Statistical Machine Learning

Lecture 04
Gaussian Process II

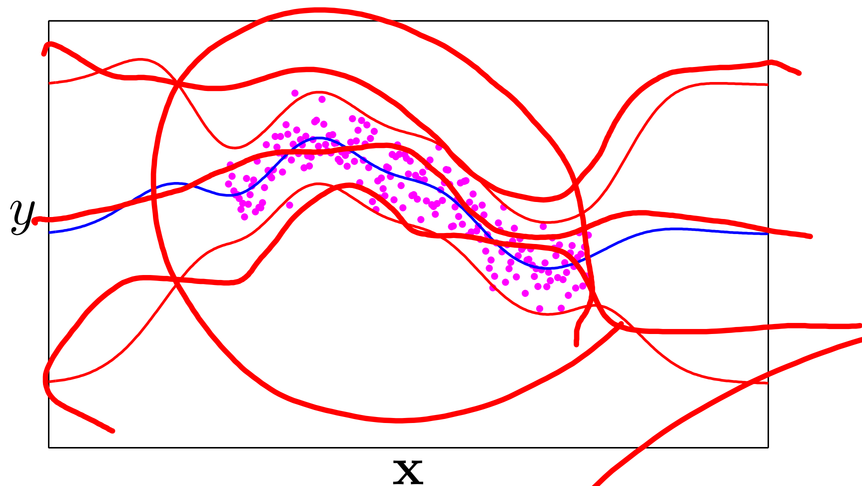
Sharif University of Technology

Spring 2021

Nonlinear regression

Consider the problem of nonlinear regression:

You want to learn a function f with error bars from data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$



GP

A Gaussian process defines a distribution over functions $p(f)$ which can be used for Bayesian regression:

GP

$$p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})}$$

likelihood
Normal

Gaussian Processes

A Gaussian process defines a distribution over functions, $p(f)$, where f is a function mapping some input space $\mathcal{X}$ to $\Re$.

R

$$f : \mathcal{X} \rightarrow \Re.$$

Notice that f can be an infinite-dimensional quantity (e.g. if $\mathcal{X} = \Re$)

Let $\mathbf{f} = (f(x_1), \ldots, f(x_n))$ be an n-dimensional vector of function values evaluated at n points $x_i \in \mathcal{X}$. Note $\mathbf{f}$ is a random variable.

Definition: $p(f)$ is a Gaussian process if for *any* finite subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the marginal distribution over that finite subset $p(\mathbf{f})$ has a multivariate Gaussian distribution.

Gaussian process covariance functions (kernels)

$p(f)$ is a Gaussian process if for *any* finite subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the marginal distribution over that finite subset $p(\mathbf{f})$ has a multivariate Gaussian distribution.

Gaussian processes (GPs) are parameterized by a mean function, $\mu(x)$, and a covariance function, or kernel, $K(x, x')$.

$$p(f(x), f(x')) = \mathsf{N}(\mu, \Sigma)$$

where

$$\mu = \begin{bmatrix} \mu(x) \\ \mu(x') \end{bmatrix} \quad \Sigma = \begin{bmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{bmatrix}$$

and similarly for $p(f(x_1), \ldots, f(x_n))$ where now μ is an $n \times 1$ vector and Σ is an $n \times n$ matrix.

Linear Kernel (covariance): $K(x, x') = x \cdot x'$

Polynomial: $K(x,x') = (x \cdot x')^d$

exponential: $K(x,x') = \exp(-\frac{1}{\lambda}|x-x'|)$

Gaussian process covariance functions

Gaussian processes (GPs) are parameterized by a mean function, $\mu(x)$, and a covariance function, $K(x, x')$.

An example covariance function:

$$\theta = (v_0, v_1, v_2, r, \alpha)$$

iid

$$K(x_i, x_j) = v_0 \exp\left\{-\left(\frac{|x_i - x_j|}{r}\right)^{\alpha}\right\} + v_1 + v_2\,\delta_{ij}$$

$$Y \sim N(\underline{f(x)}, \sigma^2 I)$$

with parameters $(v_0, v_1, v_2, r, \alpha)$

These kernel parameters are interpretable and can be learned from data:

$$f \sim GP(0, \underset{K}{\Sigma})$$

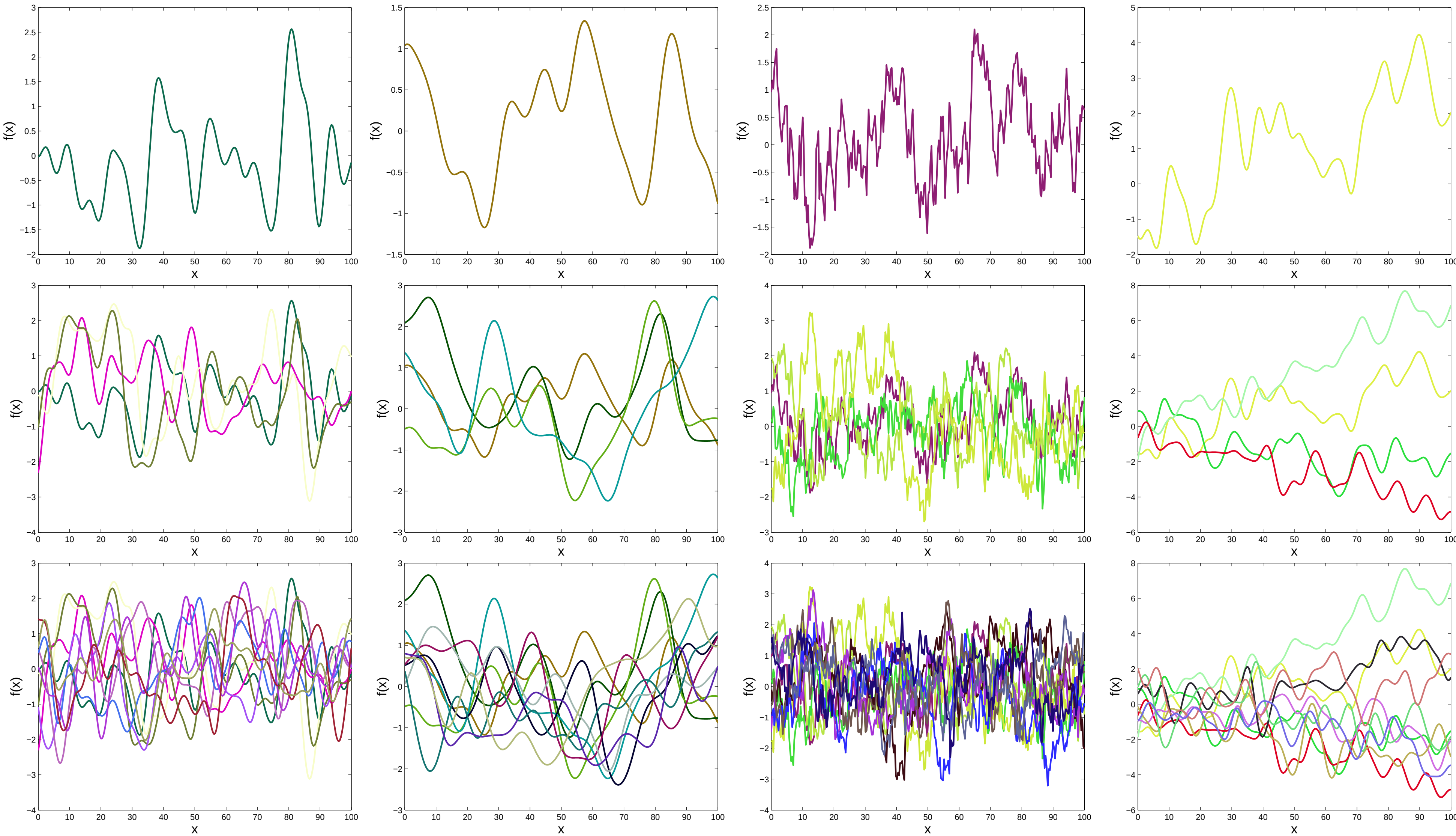
observe $\sim$ Normal iid

$$f(x) \sim GP(0, C(x, x'))$$

v_0	signal variance
v_1	variance of bias
v_2	noise variance
r	lengthscale
α	roughness

Once the mean and covariance functions are defined, everything else about GPs follows from the basic rules of probability applied to mutivariate Gaussians.

Samples from GPs with different $K(x, x')$

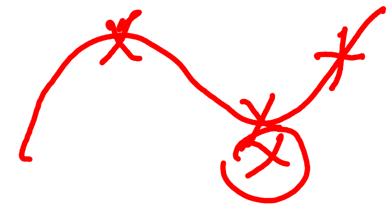


Using Gaussian processes for nonlinear regression

Imagine observing a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)_{i=1}^n\} = (\mathbf{X}, \mathbf{y})$.

Model:

$$\begin{aligned} y_i &= f(\mathbf{x}_i) + \epsilon_i \\ f &\sim \mathrm{GP}(\cdot|0, K) \\ \epsilon_i &\sim \mathrm{N}(\cdot|0, \sigma^2) \end{aligned}$$



$y_i = f(x_i) + \epsilon_i$

Prior on f is a GP, likelihood is Gaussian, therefore posterior on f is also a GP.

