From Gaussian variables to Gaussian processes

Mauricio A. Álvarez

Modelos probabilísticos profundos
AIR Institute

Gaussian processes in machine learning

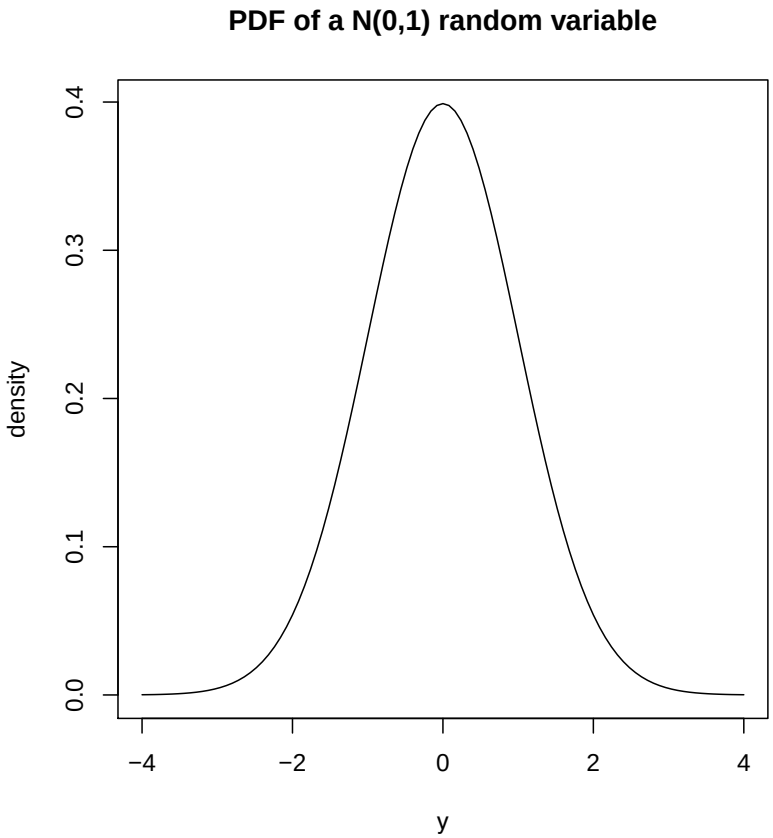
- Gaussian processes (GPs) or Gaussian random fields.
- They were introduced by George Matheron in 1960 under the name of **kriging** (geostatistics literature).
- Well known in the Statistics and Probability communities.
- Growing in populatity in machine learning since the 90s.

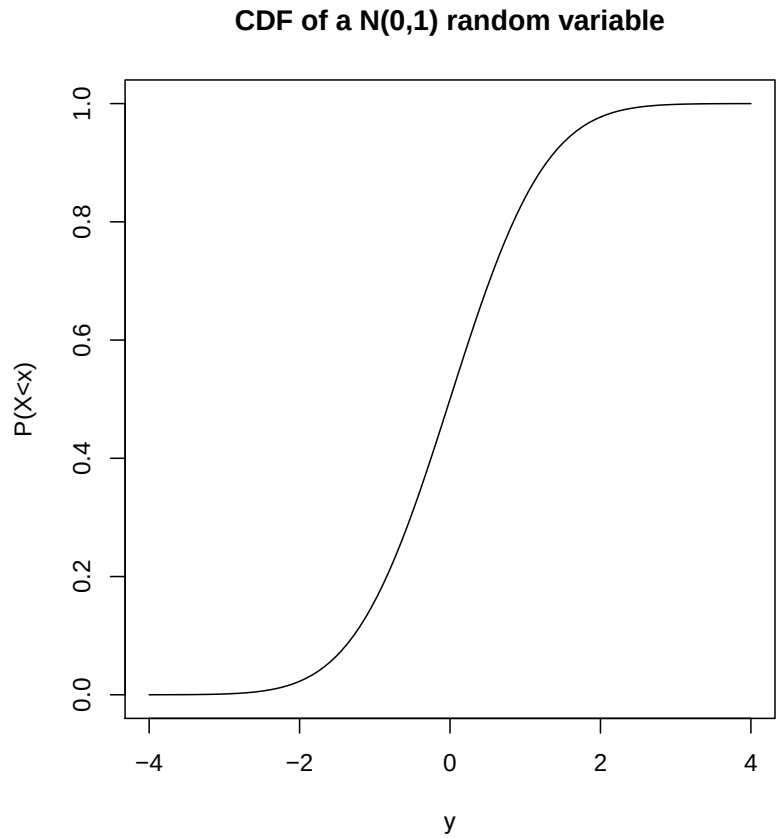
Gaussian processes in machine learning

- A Gaussian process generalises the multivariate Gaussian distribution to the infinite dimensional setting.
- Most common application is non-linear regression.
- They have been also used in pattern classification, dimensionality reduction, multi-task learning and Bayesian optimisation.

Contents

Univariate Gaussian distributions





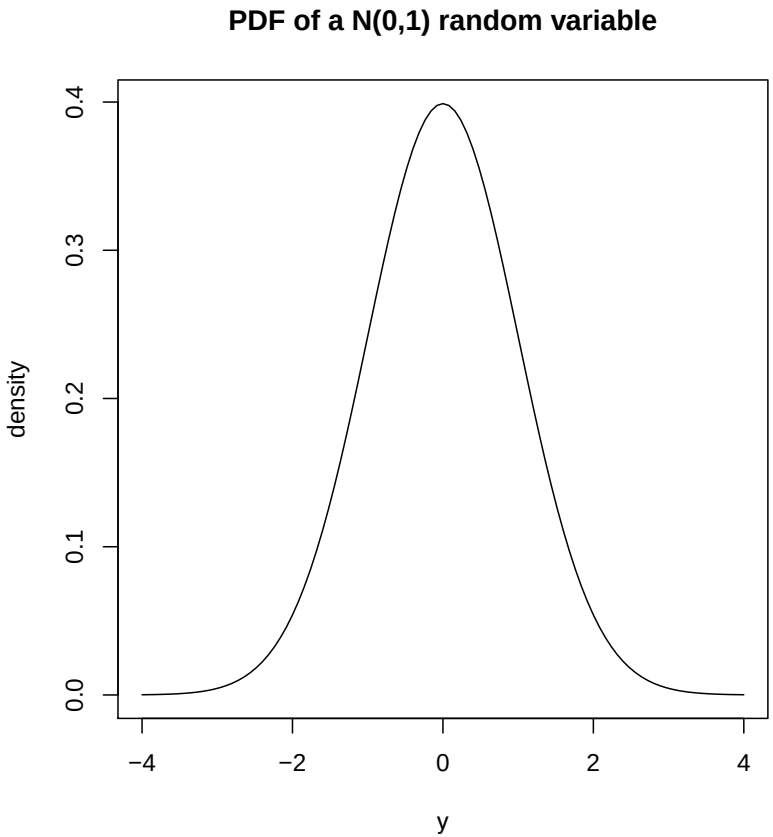
$$Y \sim N(\mu, \sigma^2)$$

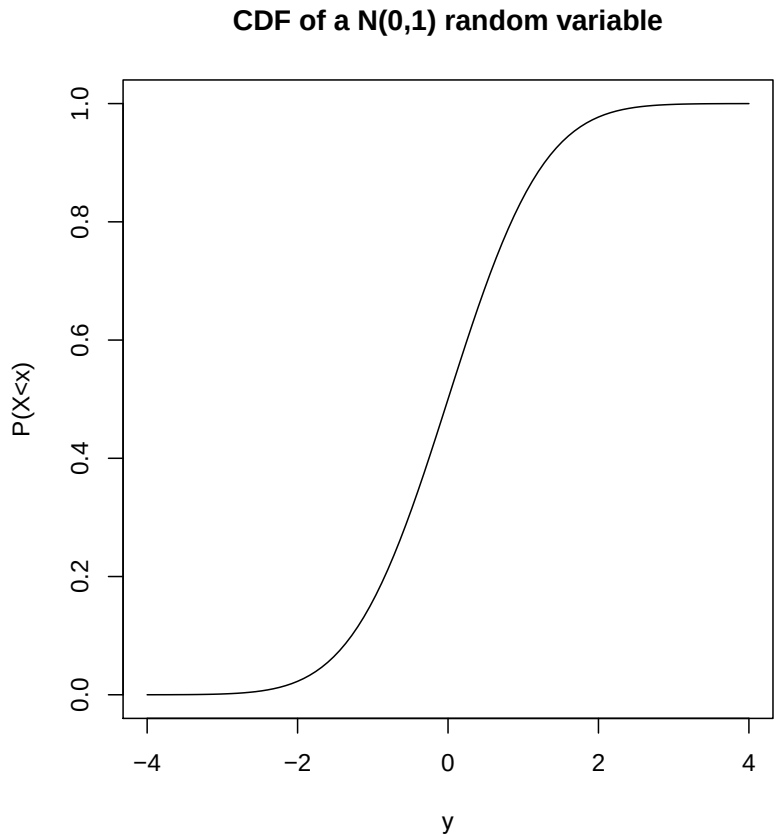
PDF: $$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

CDF: $F_Y(y) = \mathbb{P}(Y \leq y)$ not known in closed form

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

Univariate Gaussian distributions





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

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

CDF: $F_Y(y) = \mathbb{P}(Y \leq y)$ not known in closed form

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

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Central limit theorem.
- Family of normal distributions is closed under linear operations.
- If Y and Z are jointly normally distributed and are uncorrelated, then they are independent.

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Central limit theorem.
- Family of normal distributions is closed under linear operations.
- If Y and Z are jointly normally distributed and are uncorrelated, then they are independent.

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Central limit theorem.
- Family of normal distributions is closed under linear operations.
- If Y and Z are jointly normally distributed and are uncorrelated, then they are independent.

Univariate Gaussians

The normal/Gaussian distribution occurs naturally and is convenient mathematically

- Central limit theorem.
- Family of normal distributions is closed under linear operations.
- If Y and Z are jointly normally distributed and are uncorrelated, then they are independent.

