x_n) \leq y_n)$$

as these uniquely determine the law of f.
A **Gaussian process** is a stochastic process with Gaussian FDDs, i.e.,

$$(f(x_1), \ldots, f(x_n)) \sim N_n(\mu, \Sigma)$$

Gaussian processes

A stochastic process is a collection of random variables indexed by some variable $x \in \mathcal{X}$

$$f = \{f(x) : x \in \mathcal{X}\}$$

Usually $f(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ i.e. f can be thought of as a function of location x.

If $\mathcal{X} = \mathbb{R}^n$, then f is an infinite dimensional process.

Thankfully we only need consider the finite dimensional distributions (FDDs), i.e., for all $x_1, \ldots, x_n$ and for all $n \in \mathbb{N}$

$$\mathbb{P}(f(x_1) \leq y_1, \ldots, f(x_n) \leq y_n)$$

as these uniquely determine the law of f.

A **Gaussian process** is a stochastic process with Gaussian FDDs, i.e.,

$$(f(x_1), \ldots, f(x_n)) \sim N_n(\mu, \Sigma)$$

Why use Gaussian processes?

Why would we want to use this very restricted class of model?

Gaussian **distributions** have several properties that make them easy to work with:

Why use Gaussian processes?

Why would we want to use this very restricted class of model?

Gaussian **distributions** have several properties that make them easy to work with:

Proposition:

$$X \sim N_d(\mu, \Sigma) \text{ if and only if } AX \sim N_p(A\mu, A\Sigma A^\top)$$

for all $A \in \mathbb{R}^{p\times d}$.

Why use Gaussian processes?

Why would we want to use this very restricted class of model?

Gaussian **distributions** have several properties that make them easy to work with:

Proposition:

$$X \sim N_d(\mu, \Sigma) \text{ if and only if } AX \sim N_p(A\mu, A\Sigma A^\top)$$

for all $A \in \mathbb{R}^{p \times d}$.
So sums of Gaussians are Gaussian, and marginal distributions of multivariate Gaussians are still Gaussian.

Why use Gaussian processes?

Why would we want to use this very restricted class of model?

Gaussian **distributions** have several properties that make them easy to work with:

Proposition:

$$X \sim N_d(\mu, \Sigma) \text{ if and only if } AX \sim N_p(A\mu, A\Sigma A^\top)$$

for all $A \in \mathbb{R}^{p\times d}$.

So sums of Gaussians are Gaussian, and marginal distributions of multivariate Gaussians are still Gaussian.

Corollary: Σ must be positive semi-definite as $a^\top \Sigma a \geq 0$ for all $a \in \mathbb{R}^d$.

Why use Gaussian processes?

Why would we want to use this very restricted class of model?

Gaussian **distributions** have several properties that make them easy to work with:

Proposition:

$$X \sim N_d(\mu, \Sigma) \text{ if and only if } AX \sim N_p(A\mu, A\Sigma A^\top)$$

for all $A \in \mathbb{R}^{p \times d}$.
So sums of Gaussians are Gaussian, and marginal distributions of multivariate Gaussians are still Gaussian.

Corollary: Σ must be positive semi-definite as $a^\top \Sigma a \geq 0$ for all $a \in \mathbb{R}^d$.

Conversely, any matrix Σ which is positive semi-definite is a valid covariance matrix:

If $Z \sim N_d(0_d, I_d)$ then $X = \mu + \Sigma^{\frac{1}{2}} Z \sim N_d(\mu, \Sigma)$.

Where $\Sigma^{\frac{1}{2}}$ is a matrix square root of Σ.

Why use Gaussian processes?

Why would we want to use this very restricted class of model?

Gaussian **distributions** have several properties that make them easy to work with:

Proposition:

$$X \sim N_d(\mu, \Sigma) \text{ if and only if } AX \sim N_p(A\mu, A\Sigma A^\top)$$

for all $A \in \mathbb{R}^{p \times d}$.
So sums of Gaussians are Gaussian, and marginal distributions of multivariate Gaussians are still Gaussian.

Corollary: Σ must be positive semi-definite as $a^\top \Sigma a \geq 0$ for all $a \in \mathbb{R}^d$.

Conversely, any matrix Σ which is positive semi-definite is a valid covariance matrix:

If $Z \sim N_d(0_d, I_d)$ then $X = \mu + \Sigma^{\frac{1}{2}} Z \sim N_d(\mu, \Sigma)$.

Where $\Sigma^{\frac{1}{2}}$ is a matrix square root of Σ.

- Gives one way of generating multivariate Gaussians.

Property 2: Conditional distributions are still Gaussian

Suppose

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Property 2: Conditional distributions are still Gaussian

Suppose

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then

$$X_2 \mid X_1 = x_1 \sim N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Property 2: Conditional distributions are still Gaussian

Suppose

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then

$$X_2 \mid X_1 = x_1 \sim N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Proof:

$$\pi(x_2|x_1) = \frac{\pi(x_1, x_2)}{\pi(x_1)} \propto \pi(x_1, x_2)$$

Proof:

$$\pi(x_2|x_1) = \frac{\pi(x_1, x_2)}{\pi(x_1)} \propto \pi(x_1, x_2)$$

$$\propto \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

$$= \exp(-\frac{1}{2}\left[\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right)^\top \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \cdots\right]$$

where

$$\Sigma^{-1} := Q := \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$$

Proof:

$$
\begin{aligned}
\pi(x_2|x_1) &= \frac{\pi(x_1,x_2)}{\pi(x_1)} \propto \pi(x_1,x_2) \\
&\propto \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \\
&= \exp(-\frac{1}{2}\left[\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right)^\top \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \cdots \right] \\
&\propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right)
\end{aligned}
$$

where

$$
\Sigma^{-1} := Q := \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}
$$