We can use this to make predictions

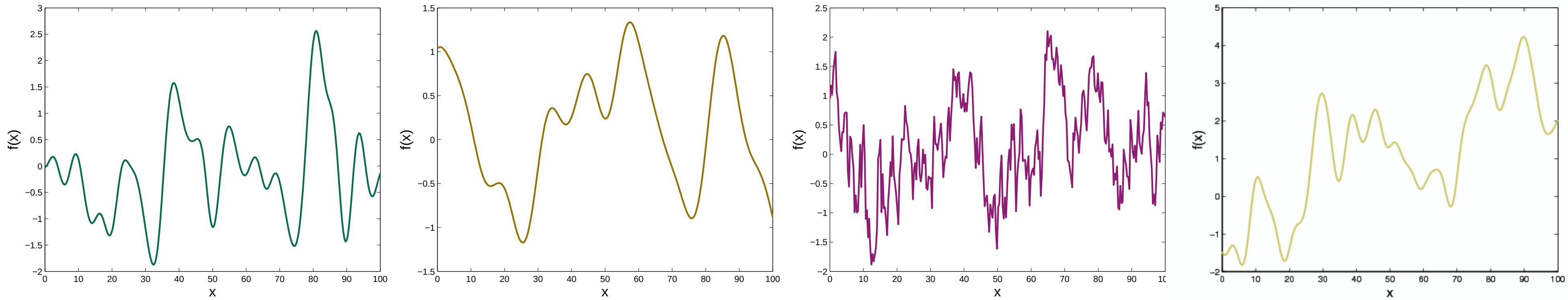
$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int p(y_*|\mathbf{x}_*, f, \mathcal{D})\, p(f|\mathcal{D})\, df$$

We can also compute the marginal likelihood (evidence) and use this to compare or tune covariance functions

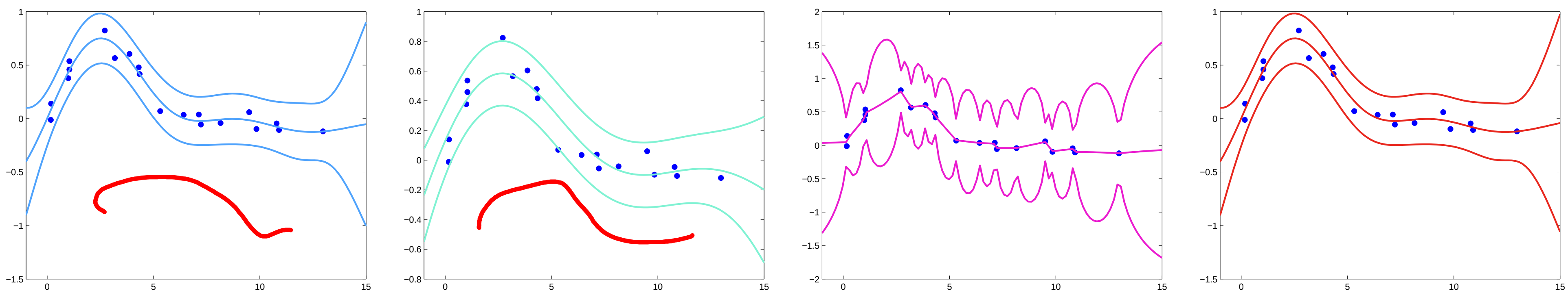
$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|f, \mathbf{X})\, p(f)\, df$$

Prediction using GPs with different $K(x, x')$

A sample from the prior for each covariance function:



Corresponding predictions, mean with two standard deviations:



`gpdemo`

GP learning the kernel

Consider the covariance function K with hyperparameters $\boldsymbol{\theta} = (v_0, v_1, r_1, \ldots, r_d, \alpha)$:

$$K_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) = v_0 \exp\left\{ -\sum_{d=1}^{D} \left(\frac{|x_i^{(d)} - x_j^{(d)}|}{r_d} \right)^{\alpha} \right\} + v_1$$

Given a data set $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, how do we learn $\boldsymbol{\theta}$?

The marginal likelihood is a function of $\boldsymbol{\theta}$

$$\boxed{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I})}$$

where its log is:

$$\ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2} \ln \det(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}) - \frac{1}{2} \mathbf{y}^\top (\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} + \text{const}$$

which can be optimized as a function of $\boldsymbol{\theta}$ and σ.

Alternatively, one can infer $\boldsymbol{\theta}$ using Bayesian methods, which is more costly but immune to overfitting.

From linear regression to GPs:

- Linear regression with inputs x_i and outputs y_i: $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
- Linear regression with M basis functions: $$y_i = \sum_{m=1}^{M} \beta_m \phi_m(x_i) + \epsilon_i$$
- Bayesian linear regression with basis functions:

$$\beta_m \sim \mathrm{N}(\cdot|0, \lambda_m) \quad \text{(independent of } \beta_\ell, \forall \ell \neq m\text{)}, \qquad \epsilon_i \sim \mathrm{N}(\cdot|0, \sigma^2)$$

- Integrating out the coefficients, β_j, we find:

$$E[y_i] = 0, \qquad Cov(y_i, y_j) = K_{ij} \stackrel{\text{def}}{=} \sum_{m=1}^{M} \lambda_m \, \phi_m(x_i) \, \phi_m(x_j) + \delta_{ij}\sigma^2$$

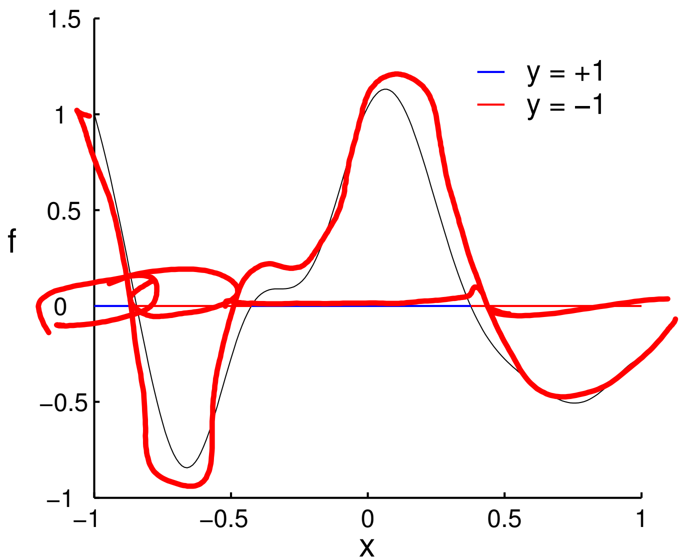
This is a Gaussian process with covariance function $K(x_i, x_j) = K_{ij}$.

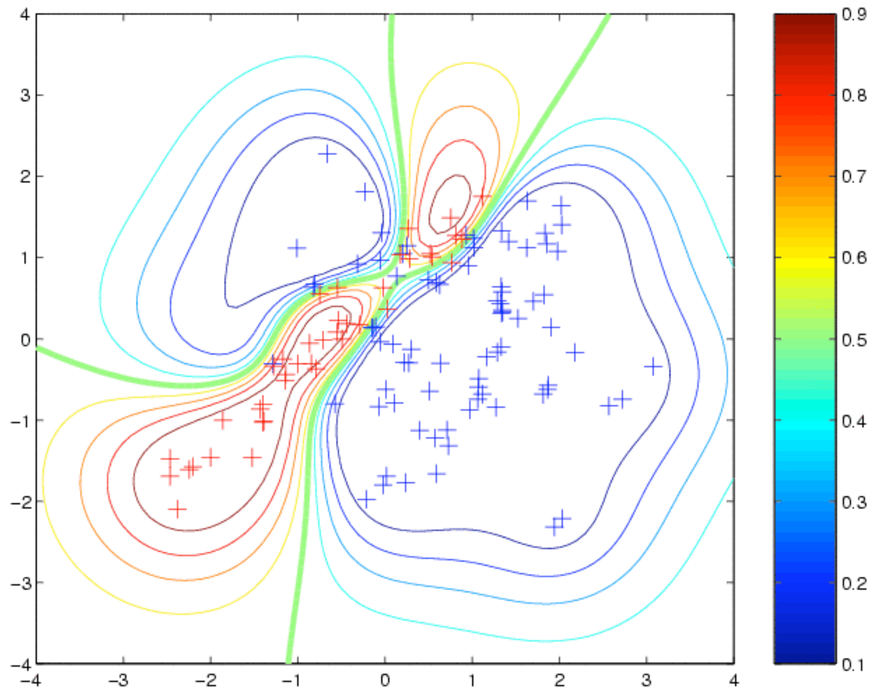
This GP has a finite number (M) of basis functions. Many useful GP kernels correspond to infinitely many basis functions (i.e. infinite-dim feature spaces).

A multilayer perceptron (neural network) with infinitely many hidden units and Gaussian priors on the weights $\rightarrow$ a GP (Neal, 1996)

Using Gaussian Processes for Classification

Binary classification problem: Given a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with binary class labels $y_i \in \{-1, +1\}$, infer class label probabilities at new points.





There are many ways to relate function values $f_i = f(\mathbf{x}_i)$ to class probabilities:

$$p(y_i|f_i) = \begin{cases} \frac{1}{1+\exp(-y_i f_i)} & \text{sigmoid (logistic)} \\ \Phi(y_i f_i) & \text{cumulative normal (probit)} \\ \boldsymbol{H}(y_i f_i) & \text{threshold} \\ \epsilon + (1-2\epsilon)\boldsymbol{H}(y_i f_i) & \text{robust threshold} \end{cases}$$

Non-Gaussian likelihood, so we need to use approximate inference methods (Laplace, EP, MCMC).

Two Simultaneous Equations

A system of two differential equations with two unknowns.

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

Two Simultaneous Equations

A system of two differential equations with two unknowns.

$$y_1 - y_2 = m(x_1 - x_2)$$

Two Simultaneous Equations

A system of two differential equations with two unknowns.

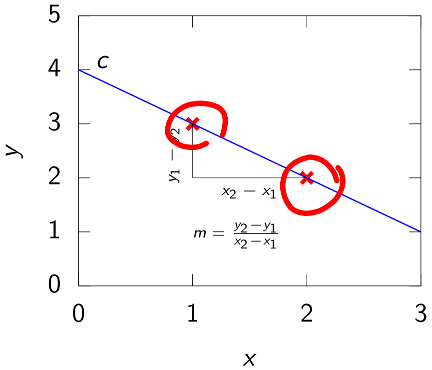
$$\frac{y_1 - y_2}{x_1 - x_2} = m$$

Two Simultaneous Equations

A system of two differential equations with two unknowns.

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

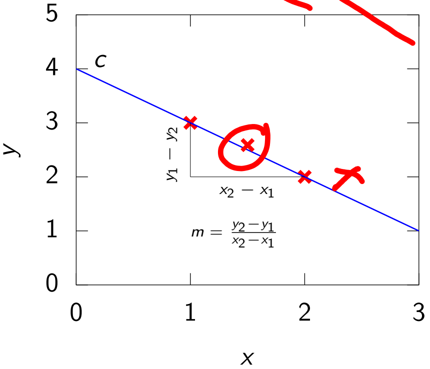
$$c = y_1 - mx_1$$



Two Simultaneous Equations

How do we deal with three simultaneous equations with only two unknowns?