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

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

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

Write

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

Bivariate Gaussian: d=2

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

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

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

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

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

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

Write

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

Bivariate Gaussian: d=2

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

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

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

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

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

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

Write

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

Bivariate Gaussian: d=2

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

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

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

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

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

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

Write

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

Bivariate Gaussian: d=2

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

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

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

Multivariate Gaussian distributions

'Multivariate' = two or more random variables

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

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

Write

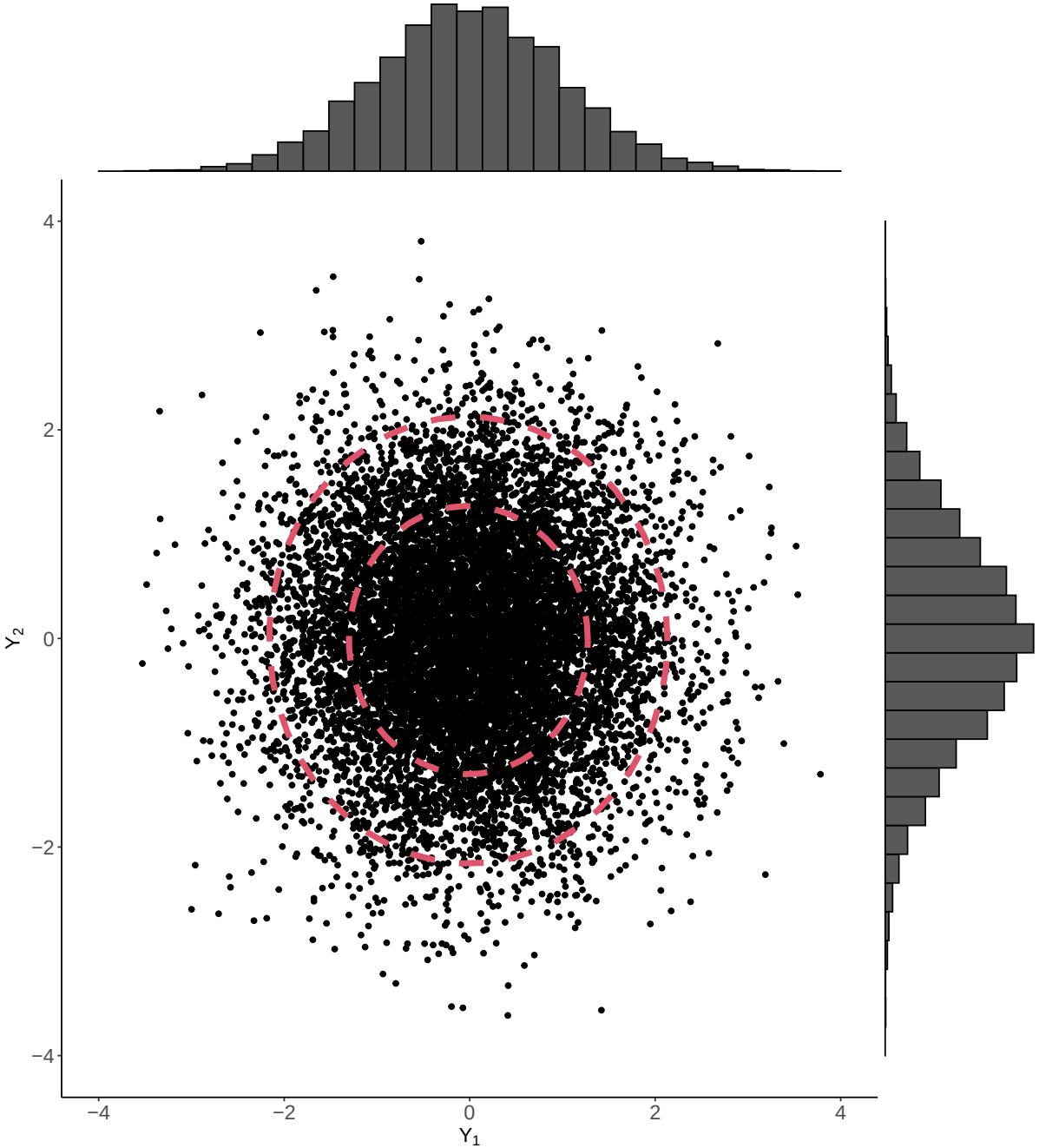
$$Y \sim N_d(\mu, \Sigma)$$

Bivariate Gaussian: d=2

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

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

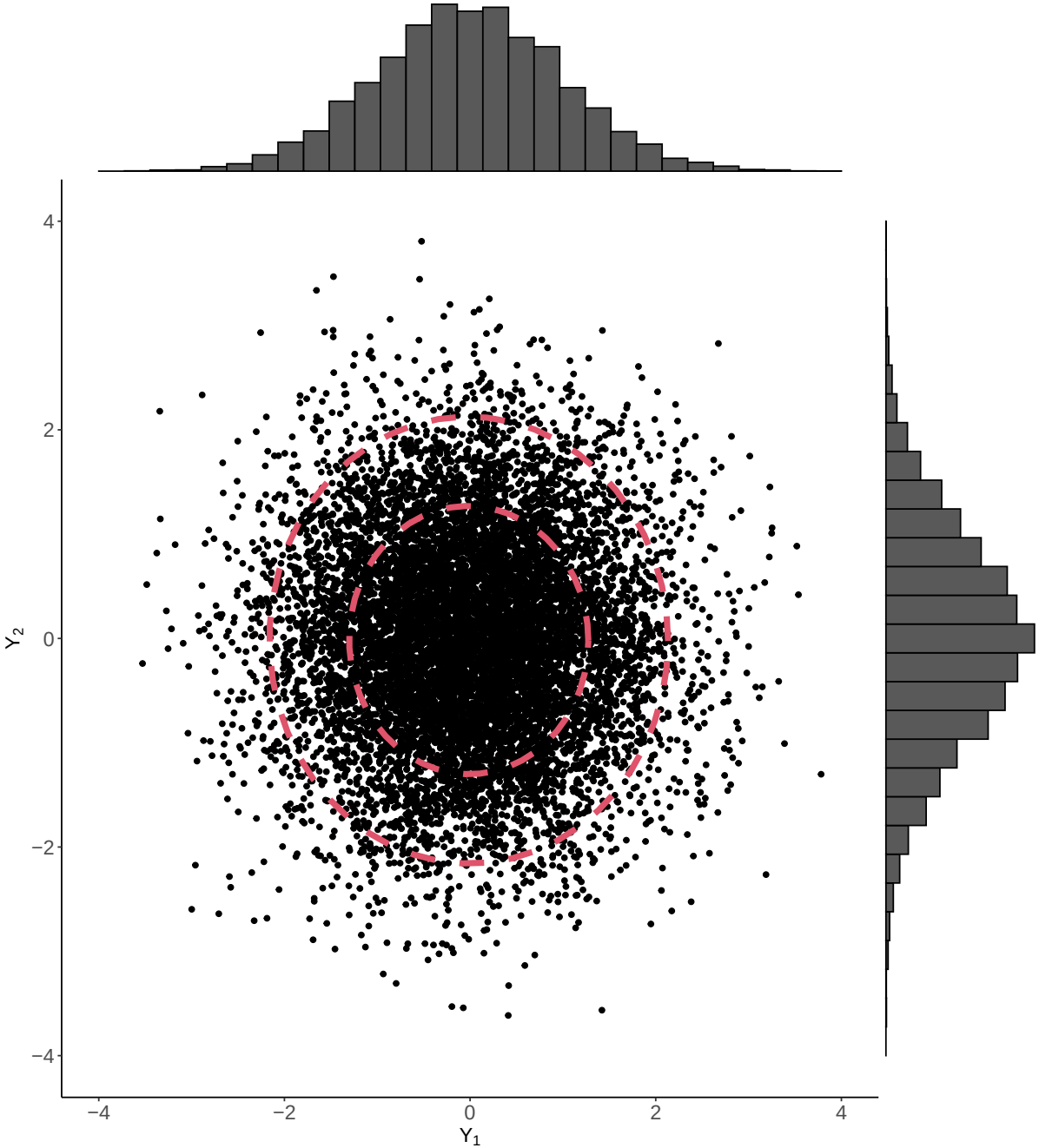
$$\text{pdf:} \quad f(y \mid \mu, \Sigma) = |\Sigma|^{-\frac{1}{2}}(2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right)$$



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

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

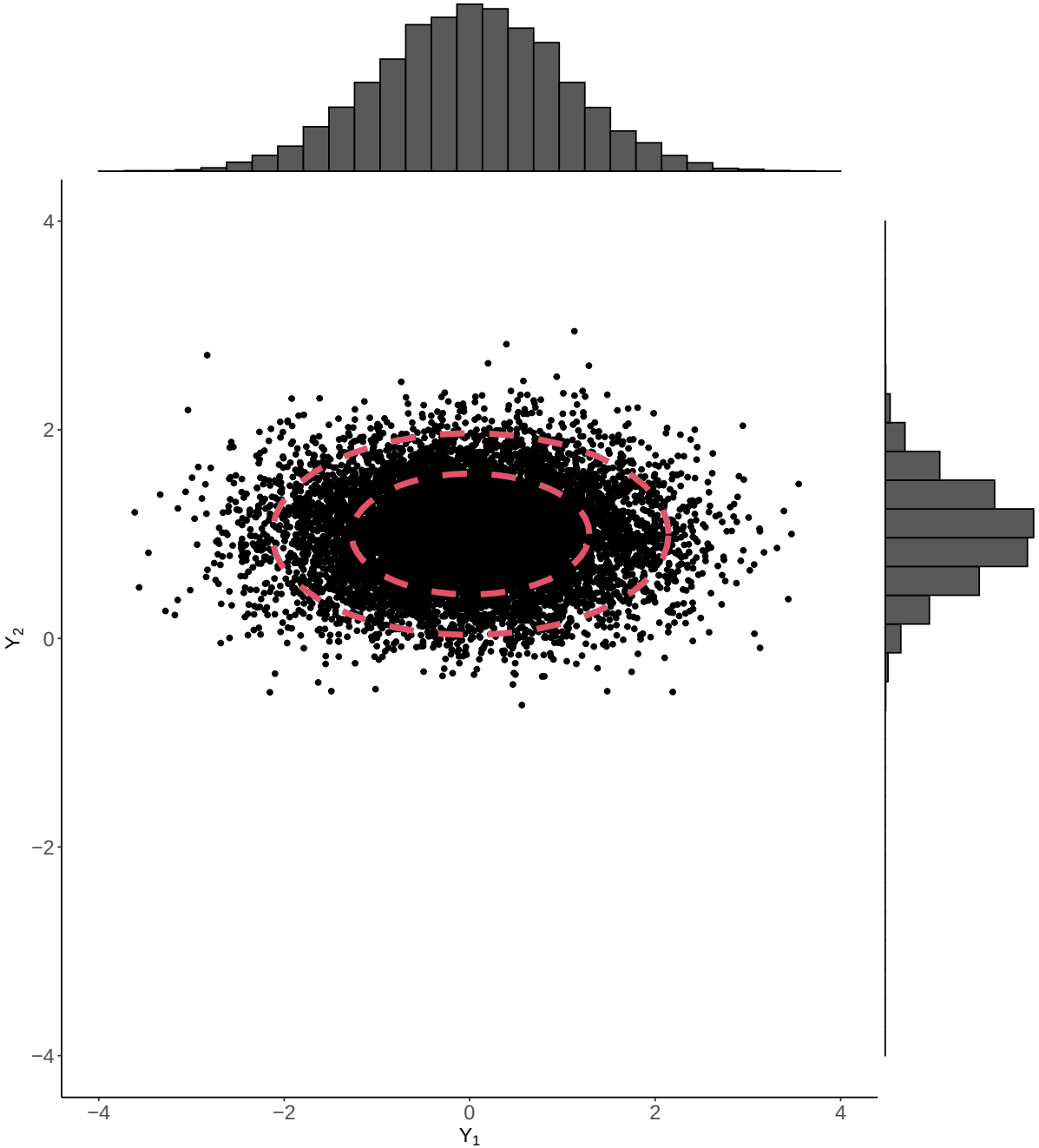
So $Cor(Y_1, Y_2) = 0$
hence Y_1
independent of Y_2



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

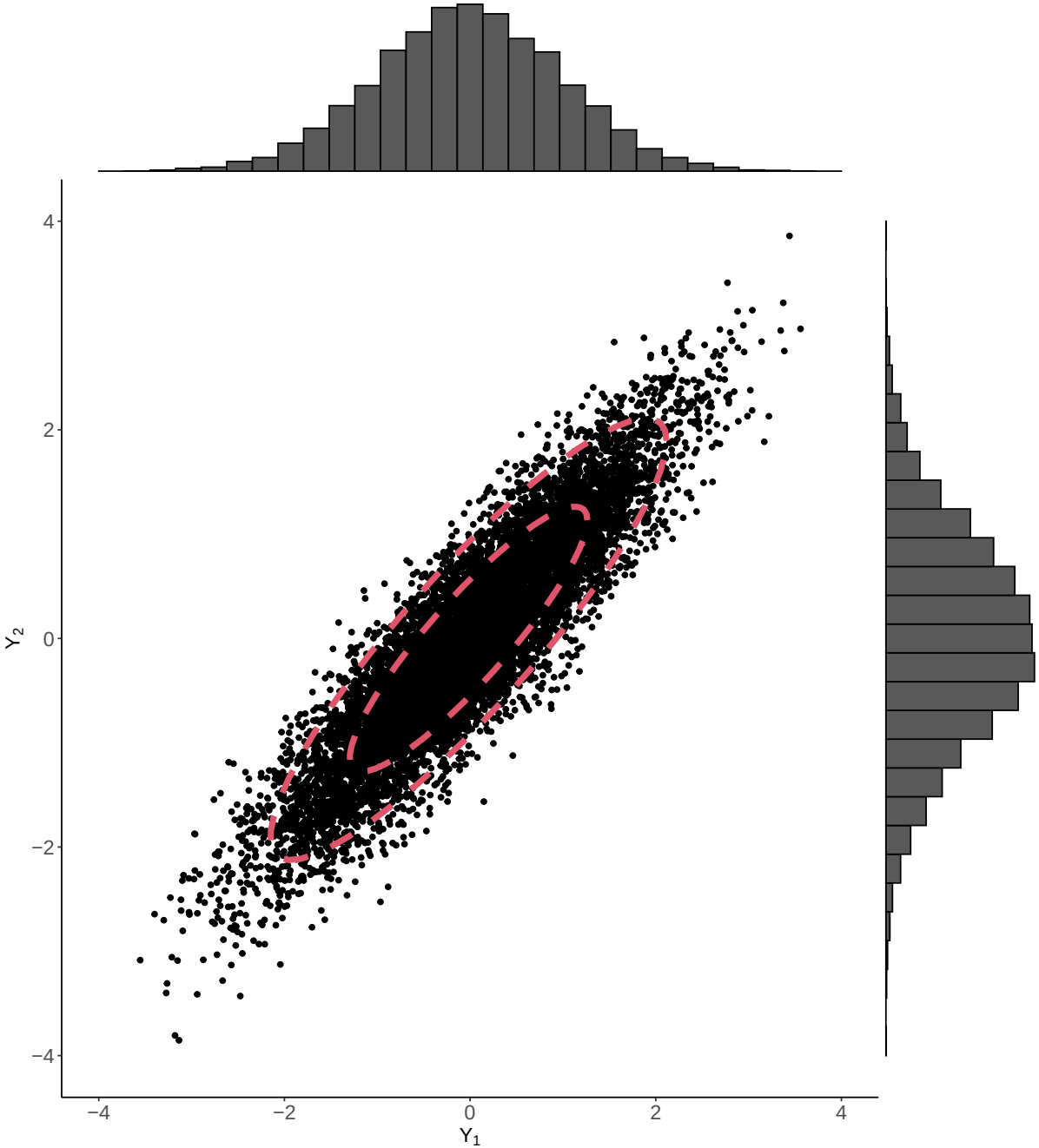
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

So $Cor(Y_1, Y_2) = 0$ hence Y_1 independent of Y_2



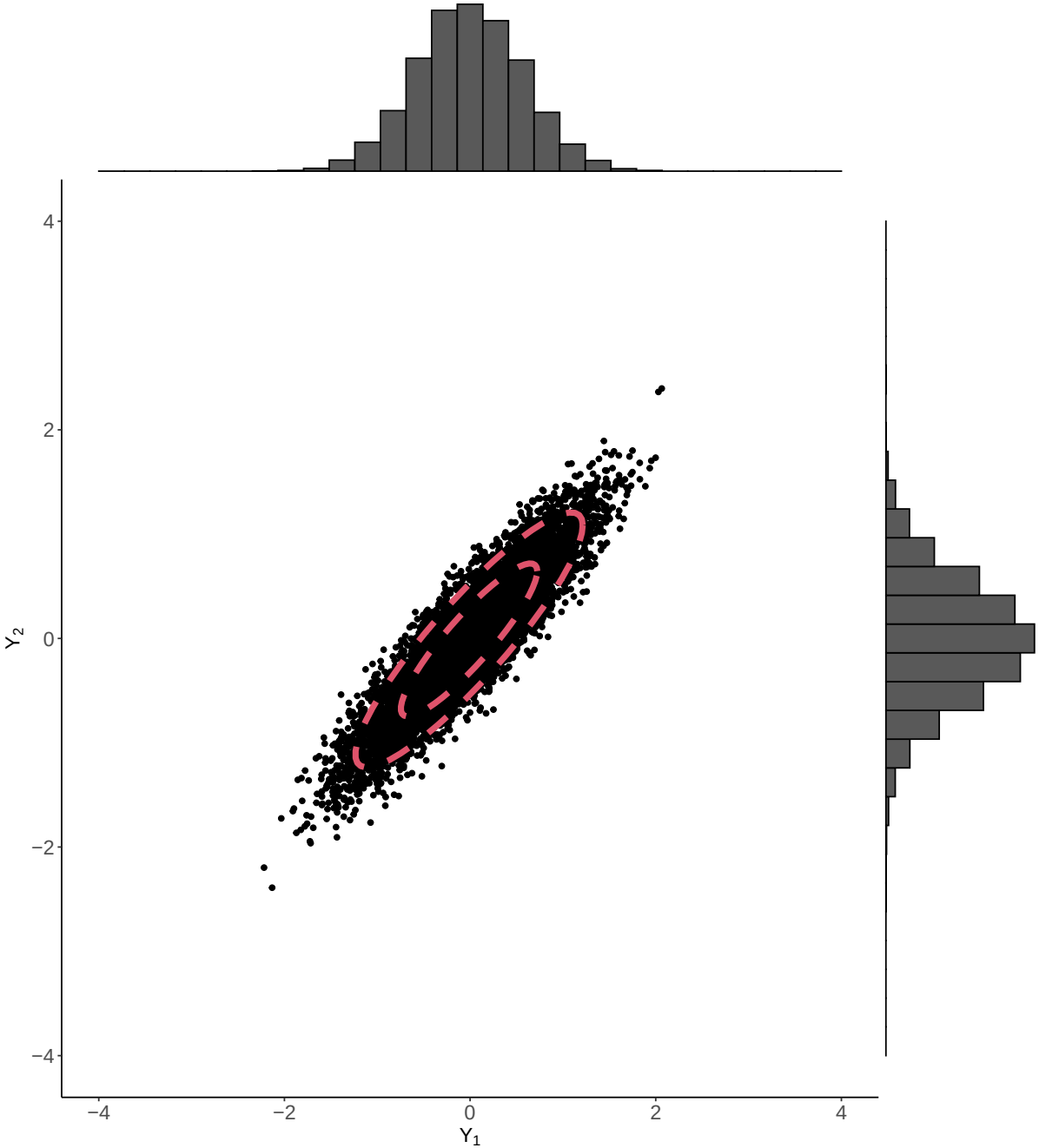
$$\mu = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