Proof:

$$
\begin{aligned}
\pi(x_2|x_1) &= \frac{\pi(x_1, x_2)}{\pi(x_1)} \propto \pi(x_1, x_2) \\
&\propto \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \\
&= \exp(-\frac{1}{2}\left[\left(\left(\begin{array}{c} x_1 \\ x_2 \end{array}\right) - \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right)\right)^\top \left(\begin{array}{cc} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{array}\right) \cdots\right] \\
&\propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right)
\end{aligned}
$$

where

$$
\Sigma^{-1} := Q := \left(\begin{array}{cc} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{array}\right)
$$

So $X_2|X_1 = x_1$ is Gaussian.

$$\pi(x_2|x_1) \propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right)$$

$$
\begin{aligned}
\pi(x_2|x_1) &\propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top \color{red}{Q_{21}(x_1-\mu_1)}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[x_2^\top Q_{22}x_2 - 2x_2^\top(Q_{22}\mu_2 + \color{red}{Q_{21}(x_1-\mu_1)})\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
\pi(x_2|x_1) &\propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[x_2^\top Q_{22}x_2 - 2x_2^\top(Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left(x_2 - Q_{22}^{-1}(Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right)^\top Q_{22}\left(x_2 - \ldots\right)\right)
\end{aligned}
$$

$$\pi(x_2|x_1) \propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right)$$
$$\propto \exp\left(-\frac{1}{2}\left[x_2^\top Q_{22}x_2 - 2x_2^\top (Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right]\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(x_2 - Q_{22}^{-1}(Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right)^\top Q_{22}\left(x_2 - \ldots\right)\right)$$

So

$$X_2|X_1 = x_1 \sim N(\mu_2 + Q_{22}^{-1}Q_{21}(x_1-\mu_1), Q_{22})$$

$$
\begin{aligned}
\pi(x_2|x_1) &\propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\left[x_2^\top Q_{22}x_2 - 2x_2^\top(Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\left(x_2 - Q_{22}^{-1}(Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right)^\top Q_{22}\left(x_2 - \ldots\right)\right)
\end{aligned}
$$

So

$$
X_2|X_1 = x_1 \sim N(\mu_2 + Q_{22}^{-1}Q_{21}(x_1-\mu_1), Q_{22})
$$

A simple matrix inversion lemma gives

$$
\begin{aligned}
Q_{22}^{-1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\\
\text{and}\, Q_{22}^{-1}Q_{21} &= \Sigma_{21}\Sigma_{11}^{-1}
\end{aligned}
$$

$$\pi(x_2|x_1) \propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2) + 2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[x_2^\top Q_{22}x_2 - 2x_2^\top (Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(x_2 - Q_{22}^{-1}(Q_{22}\mu_2 + Q_{21}(x_1-\mu_1))\right)^\top Q_{22}\left(x_2 - \dots\right)\right)$$

So

$$X_2|X_1 = x_1 \sim N(\mu_2 + Q_{22}^{-1}Q_{21}(x_1-\mu_1), Q_{22})$$

A simple matrix inversion lemma gives

$$Q_{22}^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

$$\text{and}\, Q_{22}^{-1}Q_{21} = \Sigma_{21}\Sigma_{11}^{-1}$$

giving

$$X_2|X_1 = x_1 \sim N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Conditional updates of Gaussian processes

So suppose f is a Gaussian process, then

$$f(x_1), \ldots, f(x_n), f(x) \sim N(\mu, \Sigma)$$

Conditional updates of Gaussian processes

So suppose f is a Gaussian process, then

$$f(x_1), \ldots, f(x_n), f(x) \sim N(\mu, \Sigma)$$

If we observe its value at $x_1, \ldots, x_n$ then

$$f(x) | f(x_1), \ldots, f(x_n) \sim N(\mu^*, \sigma^*)$$

where μ^* and σ^* are as on the previous slide.

Conditional updates of Gaussian processes

So suppose f is a Gaussian process, then

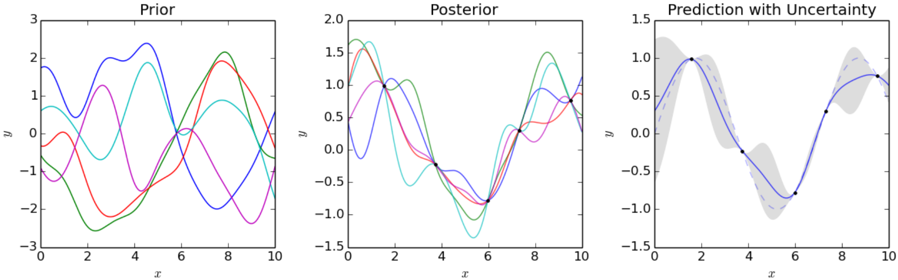
$$f(x_1), \ldots, f(x_n), f(x) \sim N(\mu, \Sigma)$$

If we observe its value at $x_1, \ldots, x_n$ then

$$f(x) | f(x_1), \ldots, f(x_n) \sim N(\mu^*, \sigma^*)$$

where μ^* and σ^* are as on the previous slide.

Note that we still believe f is a GP even though we've observed its value at a number of locations.



Why use GPs? Answer 1

The GP class of models is closed under various operations.