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$
$$y_3 = mx_3 + c$$



Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
y_1 =& mx_1 + c \\
y_2 =& mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$
y_3 = mx_3 + c
$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$
\begin{aligned}
y_1 &= mx_1 + c + \epsilon_1 \\
y_2 &= mx_2 + c + \epsilon_2 \\
y_3 &= mx_3 + c + \epsilon_3
\end{aligned}
$$

Overdetermined System

- With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

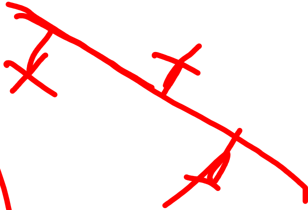
- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$y_1 = mx_1 + c + \epsilon_1$$
$$y_2 = mx_2 + c + \epsilon_2$$
$$y_3 = mx_3 + c + \epsilon_3$$

Overdetermined System



- With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$y_1 = mx_1 + c + \epsilon_1$$
$$y_2 = mx_2 + c + \epsilon_2$$
$$y_3 = mx_3 + c + \epsilon_3$$

Noise Models

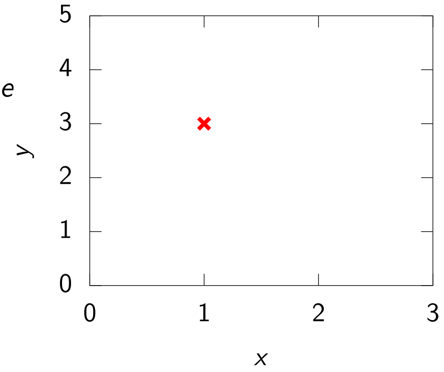
- We aren't modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to *least squares*.

H H H H
H T T H
T T T T

Underdetermined System

What about two unknowns and *one* observation?

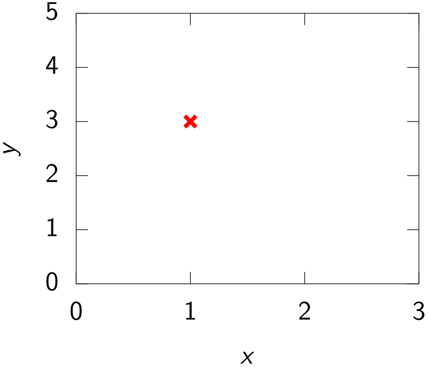
$$y_1 = mx_1 + c$$



Underdetermined System

Can compute m given c.

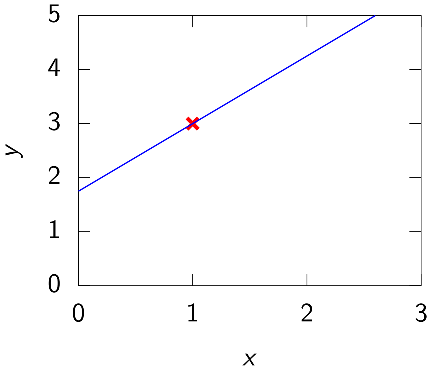
$$m = \frac{y_1 - c}{x}$$



Underdetermined System

Can compute m given c.

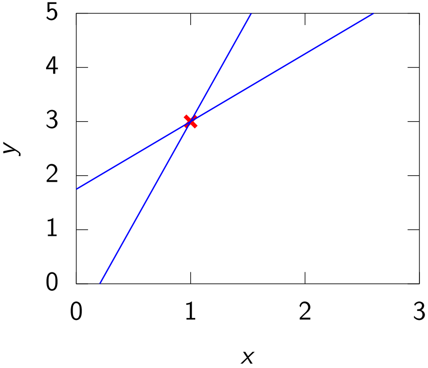
$$c = 1.75 \Longrightarrow m = 1.25$$



Underdetermined System

Can compute m given c.

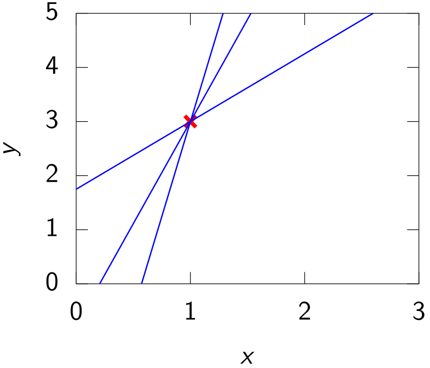
$$c = -0.777 \implies m = 3.78$$



Underdetermined System

Can compute m given c.

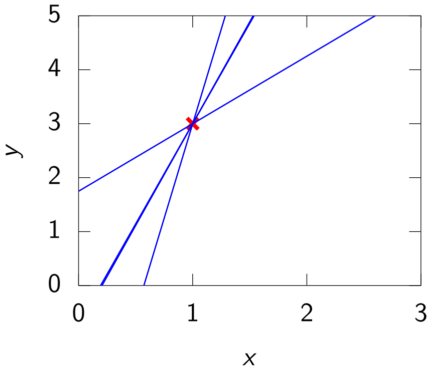
$$c = -4.01 \implies m = 7.01$$



Underdetermined System

Can compute m given c.

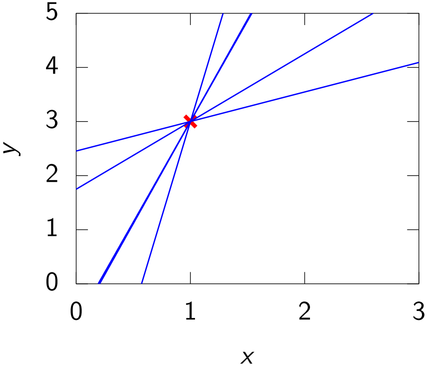
$$c = -0.718 \implies m = 3.72$$



Underdetermined System

Can compute m given c.

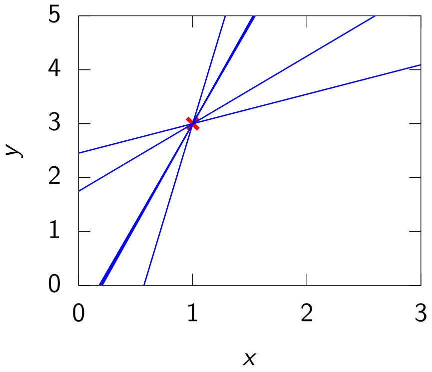
$$c = 2.45 \Longrightarrow m = 0.545$$



Underdetermined System

Can compute m given c.

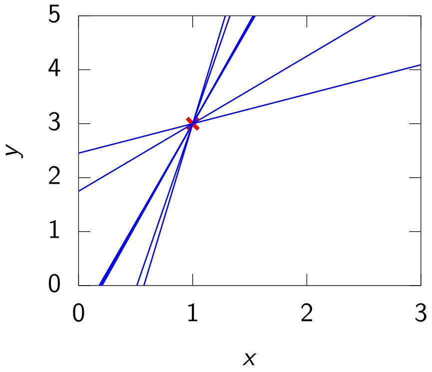
$$c = -0.657 \Longrightarrow m = 3.66$$



Underdetermined System

Can compute m given c.

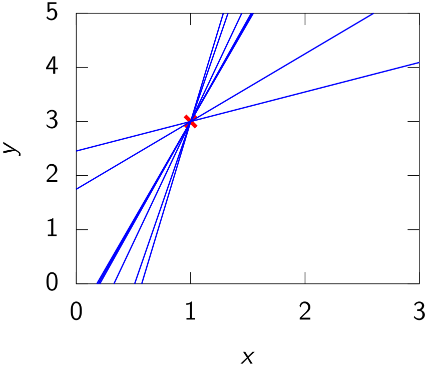
$$c = -3.13 \Longrightarrow m = 6.13$$



Underdetermined System

Can compute m given c.

$$c = -1.47 \implies m = 4.47$$

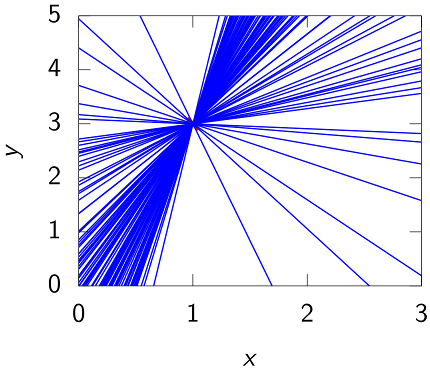


Underdetermined System

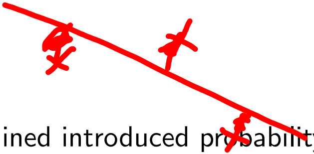
Can compute m given c.
Assume

$$c \sim \mathcal{N}(0, 4),$$

we find a distribution of solutions.



Probability for Under- and Overdetermined



- To deal with overdetermined introduced probability distribution for 'variable', ϵ_i.
- For underdetermined system introduced probability distribution for 'parameter', c.
- This is known as a Bayesian treatment.

- For general Bayesian inference need multivariate priors.
- E.g. for multivariate linear regression:

$$y_i = \sum_i w_j x_{i,j} + \epsilon_i$$

(where we've dropped c for convenience), we need a prior over $\mathbf{w}$.
- This motivates a *multivariate* Gaussian density.
- We will use the multivariate Gaussian to put a prior *directly* on the function (a Gaussian process).

- For general Bayesian inference need multivariate priors.
- E.g. for multivariate linear regression:

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

  (where we've dropped c for convenience), we need a prior over $\mathbf{w}$.
- This motivates a *multivariate* Gaussian density.
- We will use the multivariate Gaussian to put a prior *directly* on the function (a Gaussian process).