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



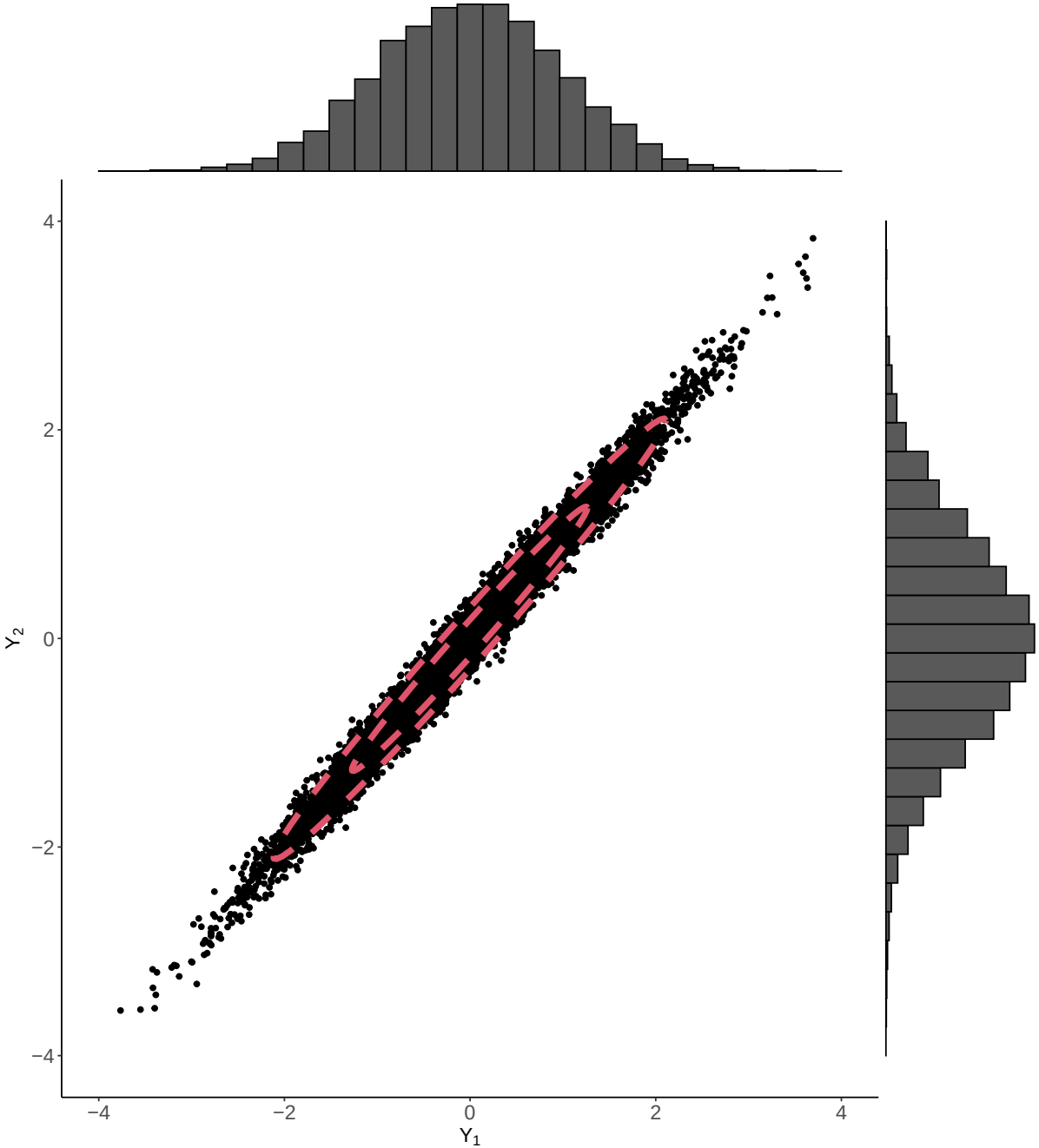
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

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



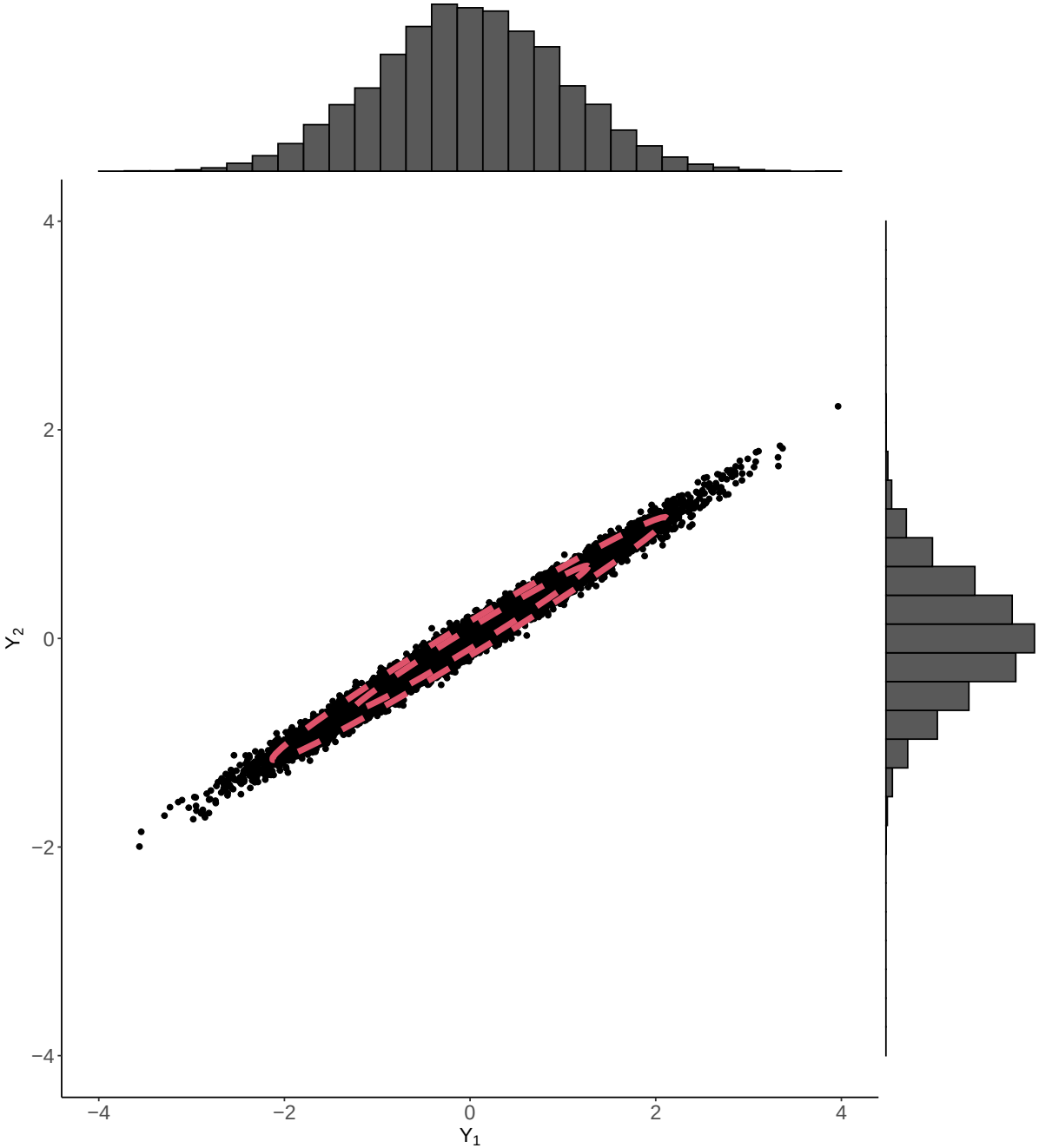
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

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



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

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

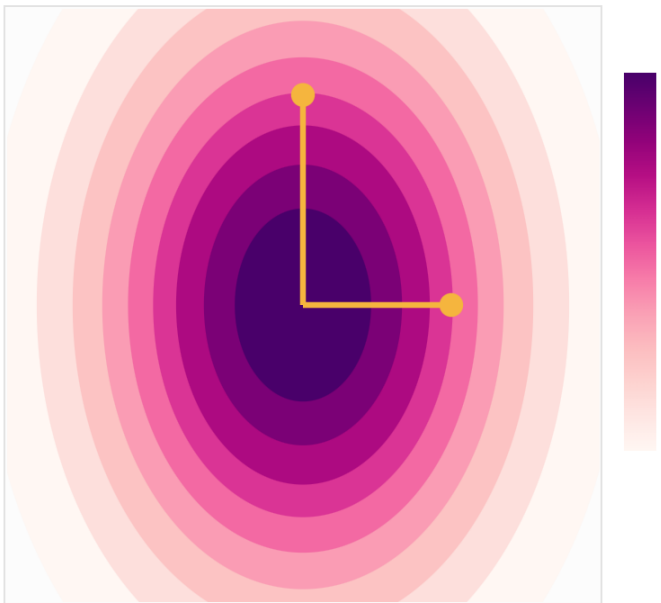


$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

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

$Cor(Y_1, Y_2) = 0.54/\sqrt{(0.3)} = 0.99$

Visual exploration



Covariance matrix (Σ)

1	0.7
0.7	2

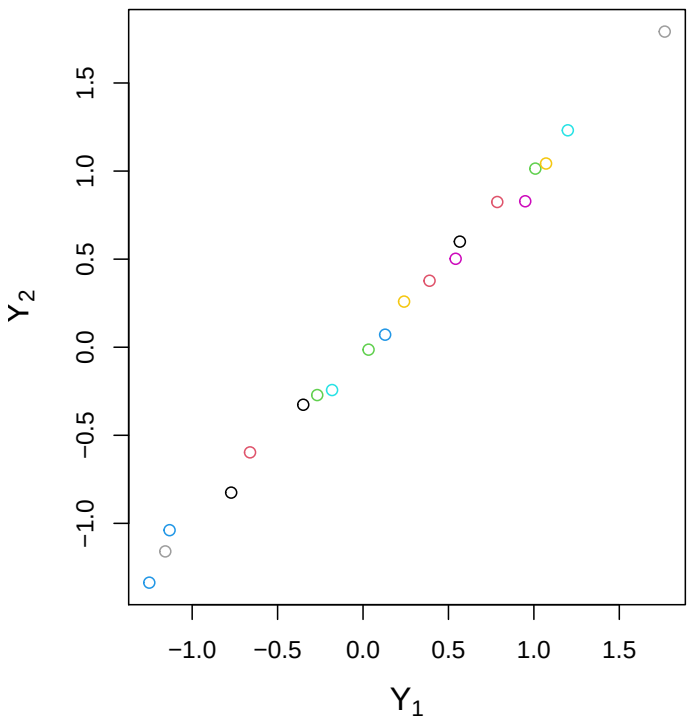
By dragging the handles you can adjust the variance along each dimension, as well as the correlation between the two random variables. *Violet* values show a high probability inside the distribution.

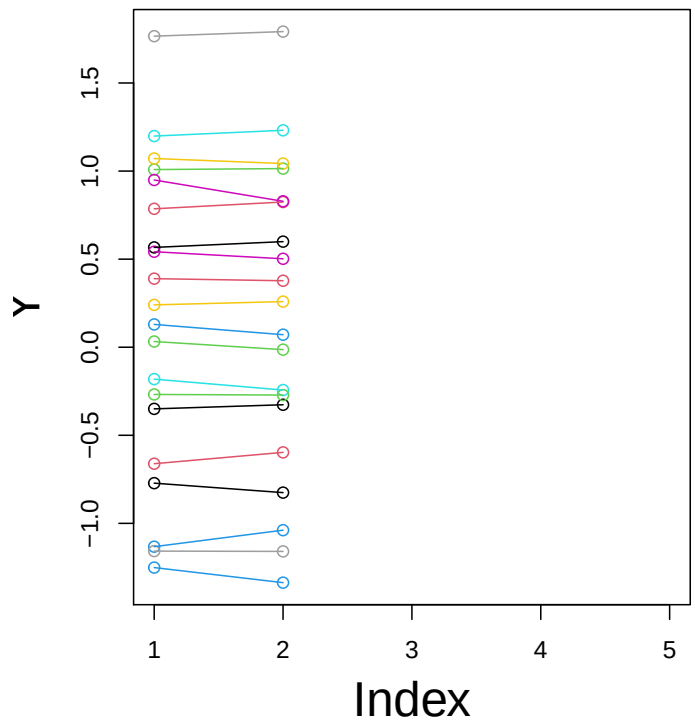
Taken from: “Visual exploration of Gaussian processes” by J Götler, R Kehlbeck and O Deussen (2019)

Visualisation in more than two dimensions

Hard to visualise in dimensions > 2, so stack points next to each other.