Why use GPs? Answer 1

The GP class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

Why use GPs? Answer 1

The GP class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

- Closed under Bayesian conditioning, i.e., if we observe

$$\mathbf{D} = (f(x_1), \ldots, f(x_n))$$

then

$$f|D \sim GP$$

but with updated mean and covariance functions.

Why use GPs? Answer 1

The GP class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

- Closed under Bayesian conditioning, i.e., if we observe

$$\mathbf{D} = (f(x_1), \ldots, f(x_n))$$

  then

$$f|D \sim GP$$

  but with updated mean and covariance functions.

- Closed under any linear operator. If $f \sim GP(m(\cdot), k(\cdot, \cdot))$, then if $\mathcal{L}$ is a linear operator

$$\mathcal{L} \circ f \sim GP(\mathcal{L} \circ m, \mathcal{L}^2 \circ k)$$

  e.g. $\frac{df}{dx}$, $\int f(x)dx$, Af are all GPs

Determining the mean and covariance function

How do we determine the mean $\mathbb{E}(f(x))$ and covariance $\mathbb{C}\text{ov}(f(x), f(x'))$?

Determining the mean and covariance function

How do we determine the mean $\mathbb{E}(f(x))$ and covariance $\mathbb{C}\text{ov}(f(x), f(x'))$?
Simplest answer is to pick values we like (found by trial and error) subject to 'the rules':

Determining the mean and covariance function

How do we determine the mean $\mathbb{E}(f(x))$ and covariance $\mathbb{C}\text{ov}(f(x), f(x'))$? Simplest answer is to pick values we like (found by trial and error) subject to 'the rules':

- We can use any mean function we want:

$$m(x) = \mathbb{E}(f(x))$$

  Most popular choices are $m(x) = 0$ or $m(x) = a$ for all x, or $m(x) = a + bx$

Determining the mean and covariance function

How do we determine the mean $\mathbb{E}(f(x))$ and covariance $\mathbb{C}\text{ov}(f(x), f(x'))$? Simplest answer is to pick values we like (found by trial and error) subject to 'the rules':

- We can use any mean function we want:

  $$m(x) = \mathbb{E}(f(x))$$

  Most popular choices are $m(x) = 0$ or $m(x) = a$ for all x, or $m(x) = a + bx$
- If mean is a linear combination of known regressor functions, e.g.,

  $$m(x) = \beta h(x) \text{ for known } h(x)$$

  and $\beta \sim N(\cdot, \cdot)$ is given a normal prior (including $\pi(\beta) \propto 1$), then $f|D, \beta \sim GP$ and

  $$f|D \sim GP$$

  with slightly modified mean and variance formulas.

Covariance functions

- We usually use a covariance function that is a function of distance between the locations

$$k(x, x') = \mathbb{C}\text{ov}(f(x), f(x')),$$

which has to be positive semi-definite, i.e., lead to valid covariance matrices.

Covariance functions

- We usually use a covariance function that is a function of distance between the locations

$$k(x, x') = \mathbb{C}\text{ov}(f(x), f(x')),$$

which has to be positive semi-definite, i.e., lead to valid covariance matrices.

  - This can be problematic (see Nicolas' talk)

Covariance functions

- We usually use a covariance function that is a function of distance between the locations

$$k(x, x') = \mathbb{C}\mathrm{ov}(f(x), f(x')),$$

which has to be positive semi-definite, i.e., lead to valid covariance matrices.

  - This can be problematic (see Nicolas' talk)

- If

$$k(x, x') = \sigma^2 c(x, x')$$

and we give σ^2 an inverse gamma prior (including $\pi(\sigma^2) \propto 1/\sigma^2$) then $f|D, \sigma^2 \sim GP$ and

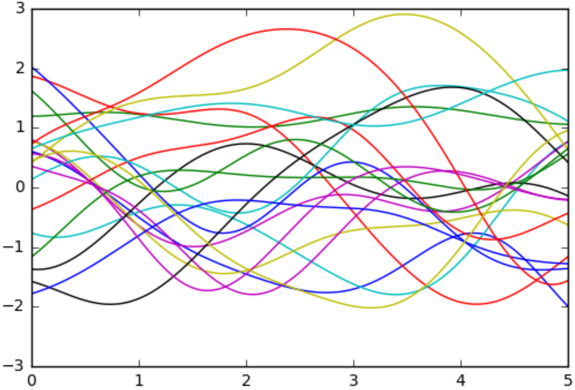
$$f|D \sim \text{t-process}$$

with $n - p$ degrees of freedom. In practice, for reasonable n, this is indistinguishable from a GP.

Examples

RBF/Squared-exponential/exponentiated quadratic

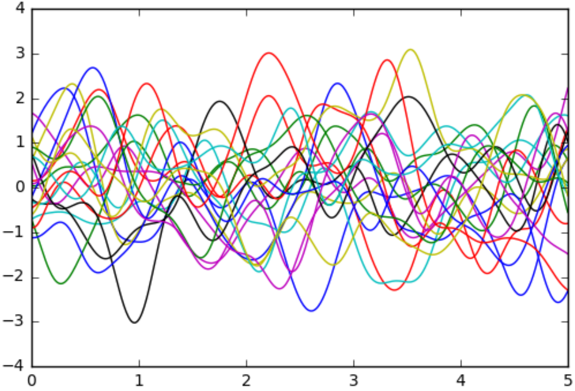
$$k(x, x') = \exp\left(-\frac{1}{2}(x - x')^2\right)$$