Multivariate Regression Likelihood

- Recall multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

- Now use a multivariate Gaussian prior:

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

Multivariate Regression Likelihood

- Recall multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

- Now use a multivariate Gaussian prior:

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^\top \mathbf{w}$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const.}$$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_w, \mathbf{C}_w)$$

$$\mathbf{C}_w = (\sigma^{-2}\mathbf{X}^\top \mathbf{X} + \alpha^{-1})^{-1} \text{ and } \boldsymbol{\mu}_w = \mathbf{C}_w \sigma^{-2}\mathbf{X}^\top \mathbf{y}$$

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^\top \mathbf{w}$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const.}$$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_w, \mathbf{C}_w\right)$$

$\mathbf{C}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1})^{-1}$ and $\boldsymbol{\mu}_w = \mathbf{C}_w \sigma^{-2}\mathbf{X}^\top\mathbf{y}$

Bayesian vs Maximum Likelihood

- Note the similarity between posterior mean

$$\boldsymbol{\mu}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1})^{-1}\sigma^{-2}\mathbf{X}^\top\mathbf{y}$$

- and Maximum likelihood solution

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Two Dimensional Gaussian

h, w

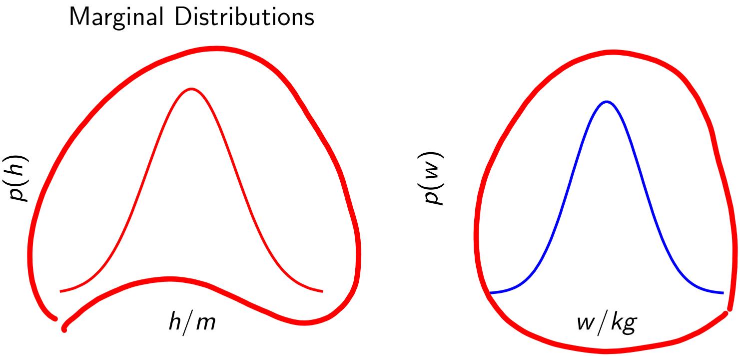
- Consider height, h/m and weight, w/kg.
- Could sample height from a distribution:

$$p(h) \sim \mathcal{N}(1.7, 0.0225)$$

- And similarly weight:

$$p(w) \sim \mathcal{N}(75, 36)$$

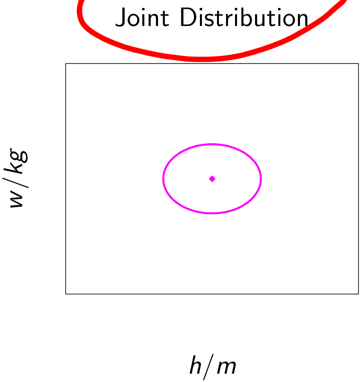
Height and Weight Models

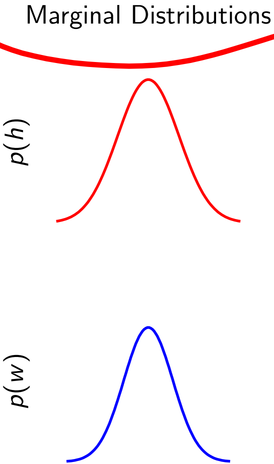


Gaussian

distributions for height and weight.

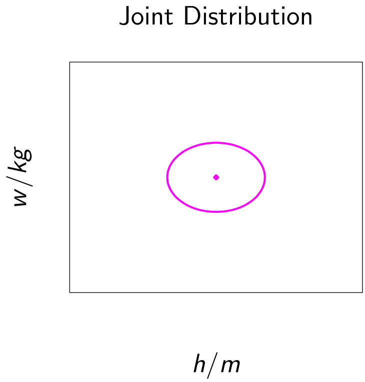
Sampling Two Dimensional Variables

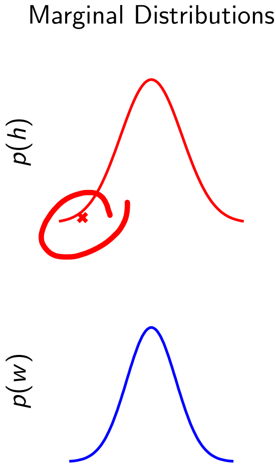




Sample height and weight one after the other and plot against each other.

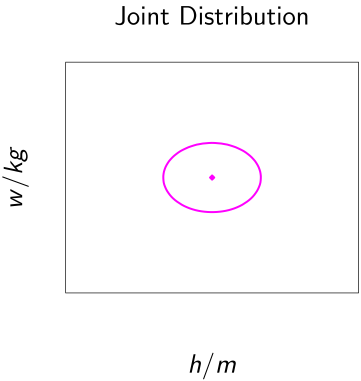
Sampling Two Dimensional Variables

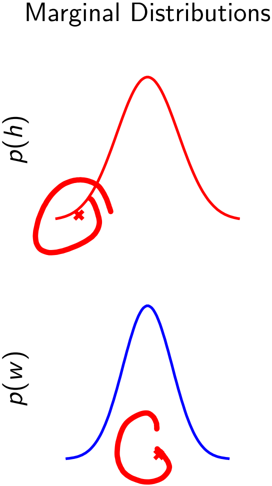




Sample height and weight one after the other and plot against each other.

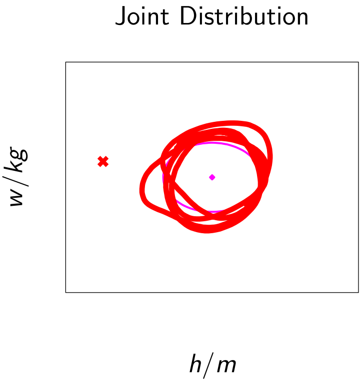
Sampling Two Dimensional Variables

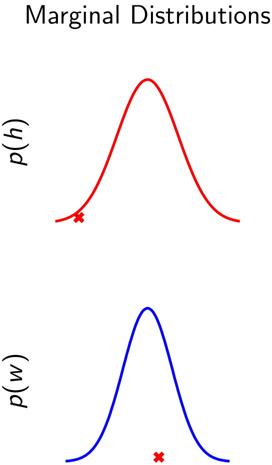




Sample height and weight one after the other and plot against each other.

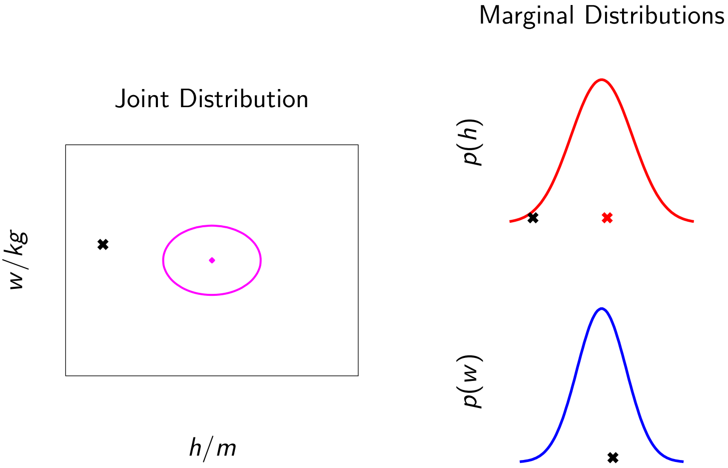
Sampling Two Dimensional Variables





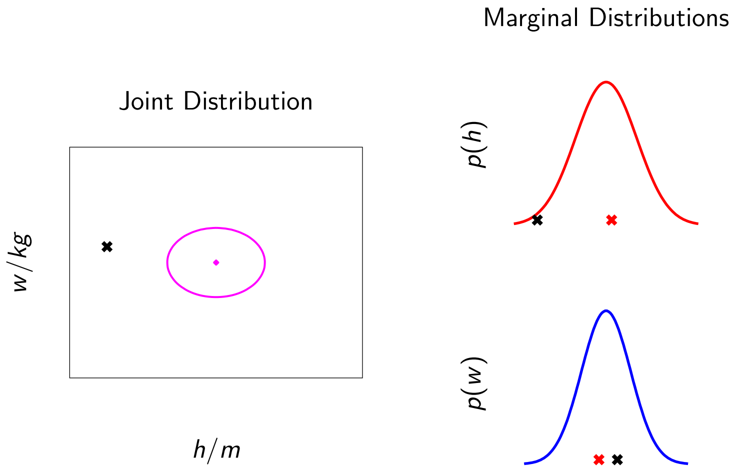
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



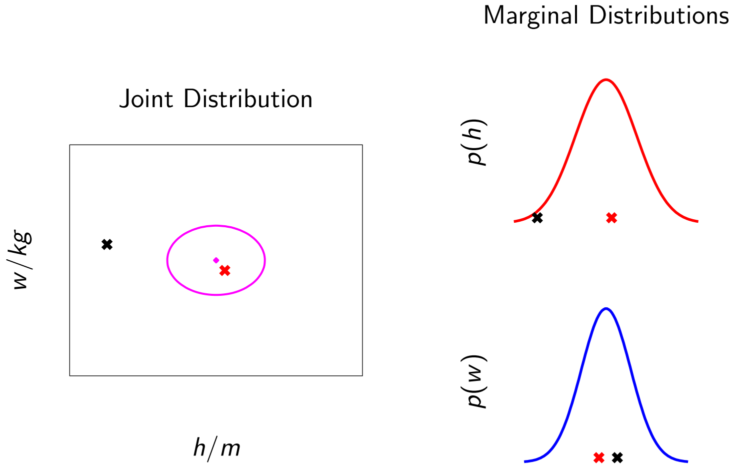
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



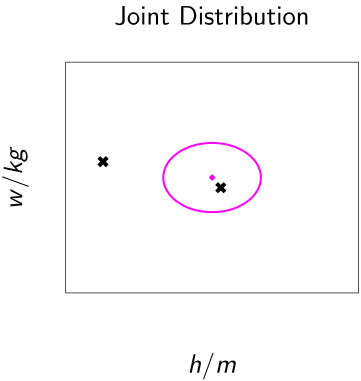
Sample height and weight one after the other and plot against each other.

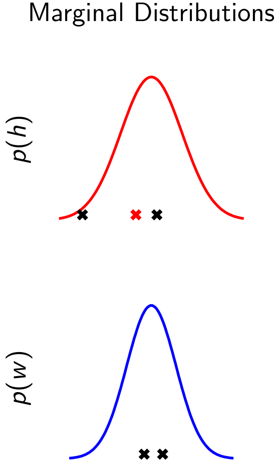
Sampling Two Dimensional Variables



Sample height and weight one after the other and plot against each other.

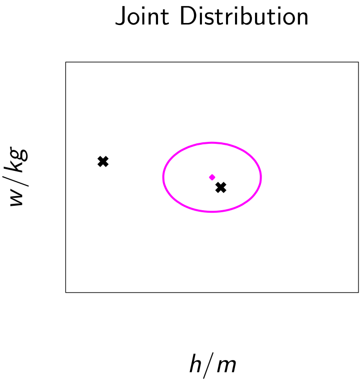
Sampling Two Dimensional Variables

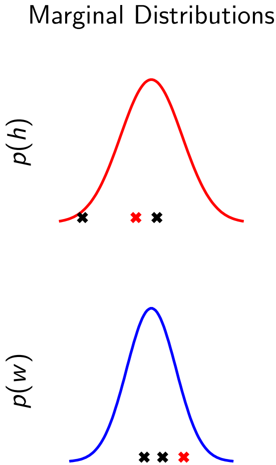




Sample height and weight one after the other and plot against each other.