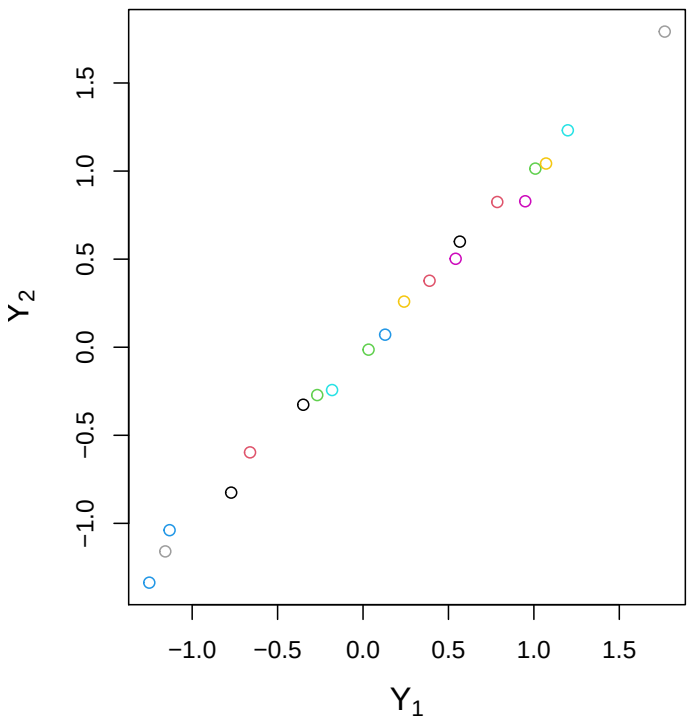
So for 2d instead of we have

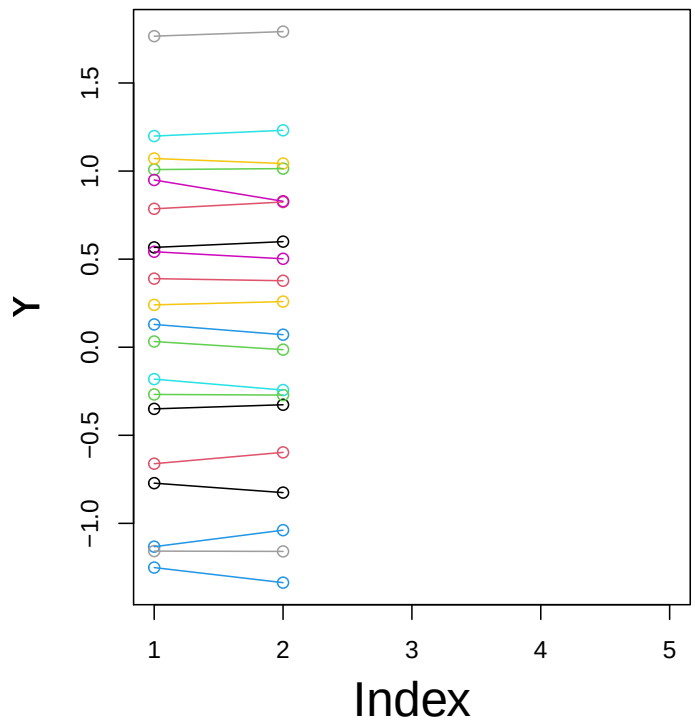




Visualisation in more than two dimensions

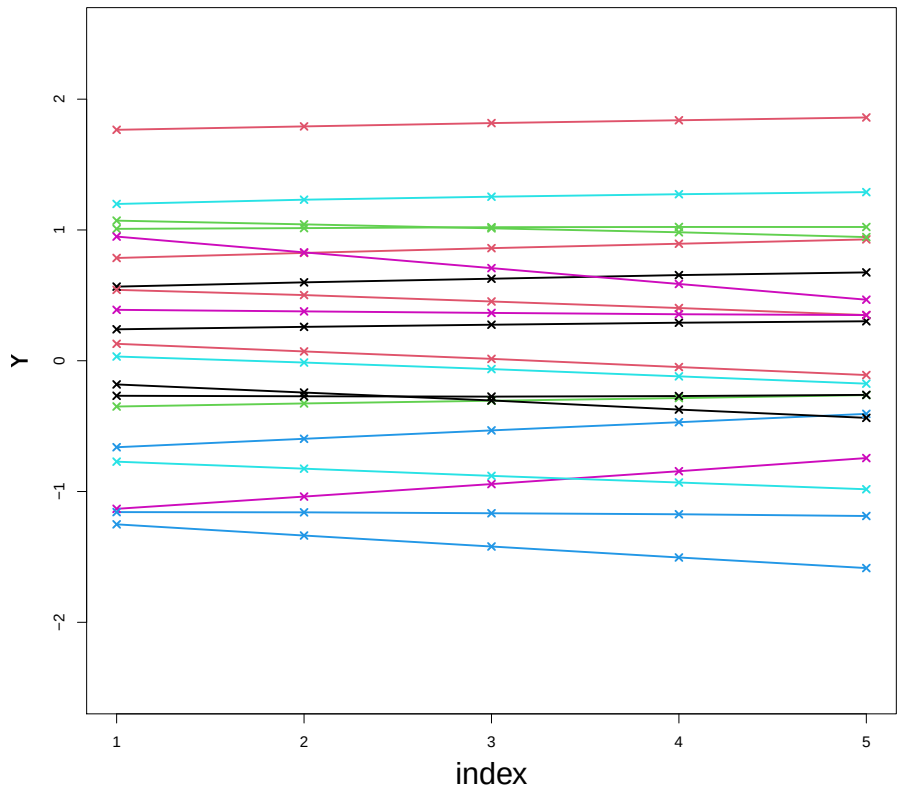
Hard to visualise in dimensions > 2, so stack points next to each other.
So for 2d instead of we have





Consider $d = 5$

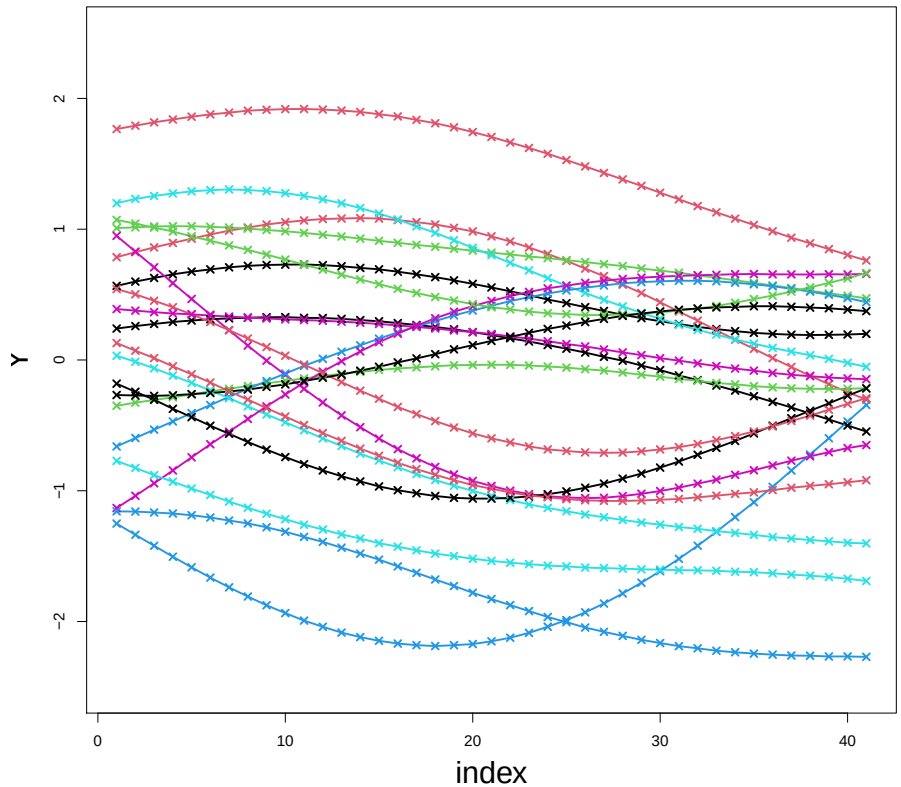
$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.99 & 0.98 & 0.97 & 0.96 \\ 0.99 & 1 & 0.99 & 0.98 & 0.97 \\ 0.98 & 0.99 & 1 & 0.99 & 0.98 \\ 0.97 & 0.98 & 0.99 & 1 & 0.99 \\ 0.96 & 0.97 & 0.98 & 0.99 & 1 \end{pmatrix}$$



Each line is one sample.

Consider $d = 50$

$$\mu = \begin{pmatrix} 0 \\ 0 \\ \\ . \\ . \\ . \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.99 & 0.98 & 0.97 & 0.96 & \dots \\ 0.99 & 1 & 0.99 & 0.98 & 0.97 & \dots \\ 0.98 & 0.99 & 1 & 0.99 & 0.98 & \dots \\ 0.97 & 0.98 & 0.99 & 1 & 0.99 & \dots \\ 0.96 & 0.97 & 0.98 & 0.99 & 1 & \dots \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{pmatrix}$$

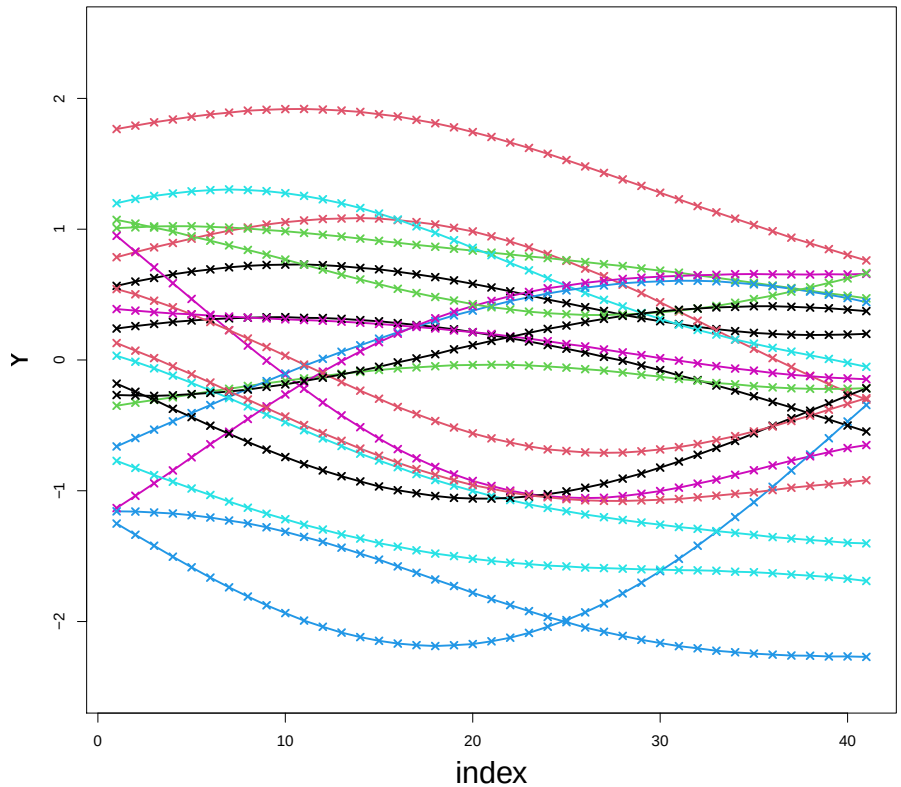


Each line is one sample.

We can think of Gaussian processes as an infinite dimensional distribution over functions - all we need to do is change the indexing

Consider $d = 50$

$$\mu = \begin{pmatrix} 0 \\ 0 \\ \\ . \\ . \\ . \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.99 & 0.98 & 0.97 & 0.96 & \dots \\ 0.99 & 1 & 0.99 & 0.98 & 0.97 & \dots \\ 0.98 & 0.99 & 1 & 0.99 & 0.98 & \dots \\ 0.97 & 0.98 & 0.99 & 1 & 0.99 & \dots \\ 0.96 & 0.97 & 0.98 & 0.99 & 1 & \dots \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{pmatrix}$$



Each line is one sample.
We can think of Gaussian processes as an infinite dimensional distribution over functions - all we need to do is change the indexing

Contents

Gaussian processes

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

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

- Usually $y(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ - think of y as a function of x.

- If $|\mathcal{X}| = \infty$, y is an infinite dimensional process.

Gaussian processes

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

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

- Usually $y(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ - think of y as a function of x.

- If $|\mathcal{X}| = \infty$, y is an infinite dimensional process.

Gaussian processes

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

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

- Usually $y(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ - think of y as a function of x.

- If $|\mathcal{X}| = \infty$, y is an infinite dimensional process.

Gaussian processes

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

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

- Usually $y(x) \in \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ - think of y as a function of x.

- If $|\mathcal{X}| = \infty$, y is an infinite dimensional process.

Gaussian processes

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

$$\mathbb{P}(y(x_1) \leq c_1, \ldots, y(x_n) \leq c_n)$$

as these uniquely determine the law of y.

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

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

Write $y(\cdot) \sim GP$ to denote that the *function* y is a GP.

Gaussian processes

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

$$\mathbb{P}(y(x_1) \leq c_1, \ldots, y(x_n) \leq c_n)$$

  as these uniquely determine the law of y.

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

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

  Write $y(\cdot) \sim GP$ to denote that the *function* y is a GP.

Gaussian processes

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

$$\mathbb{P}(y(x_1) \leq c_1, \ldots, y(x_n) \leq c_n)$$

  as these uniquely determine the law of y.

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

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

  Write $y(\cdot) \sim GP$ to denote that the *function* y is a GP.

Mean and covariance function

- To fully specify the law of a Gaussian *distribution* we only need the mean and variance.

$$Y \sim N(\mu, \Sigma)$$

- To fully specify the law of a Gaussian *process*, we need to specify mean and covariance functions.

$$y(\cdot) \sim GP(m(\cdot), k(\cdot, \cdot))$$

where

$$\mathbb{E}(y(x)) = m(x)$$
$$\mathbb{C}\text{ov}(y(x), y(x')) = k(x, x')$$

Mean and covariance function

- To fully specify the law of a Gaussian *distribution* we only need the mean and variance.

$$Y \sim N(\mu, \Sigma)$$

- To fully specify the law of a Gaussian *process*, we need to specify mean and covariance functions.

$$y(\cdot) \sim GP(m(\cdot), k(\cdot, \cdot))$$

where

$$\mathbb{E}(y(x)) = m(x)$$
$$\mathbb{C}\text{ov}(y(x), y(x')) = k(x, x')$$

Mean and covariance function

- To fully specify the law of a Gaussian *distribution* we only need the mean and variance.

$$Y \sim N(\mu, \Sigma)$$

- To fully specify the law of a Gaussian *process*, we need to specify mean and covariance functions.

$$y(\cdot) \sim GP(m(\cdot), k(\cdot, \cdot))$$

where

$$\mathbb{E}(y(x)) = m(x)$$
$$\mathbb{C}\text{ov}(y(x), y(x')) = k(x, x')$$

Specifying the mean function

- We can use any mean function we want $m(x) = \mathbb{E}(y(x))$
- Most popular choices are $m(x) = 0$ or $m(x) = const$ for all x, or $m(x) = \beta^\top x$.
- Using a neural network is another popular choice.

Specifying the mean function

- We can use any mean function we want $m(x) = \mathbb{E}(y(x))$
- Most popular choices are $m(x) = 0$ or $m(x) = const$ for all x, or $m(x) = \beta^\top x$.
- Using a neural network is another popular choice.

Specifying the mean function

- We can use any mean function we want $m(x) = \mathbb{E}(y(x))$
- Most popular choices are $m(x) = 0$ or $m(x) = const$ for all x, or $m(x) = \beta^\top x$.
- Using a neural network is another popular choice.

Covariance functions

- We usually use a covariance function that is a function of the indexes/locations

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

k must be a positive semi-definite function, i.e., lead to valid covariance matrices.

- Given locations $x_1, \ldots, x_n$, the $n \times n$ Gram matrix K with $K_{ij} = k(x_i, x_j)$ must be a positive semi-definite matrix.

Covariance functions

- We usually use a covariance function that is a function of the indexes/locations

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

  k must be a positive semi-definite function, i.e., lead to valid covariance matrices.

- Given locations $x_1, \ldots, x_n$, the $n \times n$ Gram matrix K with $K_{ij} = k(x_i, x_j)$ must be a positive semi-definite matrix.

Covariance functions

- We often assume k is a function of only the distance between locations

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

which results in a stationary processes.

- If $\mathbb{C}\text{ov}(y(x), y(x')) = k(||x - x'||)$ the covariance function is said to be isotropic.

- The covariance function determines the *nature* of the GP.

- k determines the hypothesis space/space of functions

Covariance functions

- We often assume k is a function of only the distance between locations

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

  which results in a stationary processes.

- If $\mathbb{C}\text{ov}(y(x), y(x')) = k(||x - x'||)$ the covariance function is said to be isotropic.

- The covariance function determines the *nature* of the GP.

- k determines the hypothesis space/space of functions

Covariance functions

- We often assume k is a function of only the distance between locations

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

  which results in a stationary processes.

- If $\mathbb{C}\text{ov}(y(x), y(x')) = k(||x - x'||)$ the covariance function is said to be isotropic.

- The covariance function determines the *nature* of the GP.

- k determines the hypothesis space/space of functions

Covariance functions

- We often assume k is a function of only the distance between locations

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

  which results in a stationary processes.