Examples

RBF/Squared-exponential/exponentiated quadratic

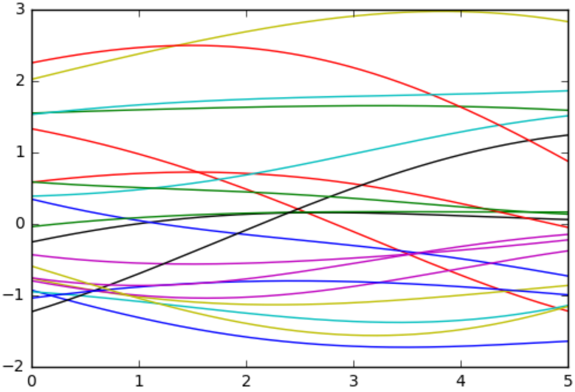
$$k(x, x') = \exp\left(-\frac{1}{2}\frac{(x - x')^2}{0.25^2}\right)$$



Examples

RBF/Squared-exponential/exponentiated quadratic

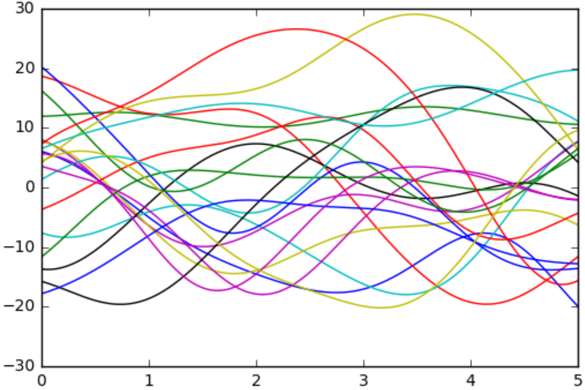
$$k(x, x') = \exp\left(-\frac{1}{2}\frac{(x - x')^2}{4^2}\right)$$



Examples

RBF/Squared-exponential/exponentiated quadratic

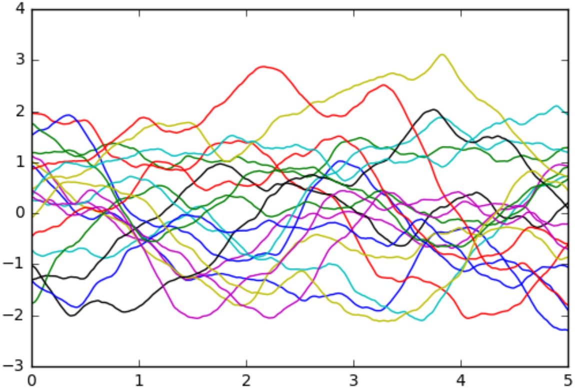
$$k(x, x') = 100 \exp\left(-\frac{1}{2}(x - x')^2\right)$$



Examples

Matern 3/2

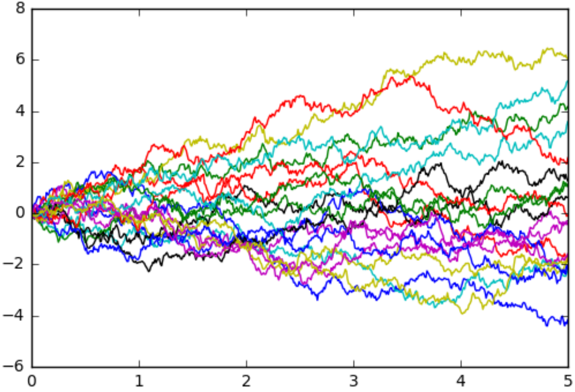
$$k(x, x') \sim (1 + |x - x'|) \exp\left(-|x - x'|\right)$$



Examples

Brownian motion

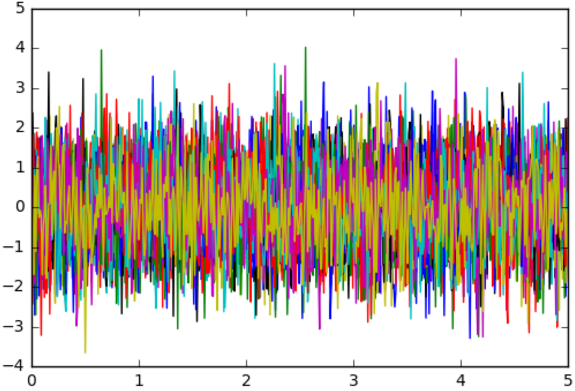
$$k(x, x') = \min(x, x')$$



Examples

White noise

$$k(x, x') = \begin{cases} 1 \text{ if } x = x' \\ 0 \text{ otherwise} \end{cases}$$



Examples

The GP inherits its properties primarily from the covariance function k.

- Smoothness
- Differentiability
- Variance

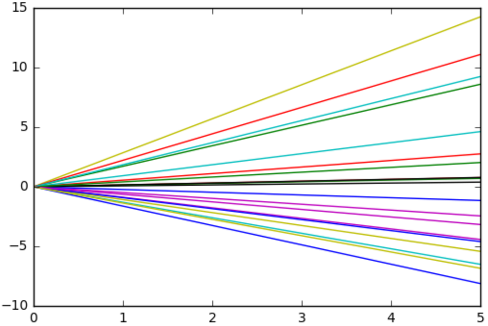
Examples

The GP inherits its properties primarily from the covariance function k.

- Smoothness
- Differentiability
- Variance

A final example

$$k(x, x') = x^\top x'$$



What is happening?