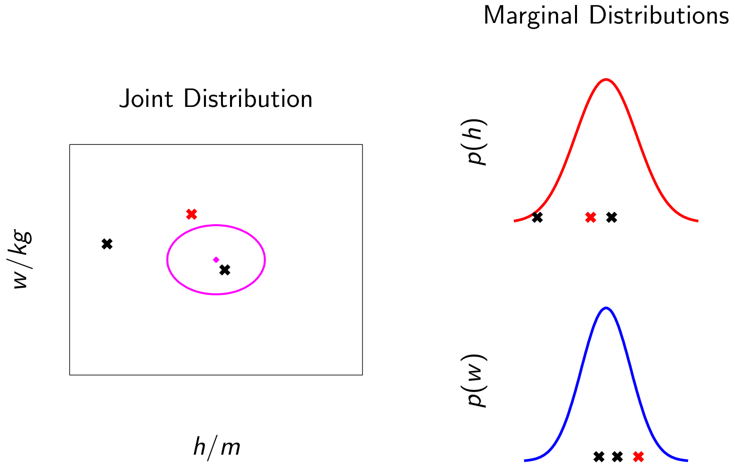
Sampling Two Dimensional Variables





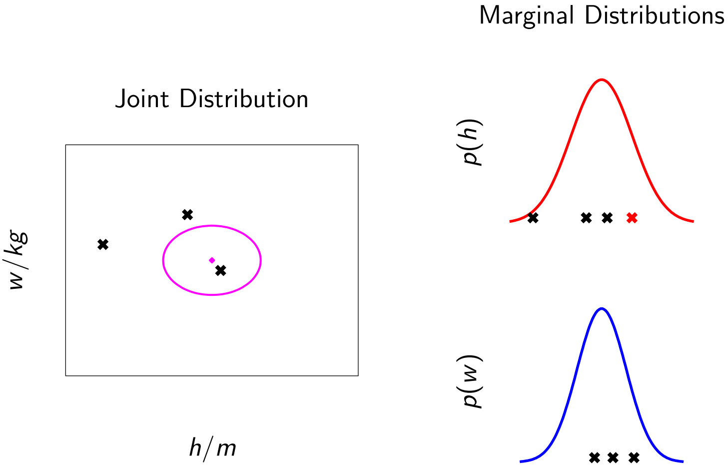
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



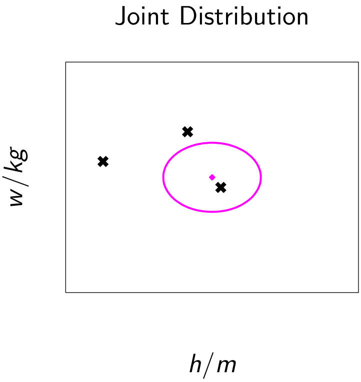
Sample height and weight one after the other and plot against each other.

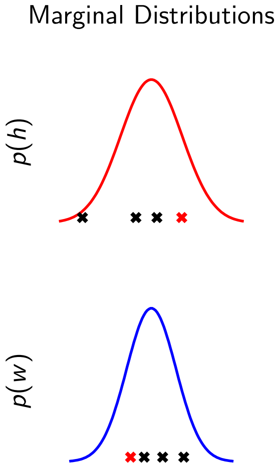
Sampling Two Dimensional Variables



Sample height and weight one after the other and plot against each other.

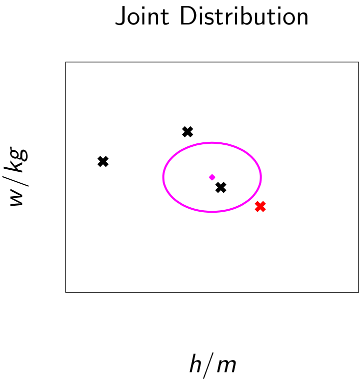
Sampling Two Dimensional Variables

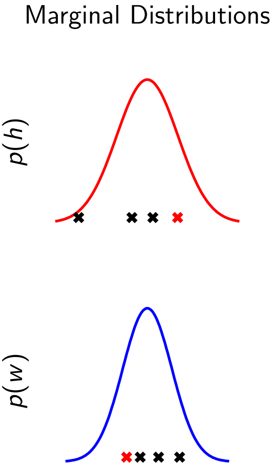




Sample height and weight one after the other and plot against each other.

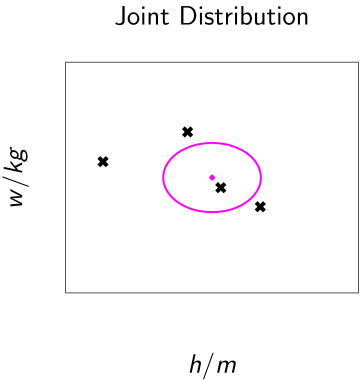
Sampling Two Dimensional Variables

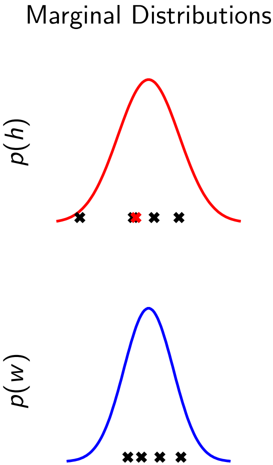




Sample height and weight one after the other and plot against each other.

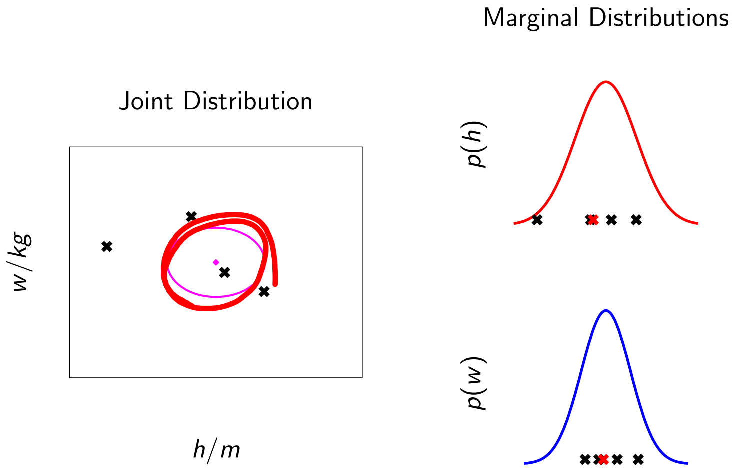
Sampling Two Dimensional Variables





Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables

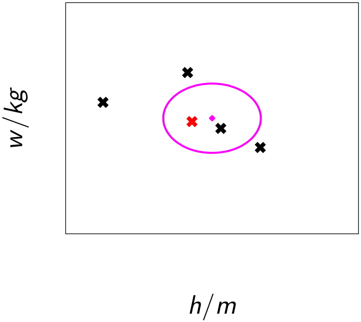


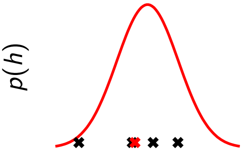
Sample height and weight one after the other and plot against each other.

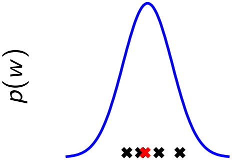
Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution

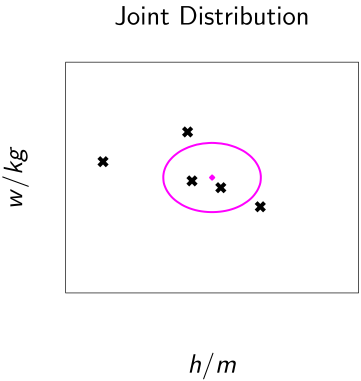


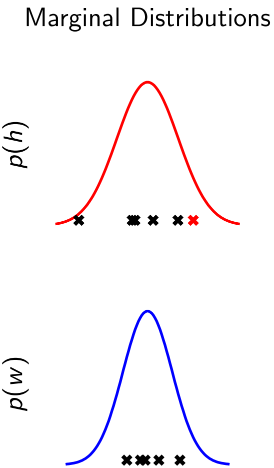




Sample height and weight one after the other and plot against each other.

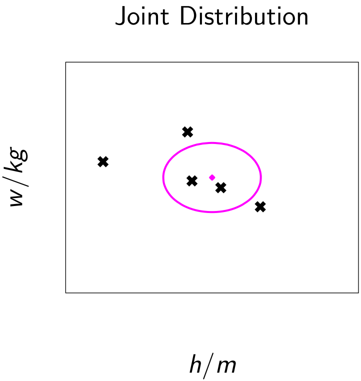
Sampling Two Dimensional Variables

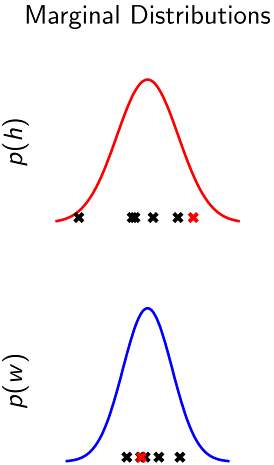




Sample height and weight one after the other and plot against each other.

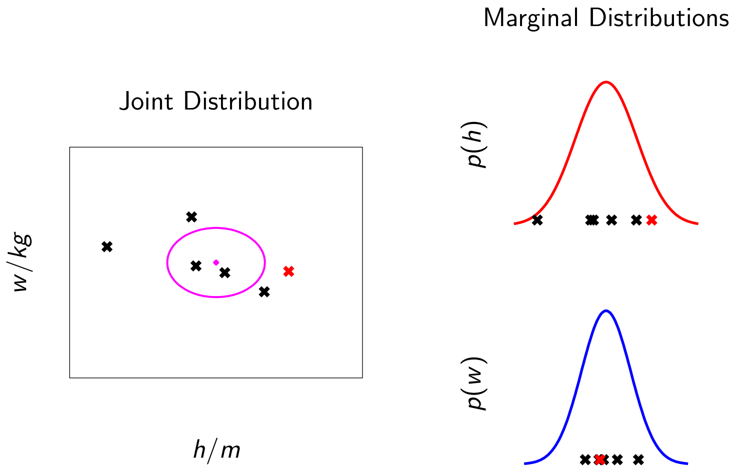
Sampling Two Dimensional Variables





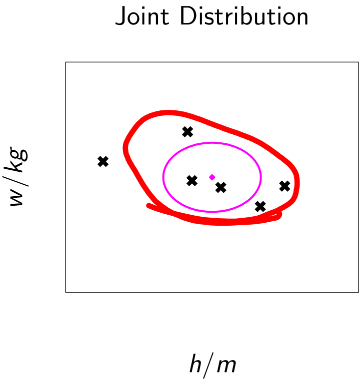
Sample height and weight one after the other and plot against each other.