- If $\mathbb{C}\text{ov}(y(x), y(x')) = k(||x - x'||)$ the covariance function is said to be isotropic.

- The covariance function determines the *nature* of the GP.

- k determines the hypothesis space/space of functions

How do we draw samples from a GP?

- Given the mean function and covariance function for a GP, we can draw samples using a multivariate Gaussian distribution.
- To sample from the multivariate Gaussian distribution, we need a mean vector and a covariance matrix.
- The mean vector is obtained from the mean function.
- The covariance matrix is obtained from the covariance function.

Sampling from a GP

- RBF/Squared-exponential/exponentiated quadratic

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

where s_f is the variance parameter and ℓ the length-scale parameter.

- If $s_f = 1$ and $\ell = 1$, we get

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

- Say we have a vector of x values, like

$$\mathbf{x} = [x_1, x_2, \cdots, x_n]^\top.$$

- These are the indexes of the stochastic process.

Sampling from a GP

- We now compute the covariance matrix

$$K_{XX} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}$$

- We assume the mean function is constant and equal to zero, $m(x) = 0$.

- To generate functions from this GP, we will then sample from

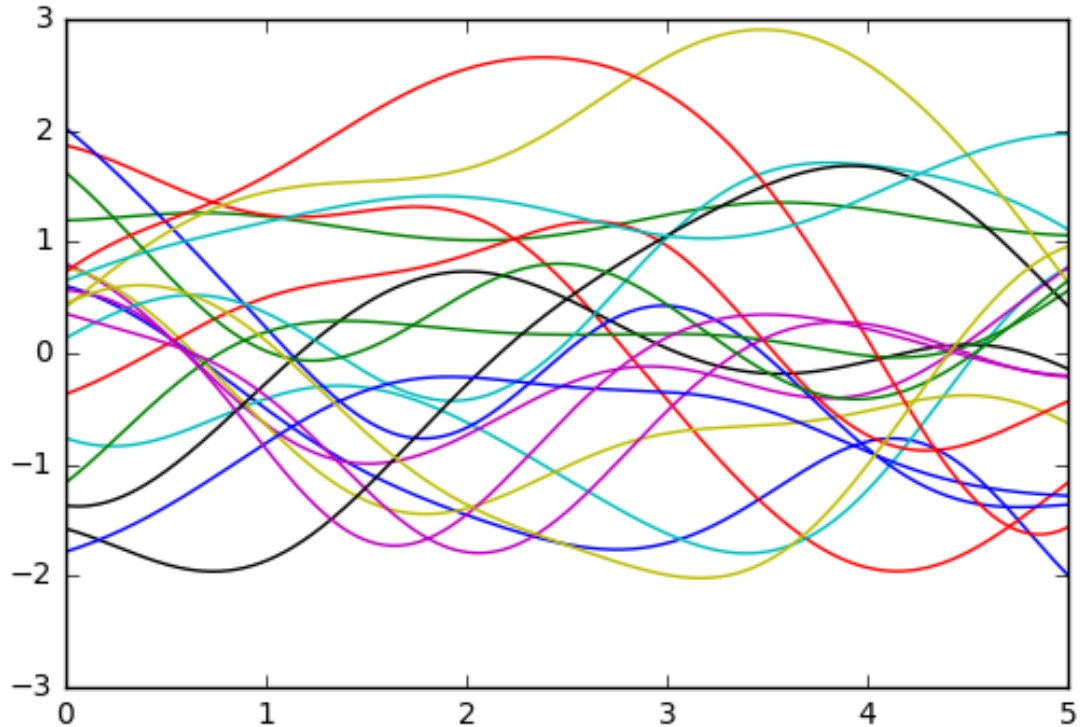
$$\mathbf{y} \sim N_n \left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix} \right)$$

- What we plot is $\mathbf{x}$ and $\mathbf{y}$.

Examples

RBF/Squared-exponential/exponentiated quadratic

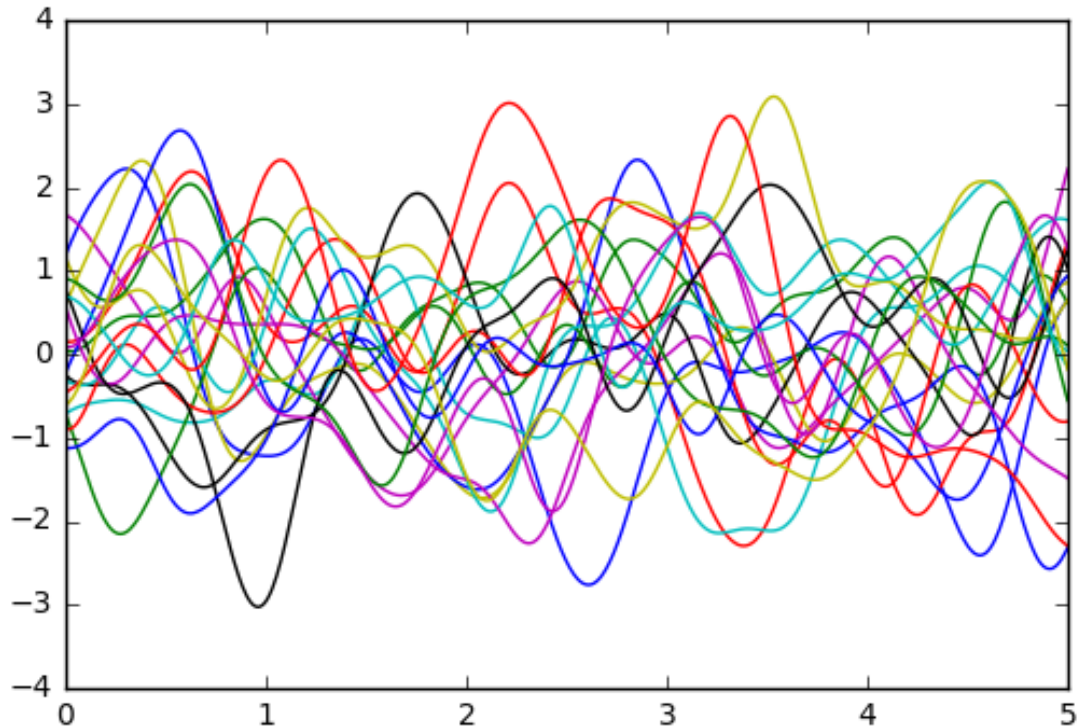
$$k(x, x') = \exp\left(-\frac{1}{2}(x - x')^2\right)$$



Examples

RBF/Squared-exponential/exponentiated quadratic

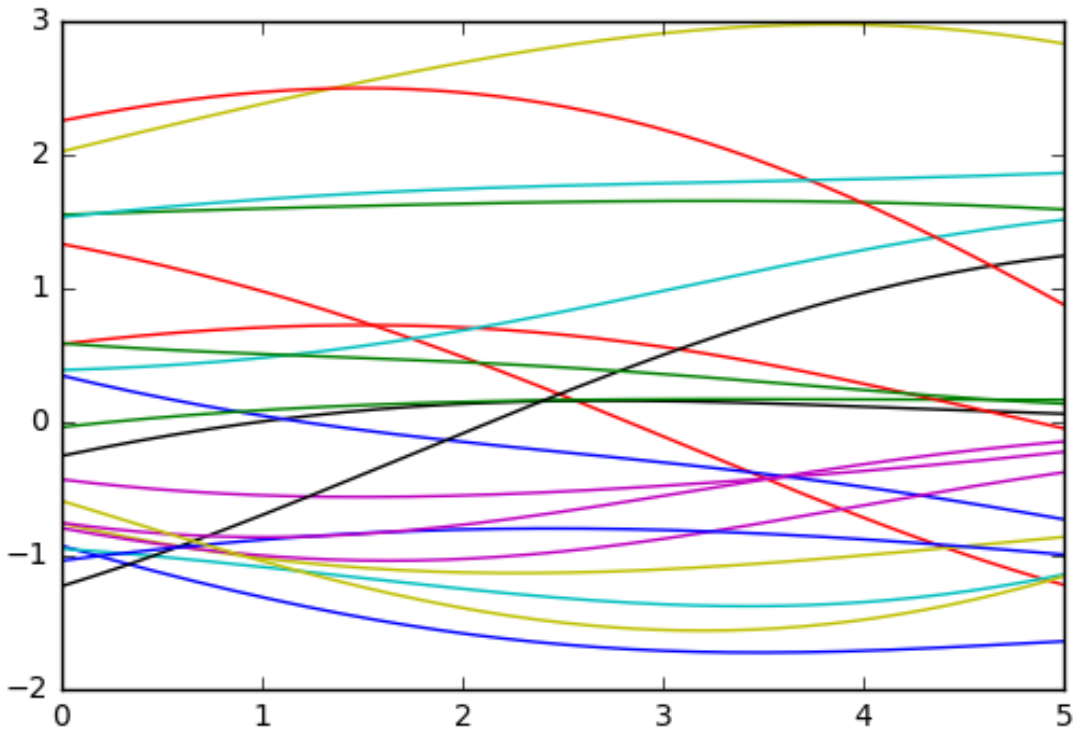
$$k(x, x') = \exp\left(-\frac{1}{2}\frac{(x - x')^2}{0.25^2}\right)$$



Examples

RBF/Squared-exponential/exponentiated quadratic

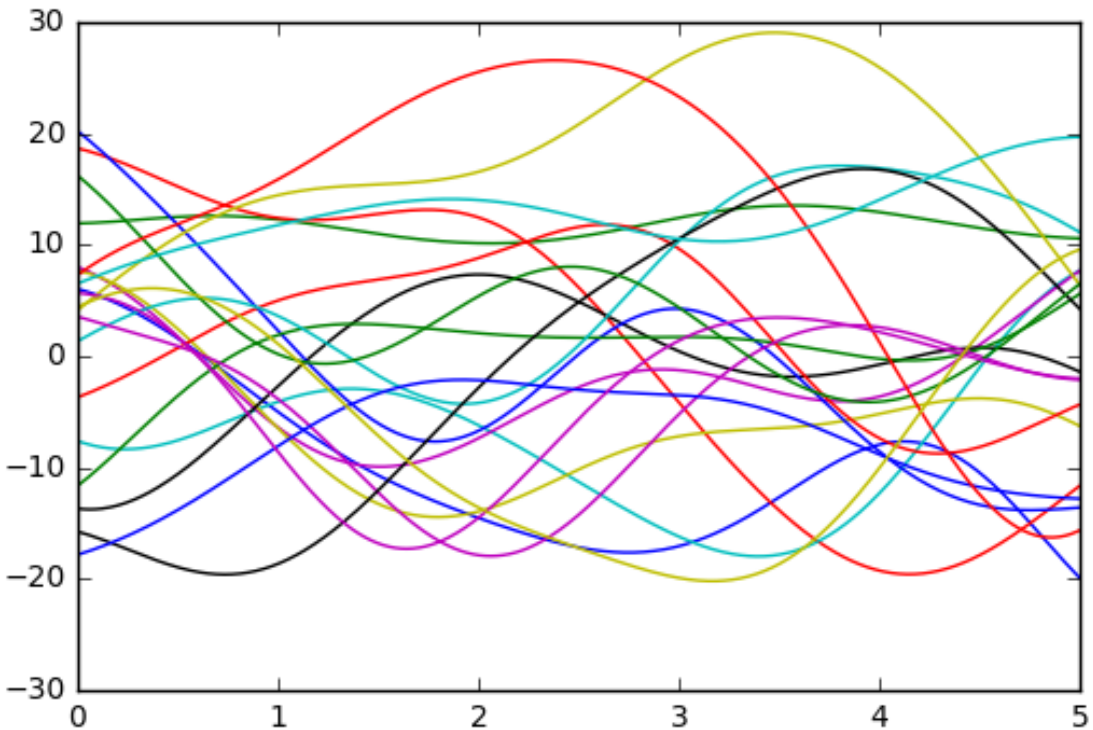
$$k(x, x') = \exp\left(-\frac{1}{2}\frac{(x - x')^2}{4^2}\right)$$



Examples

RBF/Squared-exponential/exponentiated quadratic

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



Examples

Matèrn covariance function

$$k(x, x') = s_f \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}(x - x')}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}(x - x')}{\ell} \right)$$

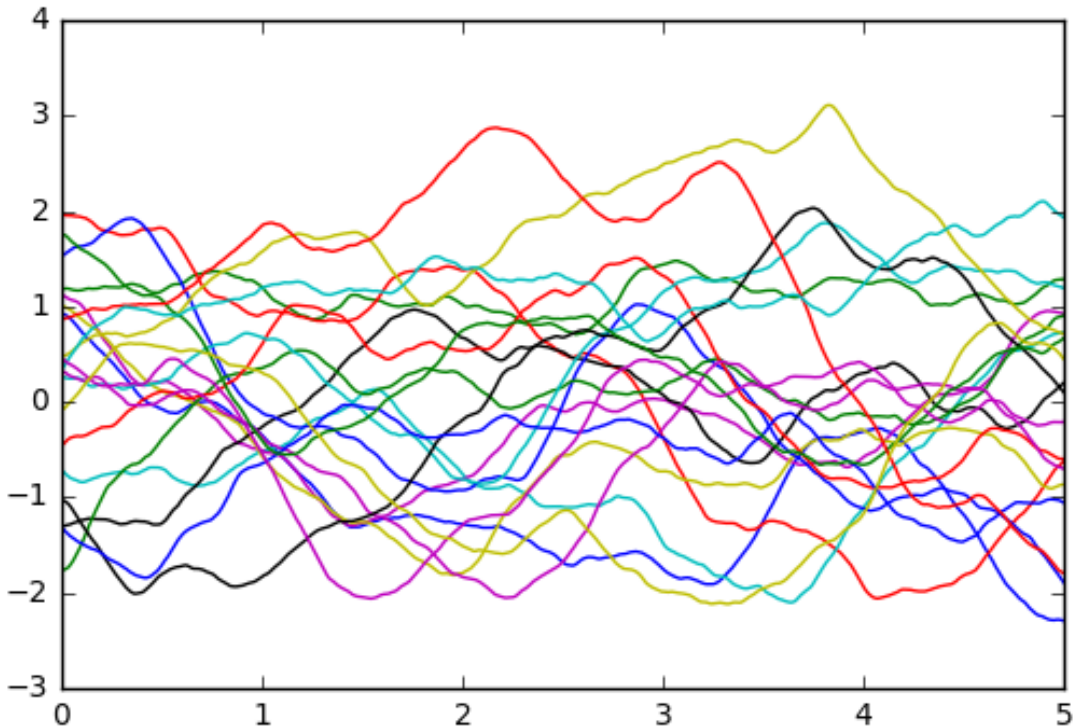
$$k(x, x') = s_f \left(1 + \frac{\sqrt{3}|x - x'|}{\ell} \right) \exp \left(- \frac{\sqrt{3}|x - x'|}{\ell} \right), \quad \nu = \frac{3}{2},$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind.