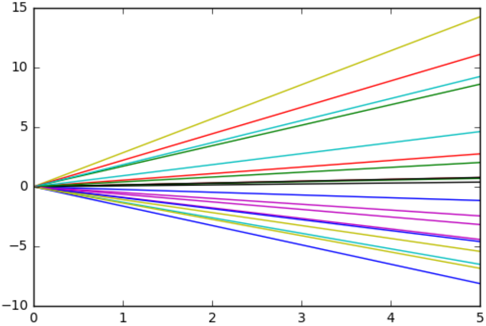
Examples

The GP inherits its properties primarily from the covariance function k.

- Smoothness
- Differentiability
- Variance

A final example

$$k(x, x') = x^\top x'$$



What is happening?
Suppose $f(x) = cx$ where $c \sim N(0, 1)$.

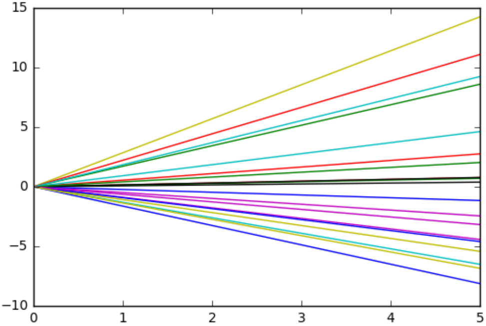
Examples

The GP inherits its properties primarily from the covariance function k.

- Smoothness
- Differentiability
- Variance

A final example

$$k(x, x') = x^\top x'$$



What is happening?

Suppose $f(x) = cx$ where $c \sim N(0, 1)$.

Then

$$\begin{aligned}\mathbb{C}\mathrm{ov}(f(x), f(x')) &= \mathbb{C}\mathrm{ov}(cx, cx') \\ &= x^\top \mathbb{C}\mathrm{ov}(c, c) x' \\ &= x^\top x'\end{aligned}$$

Conditional updates of Gaussian processes - revisited

So suppose f is a Gaussian process, then

$$f(x_1), \ldots, f(x_n), f(x) \sim N(0, \Sigma)$$

where

$$\Sigma = \begin{pmatrix} & & & & k(x_1, x) \\ & K & & & k(x_2, x) \\ & & & & \vdots \\ & & & & k(x_n, x) \\ k(x, x_1) & k(x, x_2) & \ldots & k(x, x_n) & k(x, x) \end{pmatrix}$$

where $K_{ij} = k(x_i, x_j)$ is the Gram/kernel matrix.

Conditional updates of Gaussian processes - revisited

Then

$$f(x)|f(x_1), \ldots, f(x_n) \sim N(m(x), c(x))$$

where

$$m(x) = k(x)(K + \sigma^2 I)^{-1} y$$

with

$$k(x) = \begin{pmatrix} k(x, x_1) & k(x, x_2) & \ldots & k(x, x_n) \end{pmatrix} \in \mathbb{R}^{1 \times n}$$

and

Conditional updates of Gaussian processes - revisited

Then

$$f(x)|f(x_1), \dots, f(x_n) \sim N(m(x), c(x))$$

where

$$m(x) = k(x)(K + \sigma^2 I)^{-1} y$$

with

$$k(x) = \begin{pmatrix} k(x, x_1) & k(x, x_2) & \dots & k(x, x_n) \end{pmatrix} \in \mathbb{R}^{1 \times n}$$

and

$$c(x) = k(x, x) - k(x)(K + \sigma^2 I)^{-1} k(x)^\top$$

Conditional updates of Gaussian processes - revisited

Then

$$f(x)|f(x_1),\ldots,f(x_n) \sim N(m(x),c(x))$$

where

$$m(x) = k(x)(K+\sigma^2 I)^{-1}y$$

with

$$k(x) = \begin{pmatrix} k(x,x_1) & k(x,x_2) & \ldots & k(x,x_n) \end{pmatrix} \in \mathbb{R}^{1\times n}$$

and

$$c(x) = k(x,x) - k(x)(K+\sigma^2 I)^{-1}k(x)^\top$$

Cf

$$X_2|X_1 = x_1 \sim N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Why use GPs? Answer 2: non-parametric/kernel regression

We can also view GPs as a non-parametric extension to linear regression.

k determines the space of functions that sample paths live in.

Why use GPs? Answer 2: non-parametric/kernel regression

We can also view GPs as a non-parametric extension to linear regression.

k determines the space of functions that sample paths live in.
Suppose we're given data $\{(x_i, y_i)_{i=1}^n\}$.
Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

Why use GPs? Answer 2: non-parametric/kernel regression

We can also view GPs as a non-parametric extension to linear regression.

k determines the space of functions that sample paths live in.
Suppose we're given data $\{(x_i, y_i)_{i=1}^n\}$.
Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$
\begin{aligned}
\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\
&= (X^\top X + \sigma^2 I)^{-1} X^\top y
\end{aligned}
$$

$$
\text{where} \quad X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix}
$$

Why use GPs? Answer 2: non-parametric/kernel regression

We can also view GPs as a non-parametric extension to linear regression.

k determines the space of functions that sample paths live in.
Suppose we're given data $\{(x_i, y_i)_{i=1}^n\}$.
Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= (X^\top X + \sigma^2 I)^{-1} X^\top y \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

$$\text{where} \quad X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix}$$

Why use GPs? Answer 2: non-parametric/kernel regression

We can also view GPs as a non-parametric extension to linear regression.

k determines the space of functions that sample paths live in.
Suppose we're given data $\{(x_i, y_i)_{i=1}^n\}$.
Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}
\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\
&= (X^\top X + \sigma^2 I)^{-1} X^\top y \\
&= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}
\end{aligned}$$

$$\text{as} \qquad (X^\top X + \sigma^2 I) X^\top = X^\top (XX^\top + \sigma^2 I)$$

$$\text{so} \qquad X^\top (XX^\top + \sigma^2 I)^{-1} = (X^\top X + \sigma^2 I)^{-1} X^\top$$

$$\text{where} \qquad X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix}$$