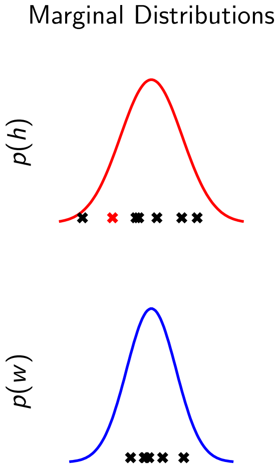
Sampling Two Dimensional Variables



Sample height and weight one after the other and plot against each other.

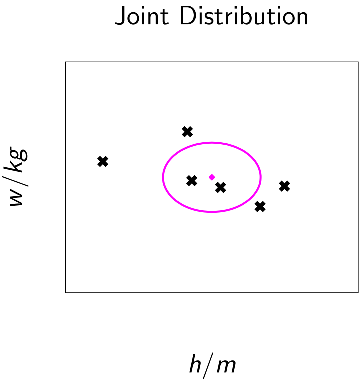
Sampling Two Dimensional Variables

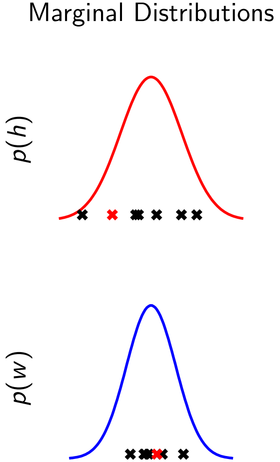




Sample height and weight one after the other and plot against each other.

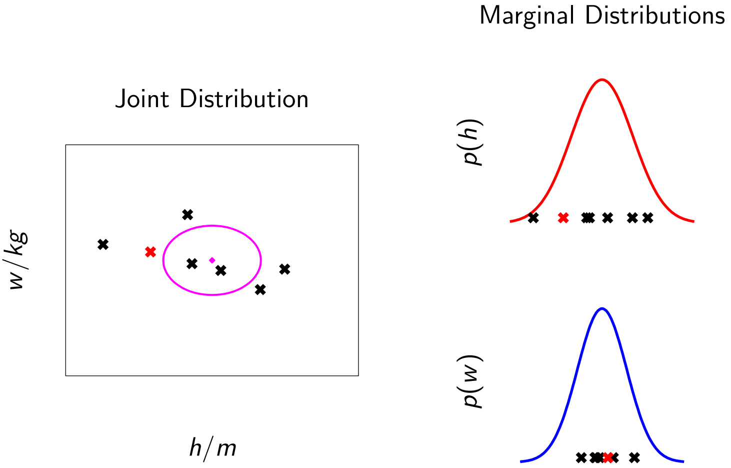
Sampling Two Dimensional Variables





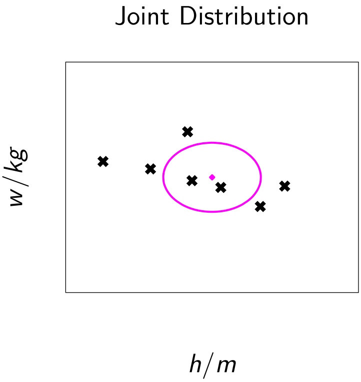
Sample height and weight one after the other and plot against each other.

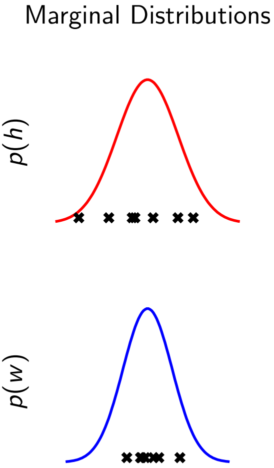
Sampling Two Dimensional Variables



Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables





Sample height and weight one after the other and plot against each other.

Independence Assumption

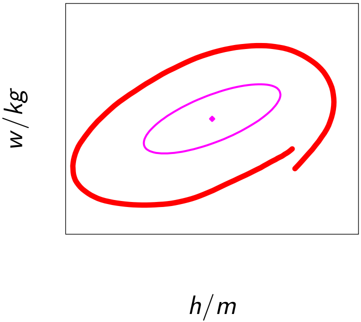
- This assumes height and weight are independent.

$$p(h, w) = p(h)p(w)$$

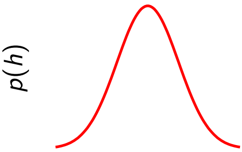
- In reality they are dependent (body mass index) $= \frac{w}{h^2}$.