Examples

Matèrn 3/2

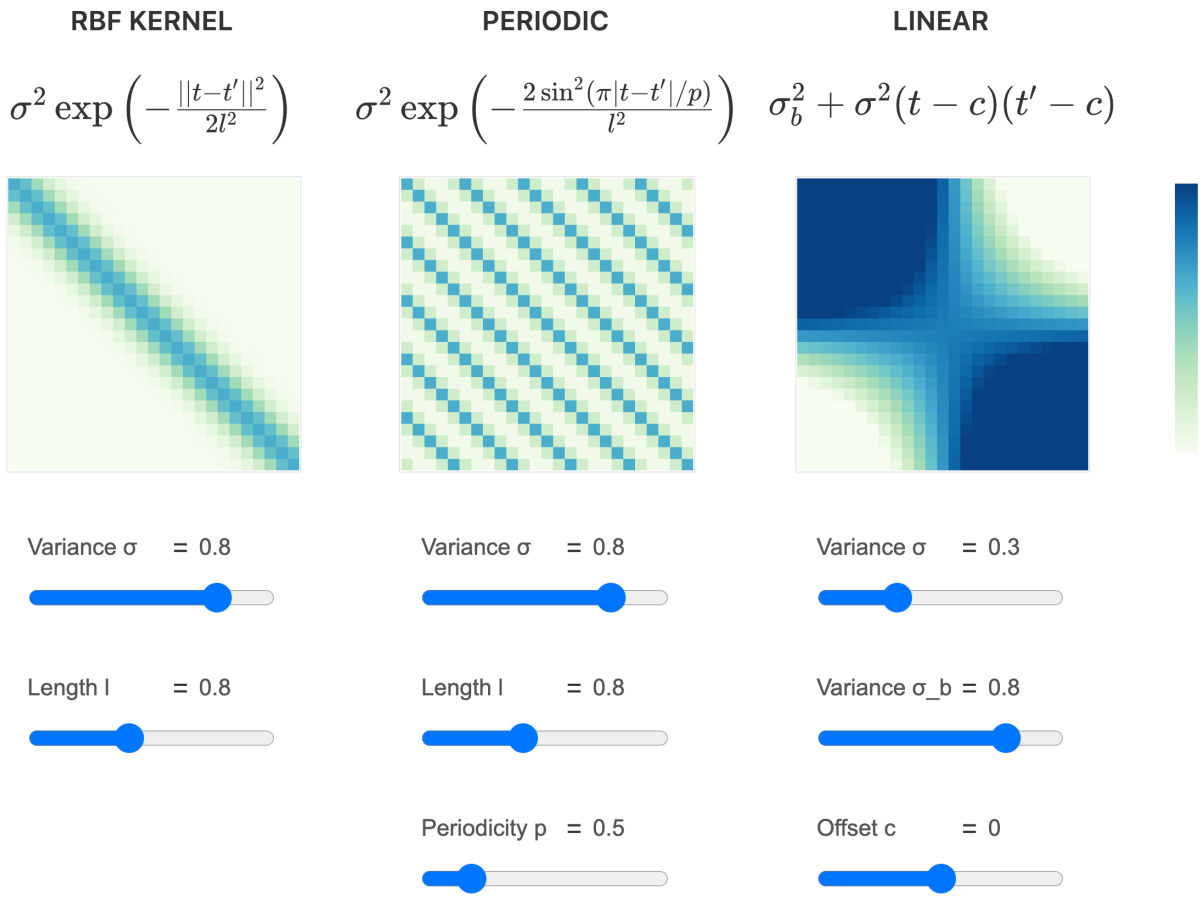
$$k(x, x') \sim (1 + |x - x'|) \exp(-|x - x'|)$$



Examples

Many other covariance functions: constant, linear, polynomial, exponential, etc.

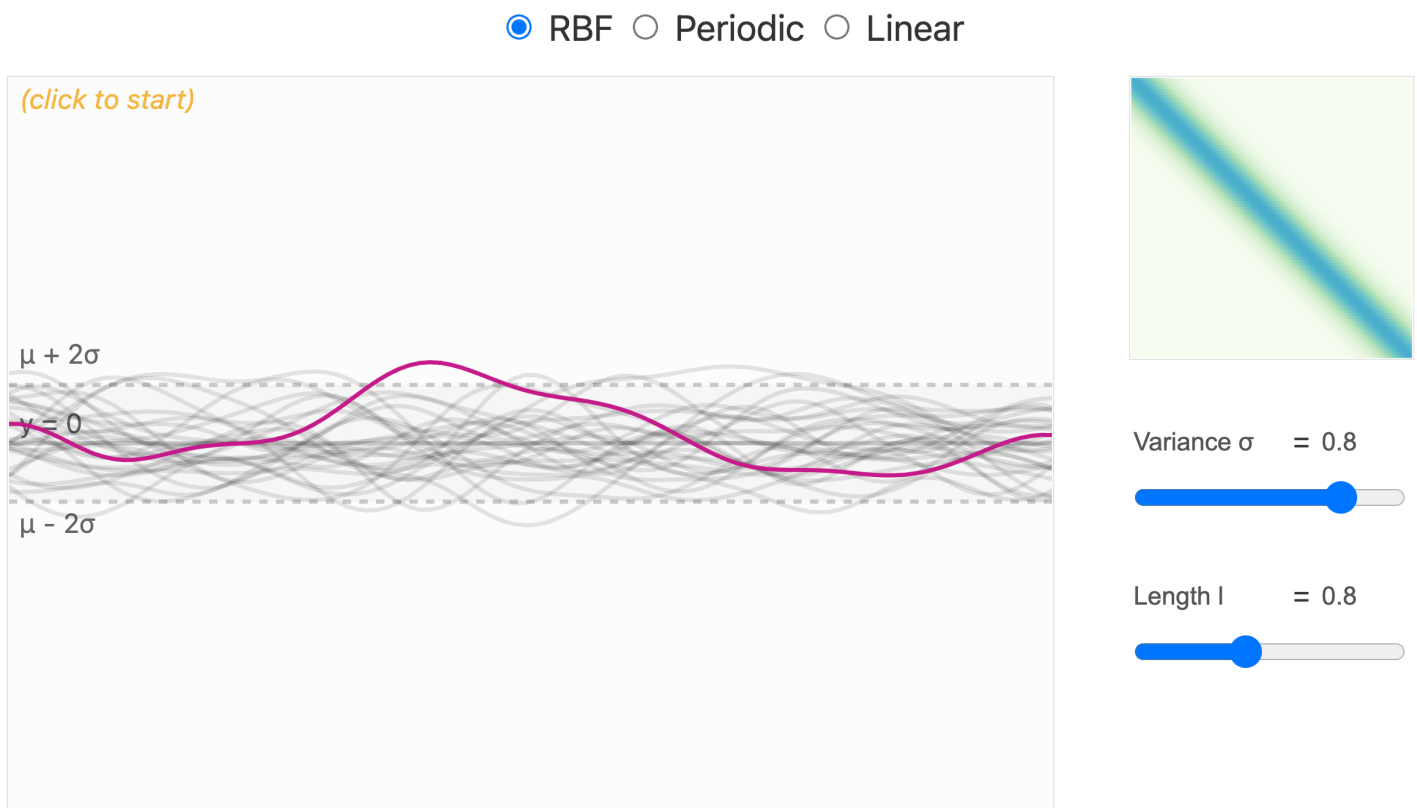
Visual exploration: covariance matrices



This figure shows various kernels that can be used with Gaussian processes. Each kernel has different parameters, which can be changed by adjusting the according sliders. When grabbing a slider, information on how the current parameter influences the kernel will be shown on the right.

Taken from: “Visual exploration of Gaussian processes” by J Götler, R Kehlbeck and O Deussen (2019)

Visual exploration: samples from GPs



Clicking on the graph results in continuous samples drawn from a Gaussian process using the selected kernel. After each draw, the previous sample fades into the background. Over time, it is possible to see that functions are distributed normally around the mean μ .

Taken from: “Visual exploration of Gaussian processes” by J Götler, R Kehlbeck and O Deussen (2019)

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

Conditional distributions are still Gaussian

Suppose

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2(\mu, \Sigma)$$

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

$$Y_2 \mid Y_1 = y_1 \sim N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Conditional distributions are still Gaussian

Suppose

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2(\mu, \Sigma)$$

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

$$Y_2 \mid Y_1 = y_1 \sim N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Conditional updates of Gaussian processes

Suppose f is a Gaussian process, then

$$f(x_1), \ldots, f(x_n), f(x) \sim N_{n+1}(0, \Sigma)$$

where

$$\Sigma = \left(\begin{array}{ccc|c} k(x_1, x_1) & \ldots & k(x_1, x_n) & k(x_1, x) \\ \vdots & & \vdots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) & k(x_n, x) \\ \hline k(x, x_1) & \ldots & k(x, x_n) & k(x, x) \end{array}\right)$$

$$= \left(\begin{array}{c|c} K_{XX} & k_X(x) \\ \hline k_X(x)^\top & k(x, x) \end{array}\right)$$

where $X = \{x_1, \ldots, x_n\}$, $[K_{XX}]_{ij} = k(x_i, x_j)$ is the Gram/kernel matrix, and $[k_X(x)]_j = k(x_j, x)$

Conditional updates of Gaussian processes

Suppose f is a Gaussian process, then

$$f(x_1), \ldots, f(x_n), f(x) \sim N_{n+1}(0, \Sigma)$$

where

$$\Sigma = \left(\begin{array}{ccc|c} k(x_1, x_1) & \ldots & k(x_1, x_n) & k(x_1, x) \\ \vdots & & \vdots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) & k(x_n, x) \\ \hline k(x, x_1) & \ldots & k(x, x_n) & k(x, x) \end{array}\right)$$

$$= \left(\begin{array}{c|c} K_{XX} & k_X(x) \\ \hline k_X(x)^\top & k(x, x) \end{array}\right)$$

where $X = \{x_1, \ldots, x_n\}$, $[K_{XX}]_{ij} = k(x_i, x_j)$ is the Gram/kernel matrix, and $[k_X(x)]_j = k(x_j, x)$

Conditional updates of Gaussian processes

Then

$$f(x)|f(x_1),\ldots,f(x_n) \sim N(\bar{m}(x), \bar{k}(x))$$

where

$$\bar{m}(x) = k_X(x)^\top K_{XX}^{-1}\mathbf{f}$$

with

$$\mathbf{f} = (f(x_1),\ldots,f(x_n))^\top$$
$$k_X(x)^\top = (k(x,x_1)\ \ k(x,x_2)\ \ \ldots\ \ k(x,x_n)) \in \mathbb{R}^{1\times n}$$

and

$$\bar{k}(x) = k(x,x) - k_X(x)^\top K_{XX}^{-1} k_X(x)$$

What this means in practice is that if we know $\mathbf{f}$, we can use it to predict $f(x)$ as a Gaussian distribution with mean $\bar{m}(x)$ and variance $\bar{k}(x)$.

Conditional updates of Gaussian processes

Then

$$f(x)|f(x_1),\ldots,f(x_n) \sim N(\bar{m}(x),\bar{k}(x))$$

where

$$\bar{m}(x) = k_X(x)^\top K_{XX}^{-1}\mathbf{f}$$

with

$$\mathbf{f} = (f(x_1),\ldots,f(x_n))^\top$$
$$k_X(x)^\top = (k(x,x_1)\;\; k(x,x_2)\;\; \ldots\;\; k(x,x_n)) \in \mathbb{R}^{1\times n}$$

and

$$\bar{k}(x) = k(x,x) - k_X(x)^\top K_{XX}^{-1} k_X(x)$$

What this means in practice is that if we know $\mathbf{f}$, we can use it to predict $f(x)$ as a Gaussian distribution with mean $\bar{m}(x)$ and variance $\bar{k}(x)$.

Conditional updates of Gaussian processes

Then

$$f(x)|f(x_1),\ldots,f(x_n) \sim N(\bar{m}(x),\bar{k}(x))$$

where

$$\bar{m}(x) = k_X(x)^\top K_{XX}^{-1}\mathbf{f}$$

with

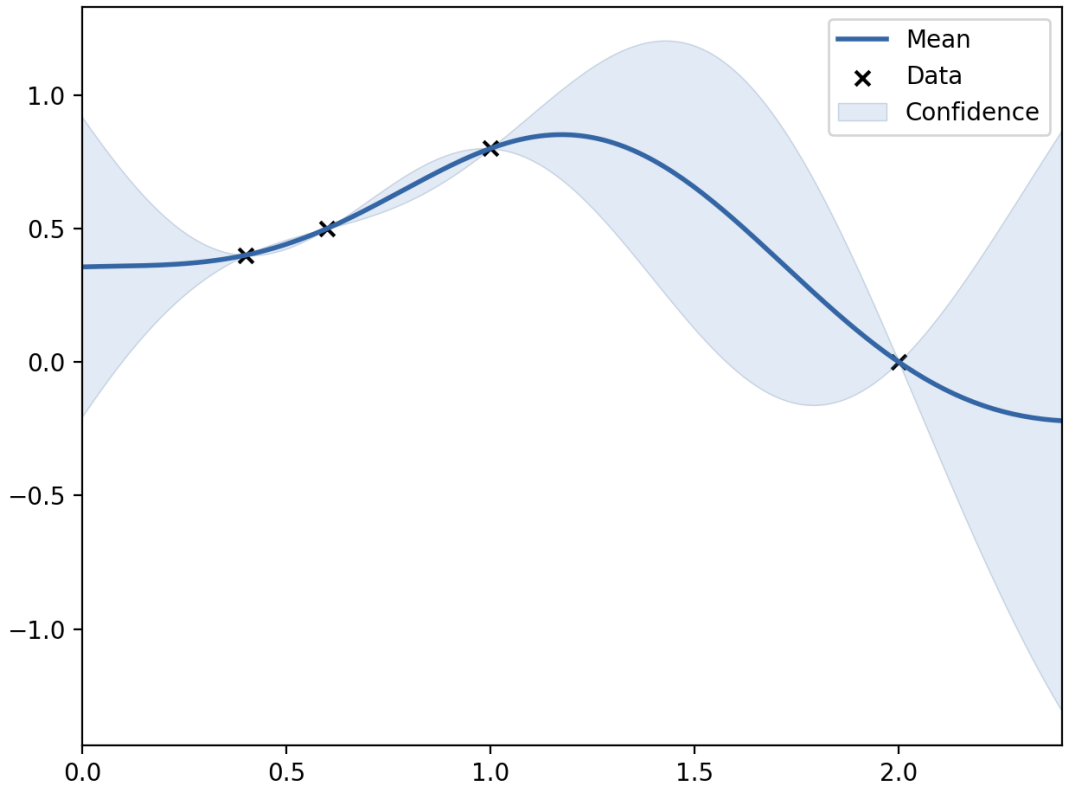
$$\mathbf{f} = (f(x_1),\ldots,f(x_n))^\top$$
$$k_X(x)^\top = (k(x,x_1)\;\; k(x,x_2)\;\; \ldots\;\; k(x,x_n)) \in \mathbb{R}^{1\times n}$$

and

$$\bar{k}(x) = k(x,x) - k_X(x)^\top K_{XX}^{-1} k_X(x)$$

What this means in practice is that if we know $\mathbf{f}$, we can use it to predict $f(x)$ as a Gaussian distribution with mean $\bar{m}(x)$ and variance $\bar{k}(x)$.