At first the dual form

$$\hat{\beta} = X^\top (XX^\top + \sigma^2 I)^{-1} y$$

looks harder to compute than the usual

$$\hat{\beta} = (X^\top X + \sigma^2 I)^{-1} X^\top y$$

- $X^\top X$ is $p \times p$ $\quad$ $p =$ number of features/parameters
- $XX^\top$ is $n \times n$ $\quad$ n is the number of data points

At first the dual form

$$\hat{\beta} = X^\top (XX^\top + \sigma^2 I)^{-1} y$$

looks harder to compute than the usual

$$\hat{\beta} = (X^\top X + \sigma^2 I)^{-1} X^\top y$$

- $X^\top X$ is $p \times p$ $\quad$ $p =$ number of features/parameters
- $XX^\top$ is $n \times n$ $\quad$ n is the number of data points

But the dual form only uses inner products.

$$XX^\top = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} (x_1 \dots x_n) = \begin{pmatrix} x_1^\top x_1 & \dots & x_1^\top x_n \\ \vdots & & \\ x_n^\top x_1 & \dots & x_n^\top x_n \end{pmatrix} = K$$

— This is useful!

Prediction

The best prediction of y at a new location x' is

$$\begin{aligned} \hat{y}' &= x'^{\top}\hat{\beta} \\ &= x'^{\top}X^{\top}(XX^{\top} + \sigma^2 I)^{-1}y \\ &= k(x')(K + \sigma^2 I)^{-1}y \end{aligned}$$

where $k(x') := (x'^{\top}x_1, \ldots, x'^{\top}x_n)$ and $K_{ij} := x_i^{\top}x_j$

Prediction

The best prediction of y at a new location x' is

$$\begin{aligned}\hat{y}' &= x'^\top \hat{\beta} \\ &= x'^\top X^\top (XX^\top + \sigma^2 I)^{-1} y \\ &= k(x')(K + \sigma^2 I)^{-1} y\end{aligned}$$

where $k(x') := (x'^\top x_1, \ldots, x'^\top x_n)$ and $K_{ij} := x_i^\top x_j$
K and $k(x)$ are kernel matrices

- every element is an inner product btwn 2 points.

Prediction

The best prediction of y at a new location x' is

$$\begin{aligned}\hat{y}' &= x'^\top \hat{\beta} \\ &= x'^\top X^\top (XX^\top + \sigma^2 I)^{-1} y \\ &= k(x')(K + \sigma^2 I)^{-1} y\end{aligned}$$

where $k(x') := (x'^\top x_1, \ldots, x'^\top x_n)$ and $K_{ij} := x_i^\top x_j$
K and $k(x)$ are kernel matrices

- every element is an inner product btwn 2 points.

Note this is exactly the GP conditional mean we derived before.

$$m(x) = t(x)(K + \sigma^2 I)^{-1} y$$

- linear regression and GP regression are equivalent when $k(x, x') = x^\top x'$.

- We know that we can replace x by a feature vector in linear regression, e.g., $\phi(x) = (1 \; x \; x^2)$ etc.
  Then
  $$K_{ij} = \phi(x_i)^\top \phi(x_j) \qquad \text{etc}$$

- For some sets of features, the inner product is equivalent to evaluating a kernel function

$$\phi(x)^\top \phi(x') \equiv k(x, x')$$

where

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is a semi-positive definite function.

- For some sets of features, the inner product is equivalent to evaluating a kernel function

$$\phi(x)^\top \phi(x') \equiv k(x, x')$$

where

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is a semi-positive definite function.
Example: If (modulo some detail)

$$\phi(x) = (e^{-\frac{(x-c_1)^2}{2\lambda^2}}, \ldots, e^{-\frac{(x-c_N)^2}{2\lambda^2}})$$

then as $N \to \infty$ then

$$\phi(x)^\top \phi(x) = \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$$

- For some sets of features, the inner product is equivalent to evaluating a kernel function

$$\phi(x)^\top \phi(x') \equiv k(x, x')$$

where

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is a semi-positive definite function.
Example: If (modulo some detail)

$$\phi(x) = (e^{-\frac{(x-c_1)^2}{2\lambda^2}}, \ldots, e^{-\frac{(x-c_N)^2}{2\lambda^2}})$$

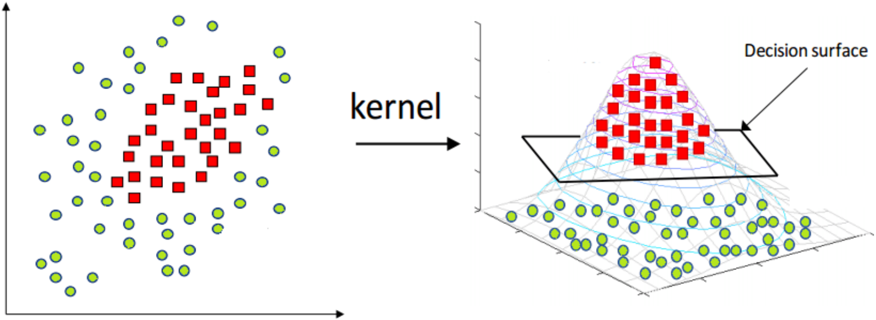
then as $N \to \infty$ then

$$\phi(x)^\top \phi(x) = \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$$

- We can use an infinite dimensional feature vector $\phi(x)$, and because linear regression can be done solely in terms of inner-products (inverting a $n \times n$ matrix in the dual form) we never need evaluate the feature vector, only the kernel.