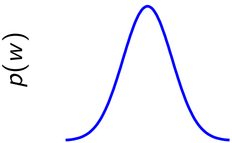
Sampling Two Dimensional Variables

Joint Distribution

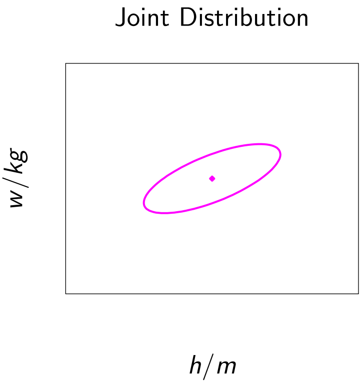


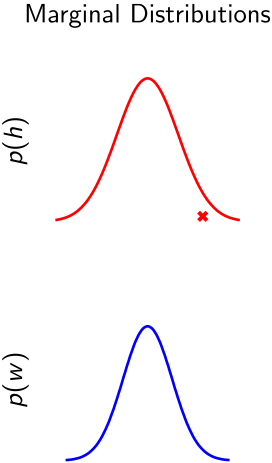
Marginal Distributions



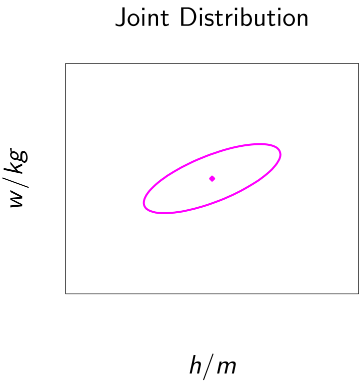


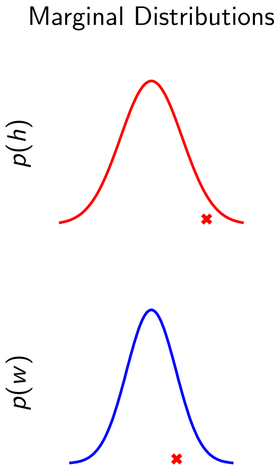
Sampling Two Dimensional Variables



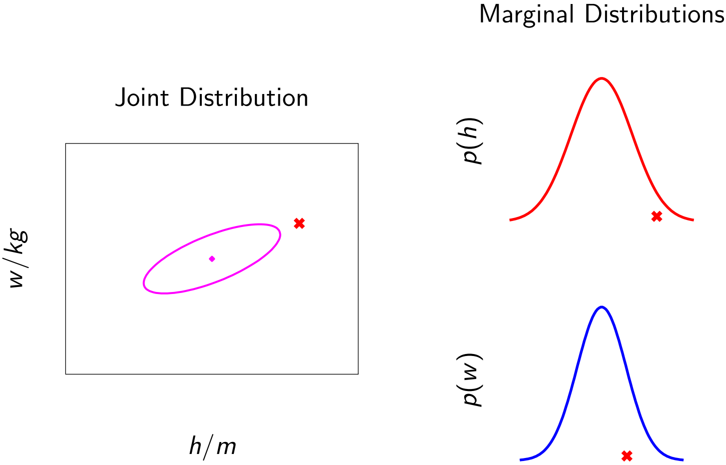


Sampling Two Dimensional Variables

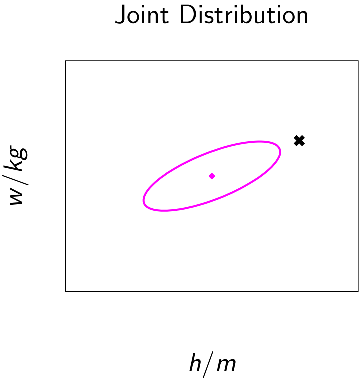


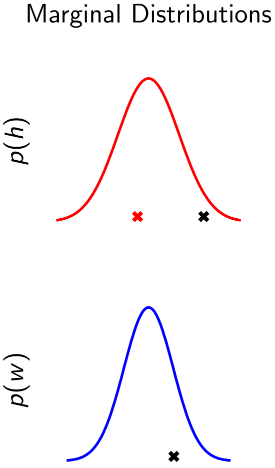


Sampling Two Dimensional Variables

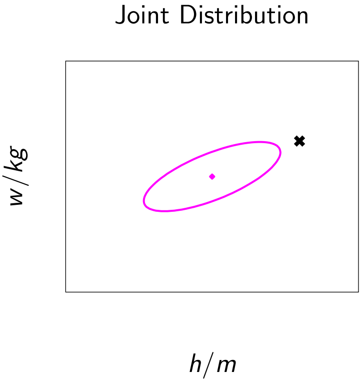


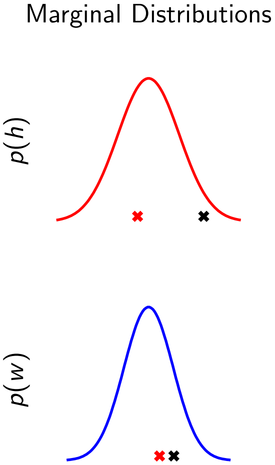
Sampling Two Dimensional Variables





Sampling Two Dimensional Variables

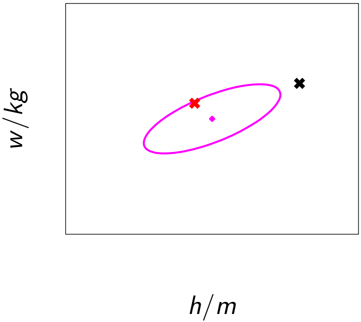


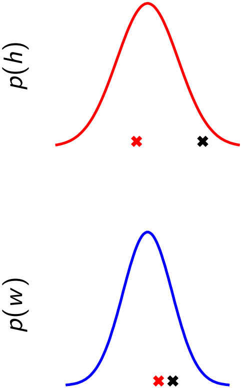


Sampling Two Dimensional Variables

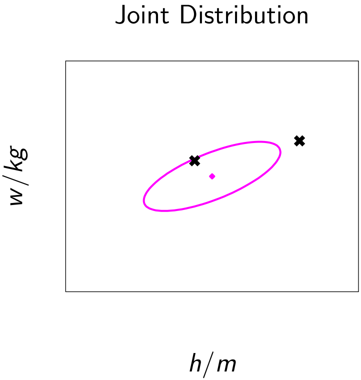
Marginal Distributions

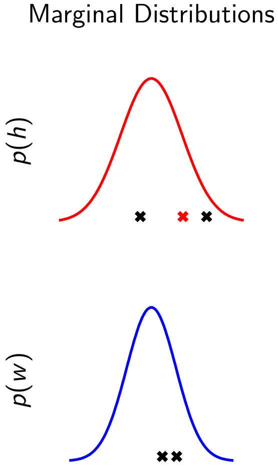
Joint Distribution



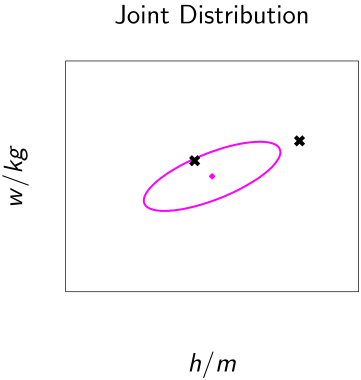


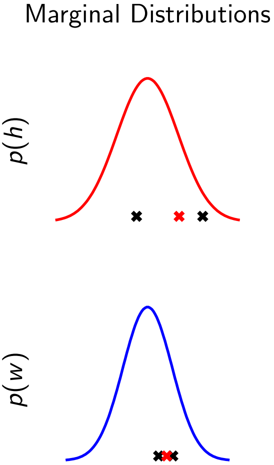
Sampling Two Dimensional Variables



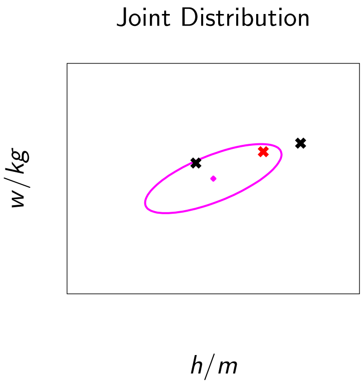


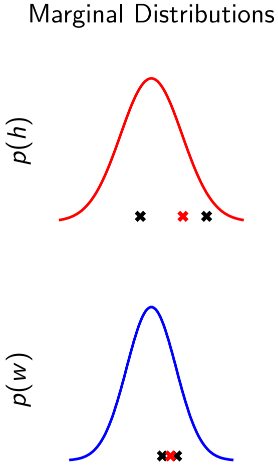
Sampling Two Dimensional Variables



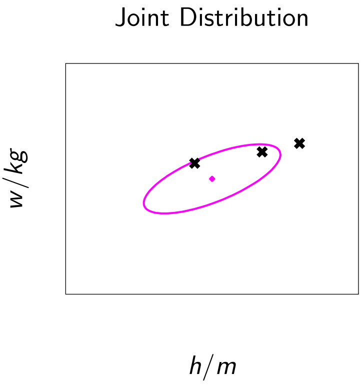


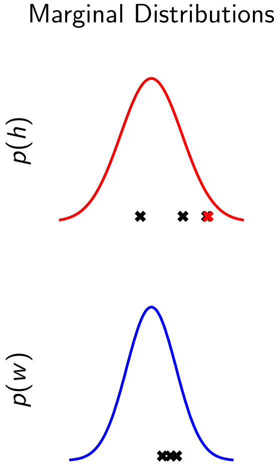
Sampling Two Dimensional Variables



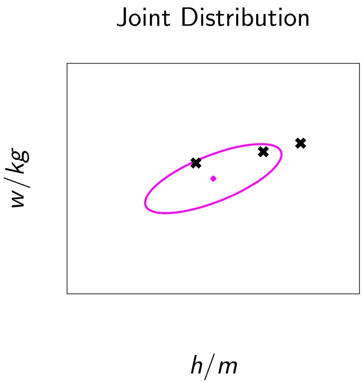


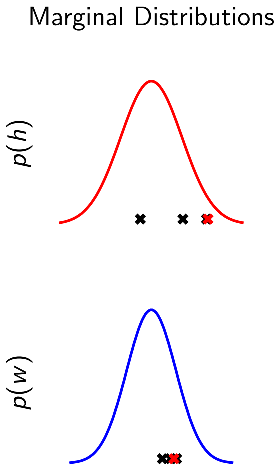
Sampling Two Dimensional Variables



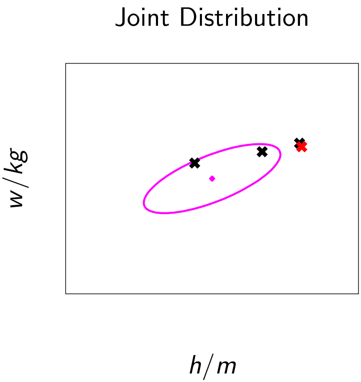


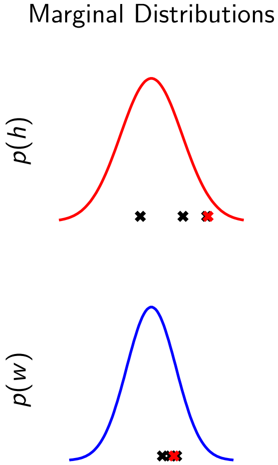
Sampling Two Dimensional Variables



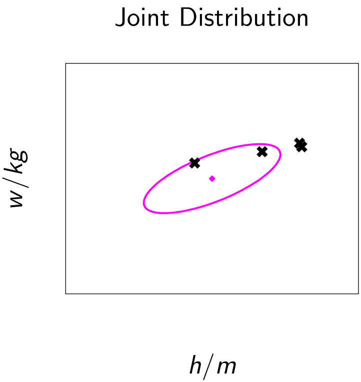


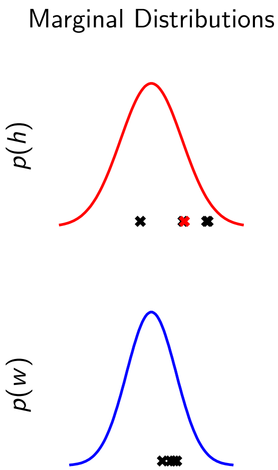
Sampling Two Dimensional Variables



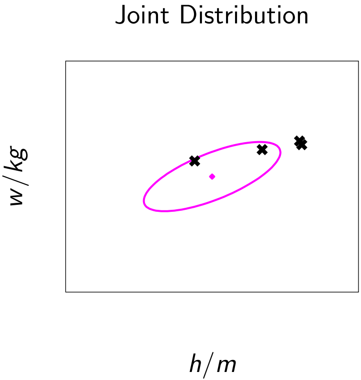


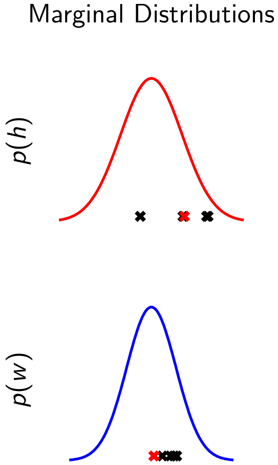
Sampling Two Dimensional Variables



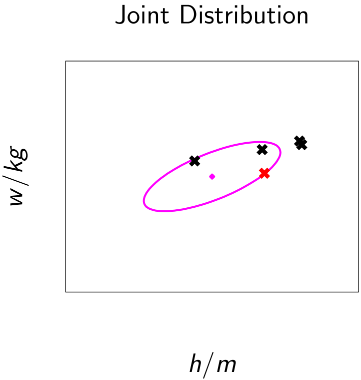


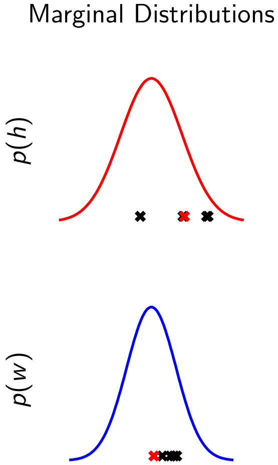
Sampling Two Dimensional Variables



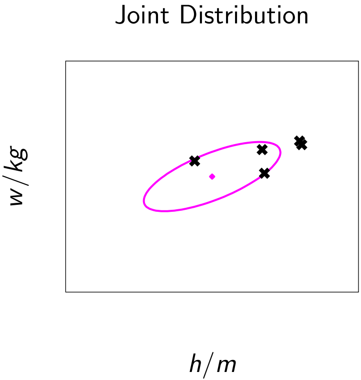


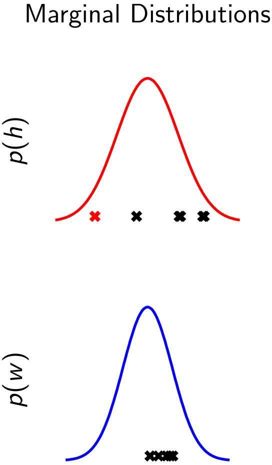
Sampling Two Dimensional Variables



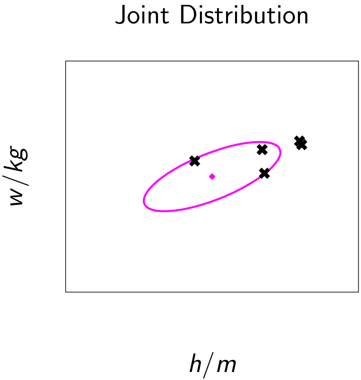


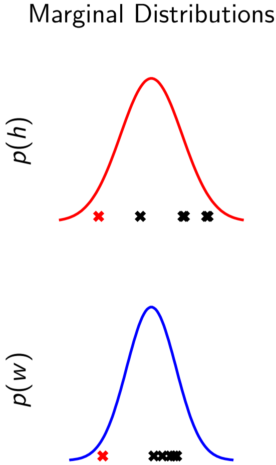
Sampling Two Dimensional Variables



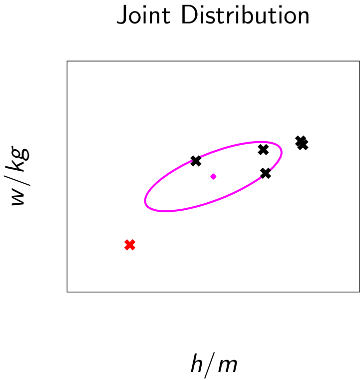


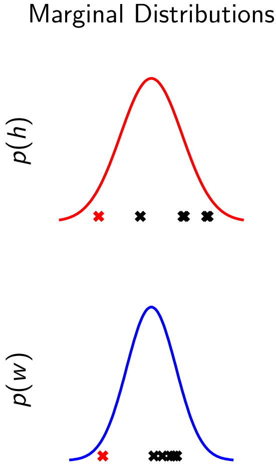
Sampling Two Dimensional Variables





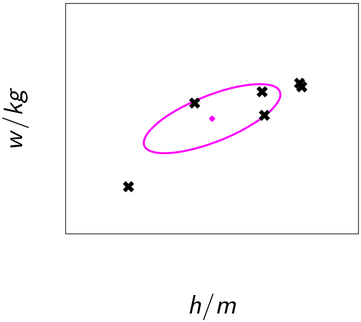
Sampling Two Dimensional Variables



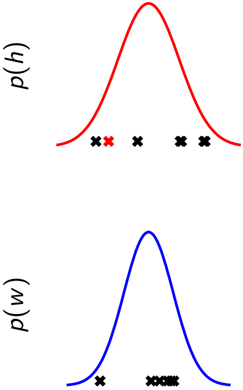


Sampling Two Dimensional Variables

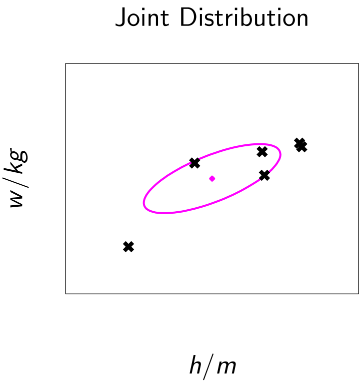
Joint Distribution

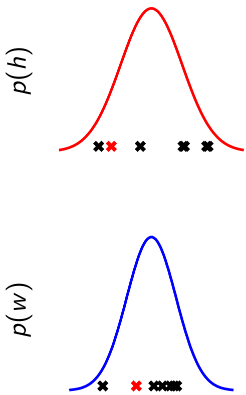


Marginal Distributions

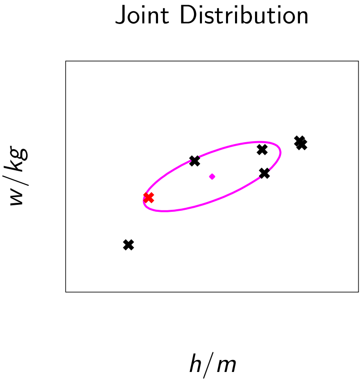


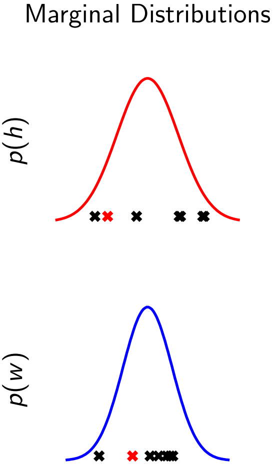
Sampling Two Dimensional Variables



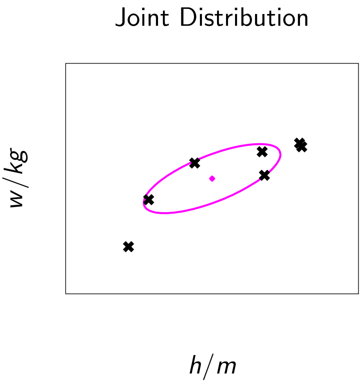


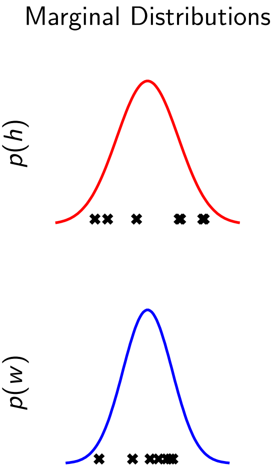