Interpolation



Solid line $\quad \bar{m}(x) = k_X^\top(x) K_{XX}^{-1} \mathbf{f}$

Shaded region $\quad \bar{m}(x) \pm 1.96\sqrt{\bar{k}(x)}$

$$\bar{k}(x) = k(x,x) - k_X^\top(x) K_{XX}^{-1} k_X(x)$$

Noisy observations - Regression

- In practice, we don't usually observe $f(x)$ directly.
- If we observe

$$y_i = f(x_i) + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ then

$$y_1, \ldots, y_n, f(x) \sim N_{n+1}(0, \Sigma)$$

where

$$\Sigma = \left(\begin{array}{ccc|c} & K_{XX} + \sigma^2 I & & \begin{matrix} k(x_1, x) \\ k(x_2, x) \\ \vdots \\ k(x_n, x) \end{matrix} \\ \hline k(x, x_1) & k(x, x_2) \quad \ldots \quad k(x, x_n) & & k(x, x) \end{array}\right)$$

Noisy observations - Regression

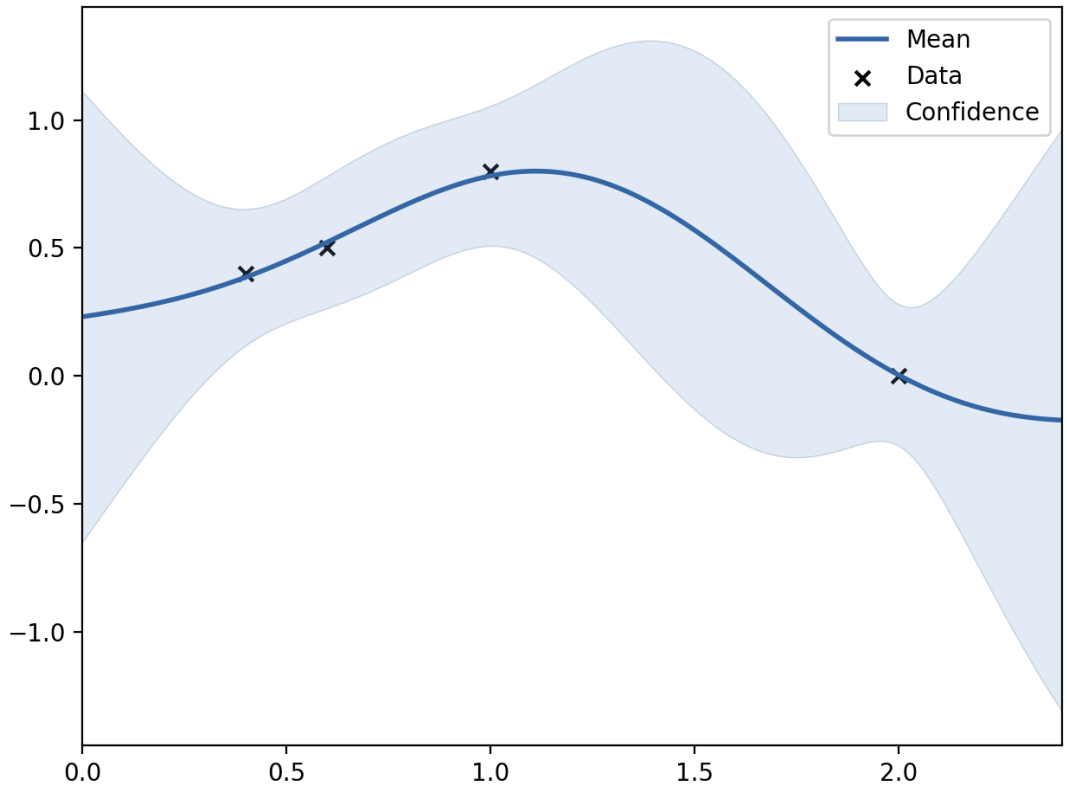
- In this way

$$f(x) \mid y_1, \ldots, y_n \sim N(\bar{m}(x), \bar{k}(x))$$

where

$$\bar{m}(x) = k_X(x)^\top (K_{XX} + \sigma^2 I)^{-1} \mathbf{y}$$
$$\bar{k}(x) = k(x, x) - k_X(x)^\top (K_{XX} + \sigma^2 I)^{-1} k_X(x)$$

Noise standard deviation $\sigma = 0.1$

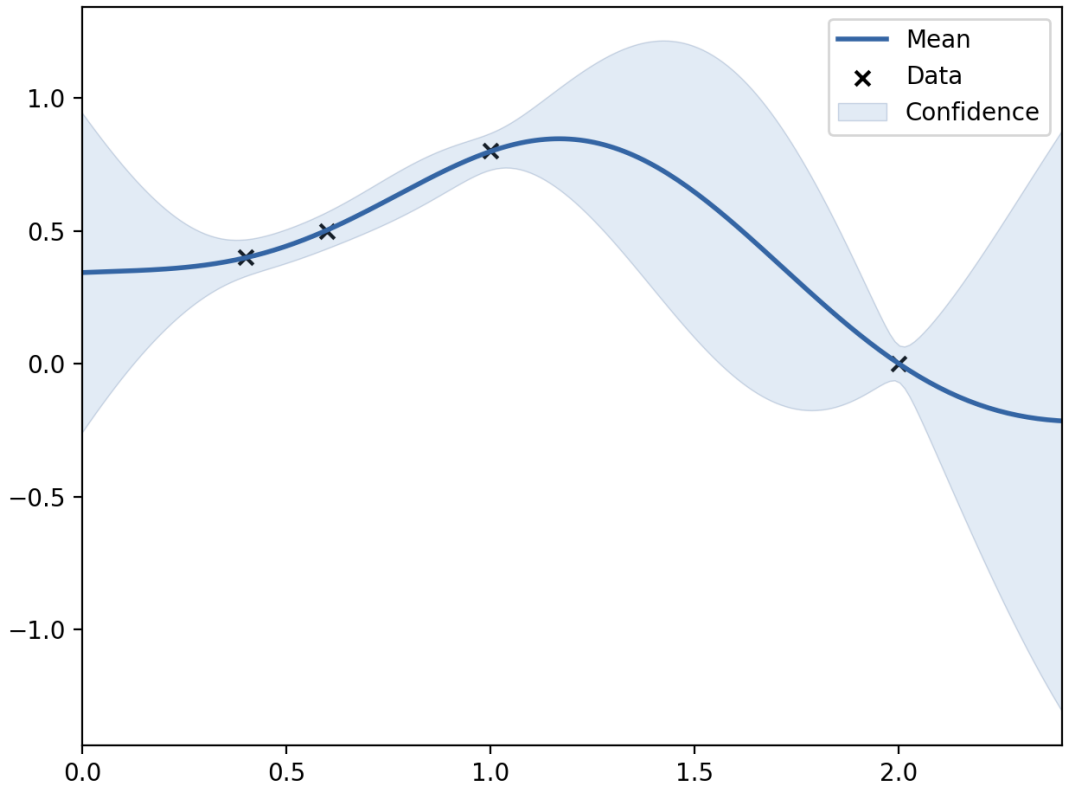


Solid line $\quad \bar{m}(x) = k_X(x)^\top K_{XX}^{-1}\mathbf{y}$

Shaded region $\quad \bar{m}(x) \pm 1.96\sqrt{\bar{k}(x)}$

$$\bar{k}(x) = k(x,x) - k_X(x)^\top (K_{XX}^{-1} + \sigma^2 I) k_X(x)$$

Noise standard deviation $\sigma = 0.025$

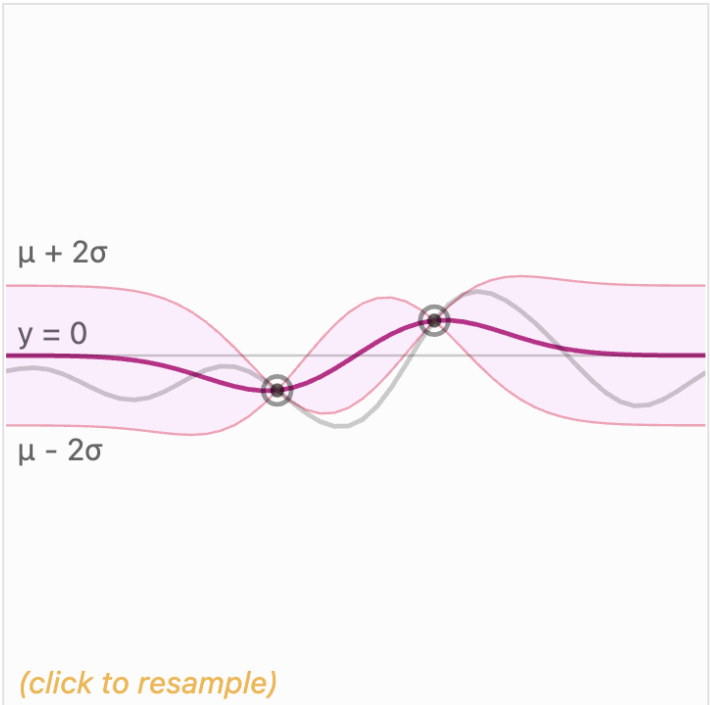


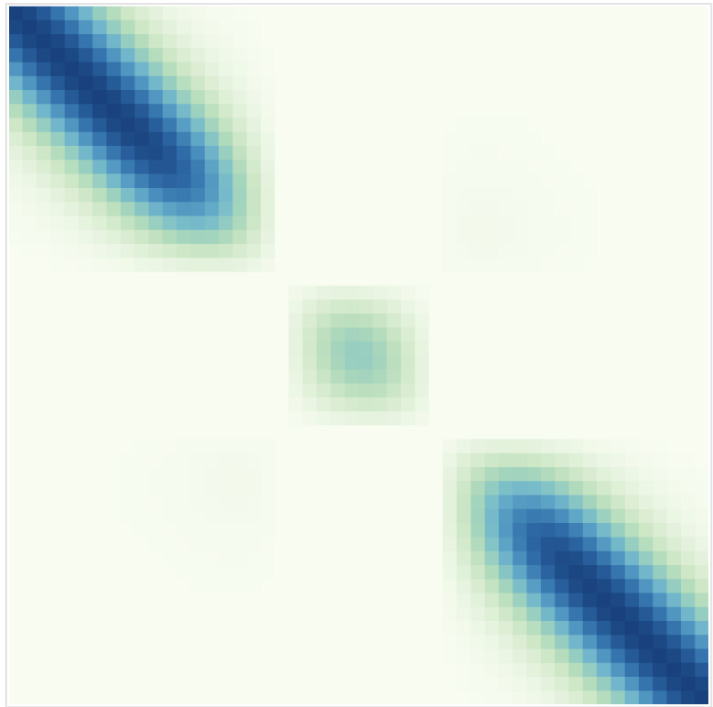
Solid line $\quad \bar{m}(x) = k_X(x)^\top K_{XX}^{-1}\mathbf{y}$

Shaded region $\quad \bar{m}(x) \pm 1.96\sqrt{\bar{k}(x)}$

$$\bar{k}(x) = k(x,x) - k_X(x)^\top (K_{XX}^{-1} + \sigma^2 I) k_X(x)$$

Visual exploration





Taken from: “Visual exploration of Gaussian processes” by J Götler, R Kehlbeck and O Deussen (2019)

Practical aspects

- If we knew the 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.

Practical aspects

- If we knew the 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.

Practical aspects

- If we knew the 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.

Practical aspects

- 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

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

Estimate these using your favourite statistical procedure (maximum likelihood, cross-validation, Bayes, expert judgement etc)

Practical aspects

- 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

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

  Estimate these using your favourite statistical procedure (maximum likelihood, cross-validation, Bayes, expert judgement etc)

Marginal likelihood

- A popular way to estimate the hyperparameters of the covariance function is through maximizing the logarithm of the marginal likelihood.

- The logarithm of the marginal likelihood is given as

$$\log p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top (K_{XX} + \sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log |K_{XX} + \sigma^2 I| - \frac{n}{2}\log 2\pi.$$

- If we know $\mathbf{x}$ and $\mathbf{y}$, the only unknowns in $\log p(\mathbf{y}|\mathbf{x})$ are the kernel hyperparameters, e.g. s_f and ℓ, and the parameter σ.

- We can then optimise $\log p(\mathbf{y}|\mathbf{x})$ wrt these parameters using a gradient-descent like procedure.

Marginal likelihood

- A popular way to estimate the hyperparameters of the covariance function is through maximizing the logarithm of the marginal likelihood.

- The logarithm of the marginal likelihood is given as

$$\log p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top (K_{xx} + \sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log |K_{xx} + \sigma^2 I| - \frac{n}{2}\log 2\pi.$$

- If we know $\mathbf{x}$ and $\mathbf{y}$, the only unknowns in $\log p(\mathbf{y}|\mathbf{x})$ are the kernel hyperparameters, e.g. s_f and ℓ, and the parameter σ.

- We can then optimise $\log p(\mathbf{y}|\mathbf{x})$ wrt these parameters using a gradient-descent like procedure.

Marginal likelihood

- A popular way to estimate the hyperparameters of the covariance function is through maximizing the logarithm of the marginal likelihood.

- The logarithm of the marginal likelihood is given as

$$\log p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top (K_{XX} + \sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log |K_{XX} + \sigma^2 I| - \frac{n}{2}\log 2\pi.$$

- If we know $\mathbf{x}$ and $\mathbf{y}$, the only unknowns in $\log p(\mathbf{y}|\mathbf{x})$ are the kernel hyperparameters, e.g. s_f and ℓ, and the parameter σ.

- We can then optimise $\log p(\mathbf{y}|\mathbf{x})$ wrt these parameters using a gradient-descent like procedure.