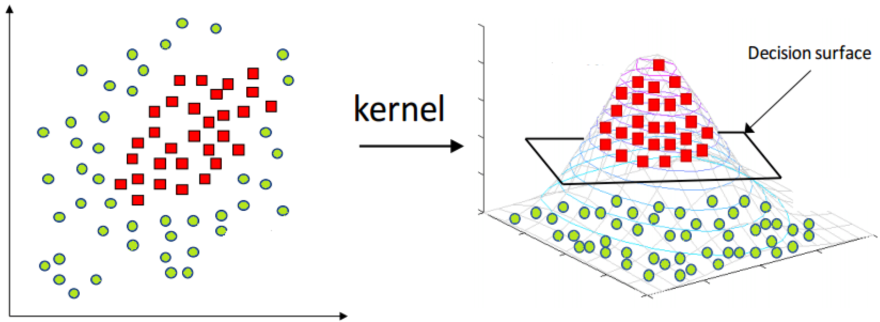
Kernel trick:

lift x into feature space by replacing inner products $x^\top x'$ by $k(x, x')$



Kernel trick:

lift x into feature space by replacing inner products $x^\top x'$ by $k(x, x')$



Kernel regression/non-parametric regression/GP regression all closely related:

$$\hat{y}' = m(x') = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

Generally, we don't think about these features, we just choose a kernel.

- $k(x, x')$ is a kernel ifF it is a positive semidefinite function

---

1

Generally, we don't think about these features, we just choose a kernel.

- $k(x, x')$ is a kernel ifF it is a positive semidefinite function

Any kernel implicitly determines a set of features (ie we can write $k(x, x') = \phi(x)^\top \phi(x')$ for some feature vector $\phi(x)$),

1

Generally, we don't think about these features, we just choose a kernel.

- $k(x, x')$ is a kernel ifF it is a positive semidefinite function

Any kernel implicitly determines a set of features (ie we can write $k(x, x') = \phi(x)^\top \phi(x')$ for some feature vector $\phi(x)$), and our model only includes functions that are linear combinations of this set of features

$$f(x) = \sum_i c_i k(x, x_i)^1$$

[1] Not quite - it lies in the completion of this set of linear combinations

Generally, we don't think about these features, we just choose a kernel.

- $k(x, x')$ is a kernel ifF it is a positive semidefinite function

Any kernel implicitly determines a set of features (ie we can write $k(x, x') = \phi(x)^\top \phi(x')$ for some feature vector $\phi(x)$), and our model only includes functions that are linear combinations of this set of features

$$f(x) = \sum_i c_i k(x, x_i)^1$$

- this space of functions is called the Reproducing Kernel Hilbert Space (RKHS) of k.

[1] Not quite - it lies in the completion of this set of linear combinations

Generally, we don't think about these features, we just choose a kernel.

- $k(x, x')$ is a kernel ifF it is a positive semidefinite function

Any kernel implicitly determines a set of features (ie we can write $k(x, x') = \phi(x)^\top \phi(x')$ for some feature vector $\phi(x)$), and our model only includes functions that are linear combinations of this set of features

$$f(x) = \sum_i c_i k(x, x_i)^1$$

- this space of functions is called the Reproducing Kernel Hilbert Space (RKHS) of k.

Although reality may not lie in the RKHS defined by k, this space is much richer than any parametric regression model (and can be dense in some sets of continuous bounded functions), and is thus more likely to contain an element close to the true functional form than any class of models that contains only a finite number of features.

This is the motivation for non-parametric methods.

---

[1] Not quite - it lies in the completion of this set of linear combinations

Why use GPs? Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

2

Why use GPs? Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

One answer might come from Bayes linear methods[2].
If we only knew the expectation and variance of some random variables, X and Y, then how should we best do statistics?

[2] Statistics without probability!

Why use GPs? Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

One answer might come from Bayes linear methods[2].
If we only knew the expectation and variance of some random variables, X and Y, then how should we best do statistics?

It has been shown, using coherency arguments, or geometric arguments, or..., that the best second-order inference we can do to update our beliefs about X given Y is

$$\mathbb{E}(X|Y) = \mathbb{E}(X) + \mathbb{C}\mathrm{ov}(X, Y)\mathbb{V}\mathrm{ar}(Y)^{-1}(Y - \mathbb{E}(Y))$$

i.e., exactly the Gaussian process update for the posterior mean.
So GPs are in some sense second-order optimal.

[2] Statistics without probability!

Why use GPs? Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

One answer might come from Bayes linear methods[2].
If we only knew the expectation and variance of some random variables, X and Y, then how should we best do statistics?

It has been shown, using coherency arguments, or geometric arguments, or..., that the best second-order inference we can do to update our beliefs about X given Y is

$$\mathbb{E}(X|Y) = \mathbb{E}(X) + \mathbb{C}\text{ov}(X, Y)\mathbb{V}\text{ar}(Y)^{-1}(Y - \mathbb{E}(Y))$$

i.e., exactly the Gaussian process update for the posterior mean.
So GPs are in some sense second-order optimal.

See also kernel Bayes and kriging/BLUP.

[2] Statistics without probability!

Why use GPs? Answer 4: Uncertainty estimates from emulators

We often think of our prediction as consisting of two parts

- point estimate
- uncertainty in that estimate

That GPs come equipped with the uncertainty in their prediction is seen as one of their main advantages.

Why use GPs? Answer 4: Uncertainty estimates from emulators

We often think of our prediction as consisting of two parts

- point estimate
- uncertainty in that estimate

That GPs come equipped with the uncertainty in their prediction is seen as one of their main advantages.

It is important to check both aspects.

Why use GPs? Answer 4: Uncertainty estimates from emulators

We often think of our prediction as consisting of two parts

- point estimate
- uncertainty in that estimate

That GPs come equipped with the uncertainty in their prediction is seen as one of their main advantages.

It is important to check both aspects.

Warning: the uncertainty estimates from a GP can be flawed. Note that given data $D = \{X, y\}$

$$\mathbb{V}\mathrm{ar}(f(x)|X, y) = k(x, x) - k(x, X)k(X, X)^{-1}k(X, x)$$

so that the posterior variance of $f(x)$ does not depend upon y!

The variance estimates are particularly sensitive to the hyper-parameter estimates.

Difficulties of using GPs

If we know what RKHS $\equiv$ what covariance function we should use, GPs work great!

Difficulties of using GPs

If we know what RKHS $\equiv$ what covariance function we should use, GPs work great!
Unfortunately, we don't usually know this.

- We pick a covariance function from a small set, based usually on differentiability considerations.

Difficulties of using GPs

If we know what RKHS $\equiv$ what covariance function we should use, GPs work great!
Unfortunately, we don't usually know this.

- We pick a covariance function from a small set, based usually on differentiability considerations.
- Possibly try a few (plus combinations of a few) covariance functions, and attempt to make a good choice using some sort of empirical evaluation.

Difficulties of using GPs

If we know what RKHS $\equiv$ what covariance function we should use, GPs work great!
Unfortunately, we don't usually know this.

- We pick a covariance function from a small set, based usually on differentiability considerations.
- Possibly try a few (plus combinations of a few) covariance functions, and attempt to make a good choice using some sort of empirical evaluation.
- Covariance functions often contain hyper-parameters. E.g
  - RBF kernel
  $$k(x, x') = \sigma^2 \exp\left(-\frac{1}{2}\frac{(x - x')^2}{\lambda^2}\right)$$

  Estimate these using some standard procedure (maximum likelihood, cross-validation, Bayes etc)

Difficulties of using GPs

Gelman *et al.* 2017

Assuming a GP model for your data imposes a complex structure on the data.

Difficulties of using GPs

Gelman *et al.* 2017

Assuming a GP model for your data imposes a complex structure on the data.

The number of parameters in a GP is essentially infinite, and so they are not identified even asymptotically.

Difficulties of using GPs

Gelman *et al.* 2017

Assuming a GP model for your data imposes a complex structure on the data.

The number of parameters in a GP is essentially infinite, and so they are not identified even asymptotically.

So the posterior can concentrate not on a point, but on some submanifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

Difficulties of using GPs

Gelman *et al.* 2017

Assuming a GP model for your data imposes a complex structure on the data.

The number of parameters in a GP is essentially infinite, and so they are not identified even asymptotically.

So the posterior can concentrate not on a point, but on some submanifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

E.g. consider a zero mean GP on $[0, 1]$ with covariance function

$$k(x, x') = \sigma^2 \exp(-\kappa^2 |x - x|)$$

We can consistently estimate $\sigma^2\kappa$, but not σ^2 or κ, even as $n \to \infty$.

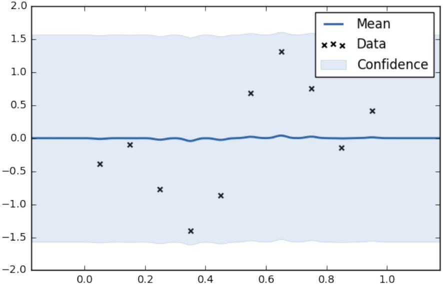
Problems with hyper-parameter optimization

As well as problems of identifiability, the likelihood surface that is being maximized is often flat and multi-modal, and thus the optimizer can sometimes fail to converge, or gets stuck in local-maxima.

Problems with hyper-parameter optimization

As well as problems of identifiability, the likelihood surface that is being maximized is often flat and multi-modal, and thus the optimizer can sometimes fail to converge, or gets stuck in local-maxima.

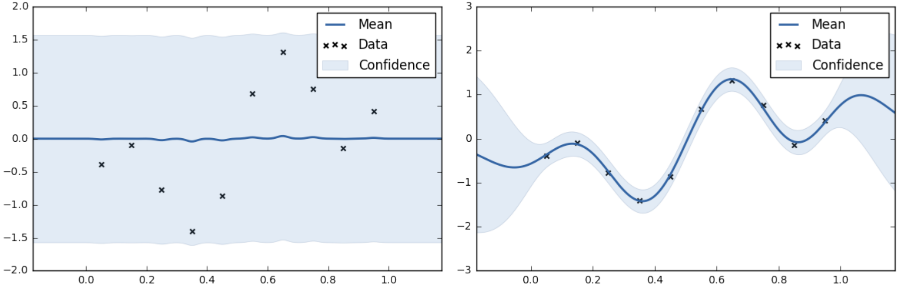
In practice, it is not uncommon to optimize hyper parameters and find solutions such as



Problems with hyper-parameter optimization

As well as problems of identifiability, the likelihood surface that is being maximized is often flat and multi-modal, and thus the optimizer can sometimes fail to converge, or gets stuck in local-maxima.

In practice, it is not uncommon to optimize hyper parameters and find solutions such as



We often work around these problems by running the optimizer multiple times from random start points, using prior distributions, constraining or fixing hyper-parameters, or adding white noise.