Log-marginal likelihood surface (σ and ℓ)

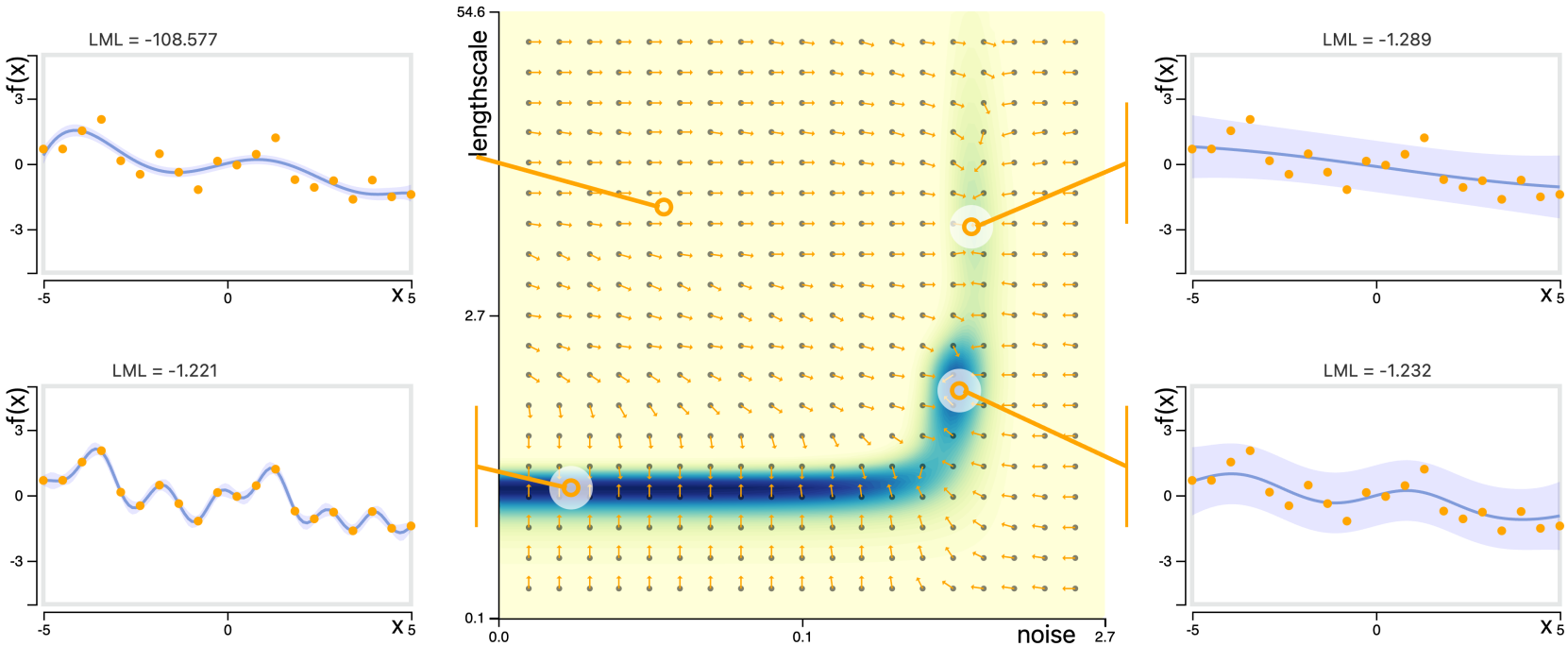


FIGURE 5: Training data (orange discs), log-marginal likelihood contour and three possible GP fits (left-bottom and right-hand panels) corresponding to the three local optima at $(\sigma_n, l) = (0.02, 0.36)$, $(0.97, 5.80)$, $(0.76, 1.13)$, respectively. All three hyperparameter settings and, therefore, the corresponding GP models make sense: While the bottom left GP model explains the (noisy) data well, the top right GP fit explains the data by a near-linear function (long lengthscale) and an high noise level. The global optimum (bottom right) has a slightly better log-marginal likelihood value and is a compromise between the other two other local optima, discovering the latent sinusoidal wave that generated the data while accounting for a fairly high level of measurement noise.

Taken from: “A Practical Guide to Gaussian Processes” by M. Deinsenroth, Y. Luo and M. van der Wilk (2013)

Computational cost

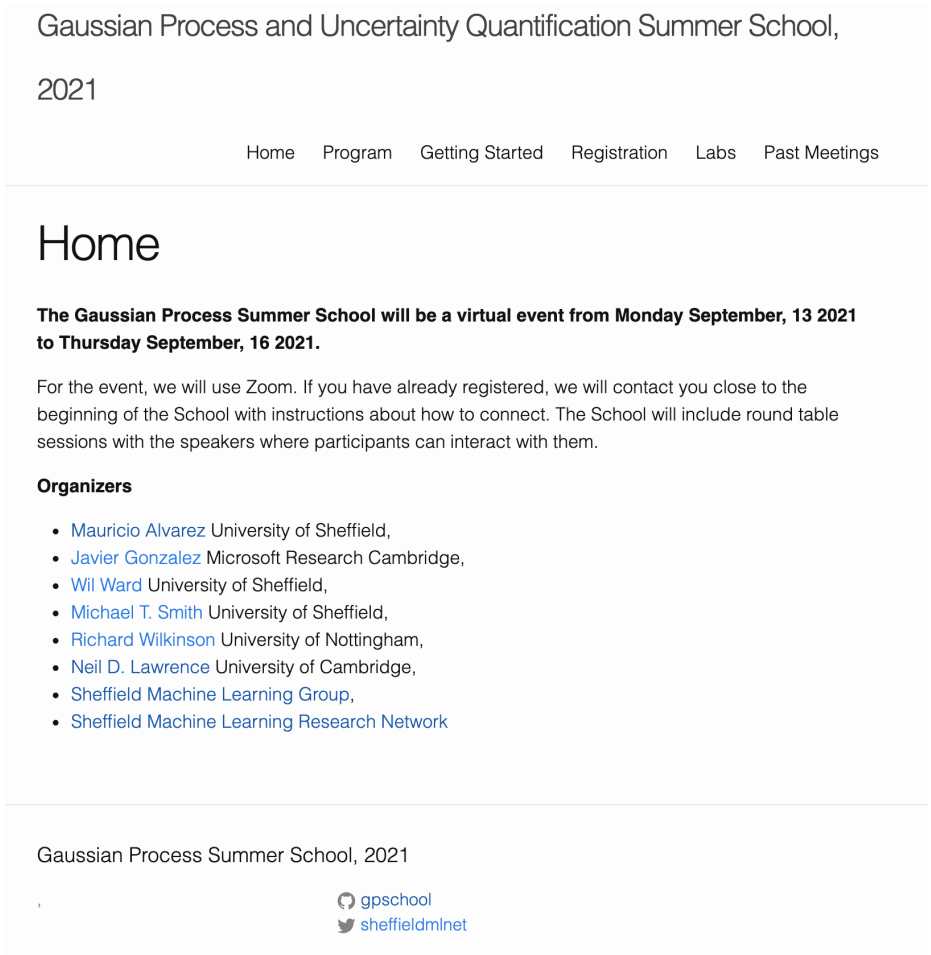
- One difficulty with GPs is the computational cost of training them: $O(n^3)$ (and $O(n^2)$ memory).
- They work our of the box for n in the order of a few thousands.
- There are many ways to side-step this cost: inducing inputs, efficient matrix-vector multiplications, random features, etc.
- These days we can use GPs for n in the order of tens of millions.

Computational cost

- One difficulty with GPs is the computational cost of training them: $O(n^3)$ (and $O(n^2)$ memory).
- They work our of the box for n in the order of a few thousands.
- There are many ways to side-step this cost: inducing inputs, efficient matrix-vector multiplications, random features, etc.
- These days we can use GPs for n in the order of tens of millions.

Contents

Gaussian process summer school



Gaussian Process and Uncertainty Quantification Summer School, 2021

Home Program Getting Started Registration Labs Past Meetings

Home

The Gaussian Process Summer School will be a virtual event from Monday September, 13 2021 to Thursday September, 16 2021.

For the event, we will use Zoom. If you have already registered, we will contact you close to the beginning of the School with instructions about how to connect. The School will include round table sessions with the speakers where participants can interact with them.

Organizers

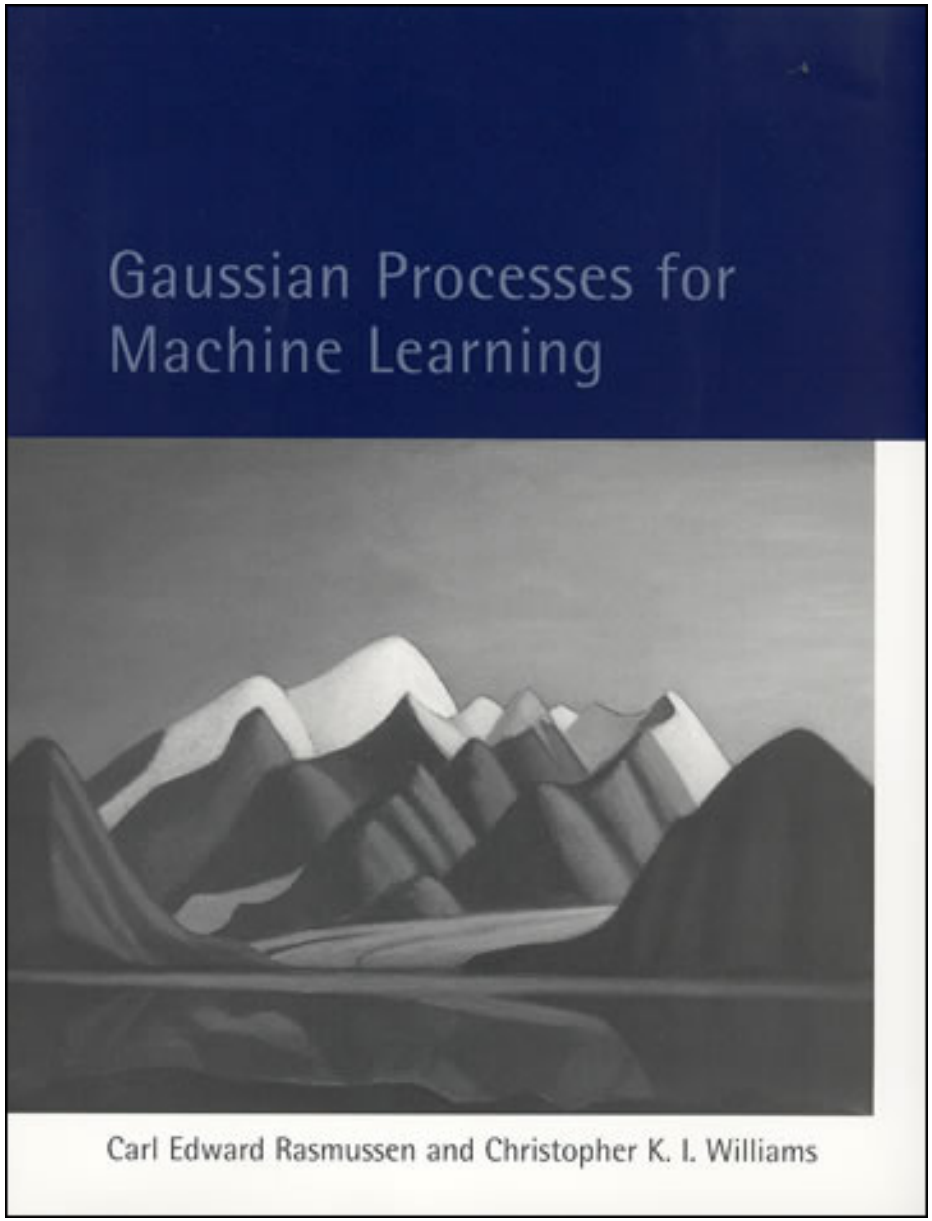
- Mauricio Alvarez University of Sheffield,
- Javier Gonzalez Microsoft Research Cambridge,
- Wil Ward University of Sheffield,
- Michael T. Smith University of Sheffield,
- Richard Wilkinson University of Nottingham,
- Neil D. Lawrence University of Cambridge,
- Sheffield Machine Learning Group,
- Sheffield Machine Learning Research Network

Gaussian Process Summer School, 2021

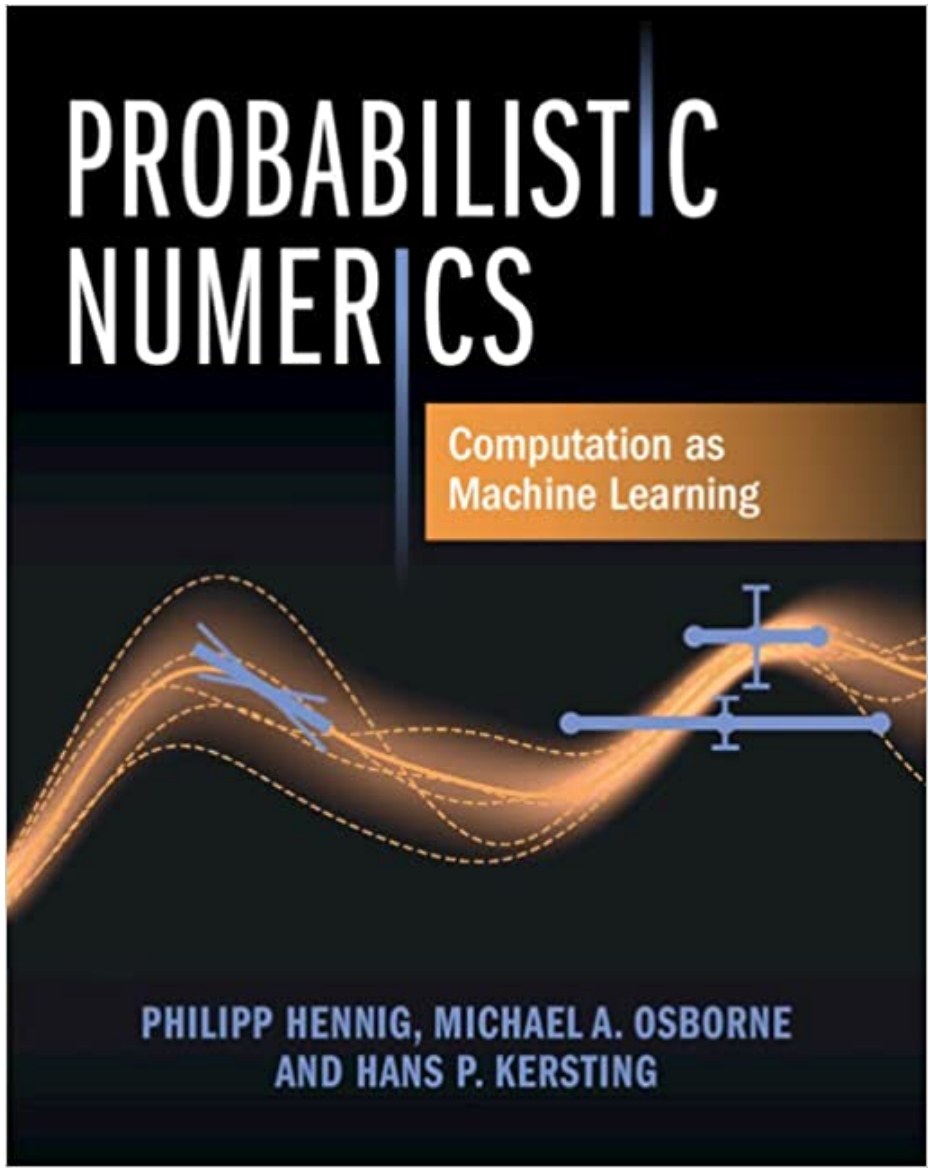
gpschool
sheffieldmlnet

Run the second week of September, go to http://gpss.cc

Book


Gaussian Processes for
Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams

Probabilistic numerics



Contents

Summary

- GPs are ubiquitous in statistics/ML.

- Popularity stems from
  - Naturalness of the framework
  - Mathematical tractability
  - Empirical success

Acknowledgements

- Prof. Richard Wilkinson from the University of Nottingham for providing some of the material in these slides.