Sampling Two Dimensional Variables





Sampling Two Dimensional Variables





Independent Gaussians

$$p(w, h) = p(w)p(h)$$

Independent Gaussians

$$p(w, h) = \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{(w-\mu_1)^2}{\sigma_1^2} + \frac{(h-\mu_2)^2}{\sigma_2^2}\right)\right)$$

Independent Gaussians

$$p(w, h) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^\top \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)\right)$$

Independent Gaussians

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{R}^\top \mathbf{y} - \mathbf{R}^\top \boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{R}^\top \mathbf{y} - \mathbf{R}^\top \boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C}^{-1} = \mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{C}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{C}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C} = \mathbf{R}\mathbf{D}\mathbf{R}^{\top}$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Multivariate Consequence

- If
$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$
- And
$$\mathbf{y} = \mathbf{W}\mathbf{x}$$
- Then
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\mu, \mathbf{W}\Sigma\mathbf{W}^{\top}\right)$$

Multivariate Consequence

- If
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
- And
$$\mathbf{y} = \mathbf{W}\mathbf{x}$$
- Then
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\mu, \mathbf{W}\Sigma\mathbf{W}^{\top}\right)$$

Multivariate Consequence

- If

$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

- And

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

- Then

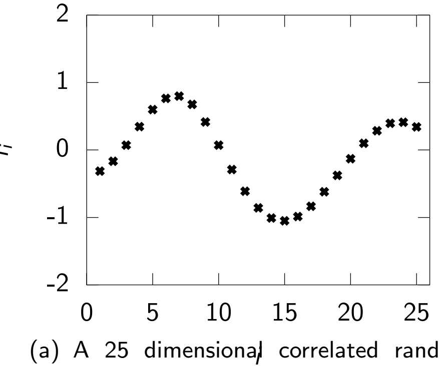
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\mu, \mathbf{W}\Sigma\mathbf{W}^{\top}\right)$$

Sampling a Function

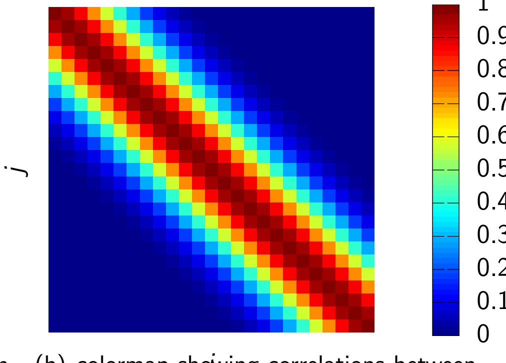
Multi-variate Gaussians

- We will consider a Gaussian with a particular structure of covariance matrix.
- Generate a single sample from this 25 dimensional Gaussian distribution, $\mathbf{f} = [f_1, f_2 \ldots f_{25}]$.
- We will plot these points against their index.

Gaussian Distribution Sample



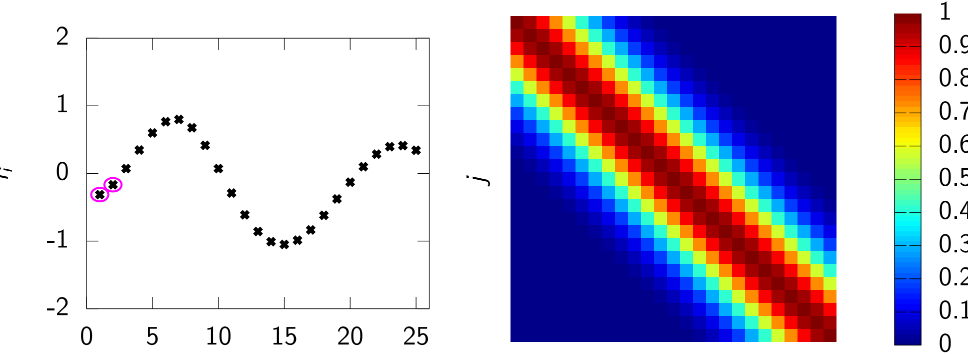
(a) A 25 dimensional correlated random variable (values ploted against index)



(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

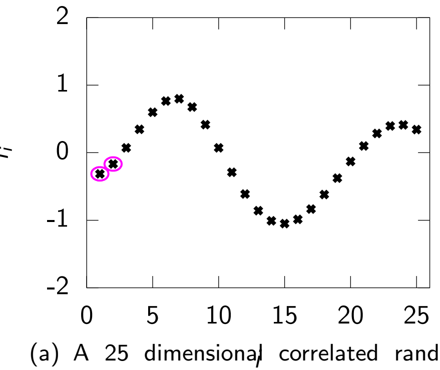


(a) A 25 dimensional correlated random variable (values ploted against index)

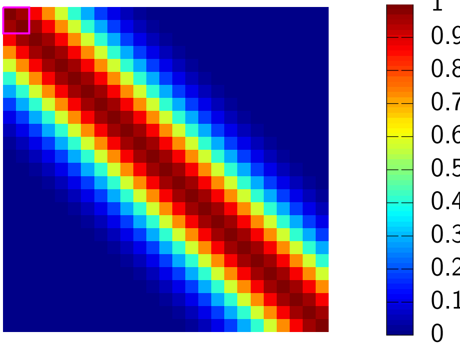
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



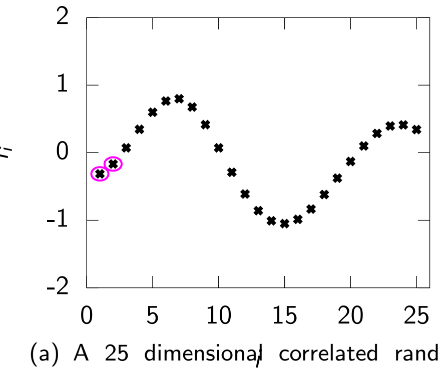
(a) A 25 dimensional correlated random variable (values ploted against index)



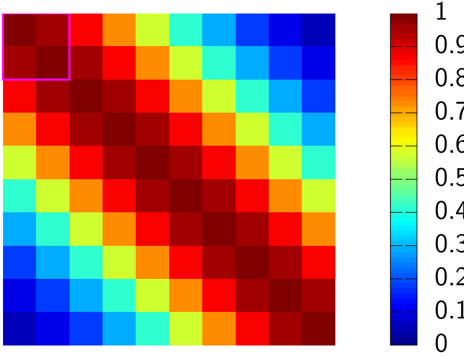
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



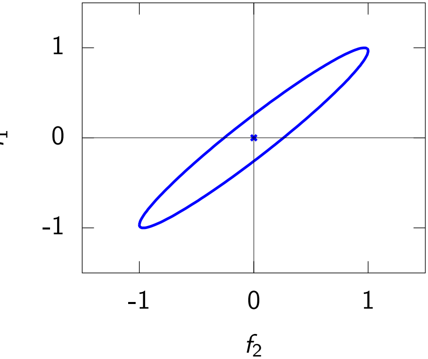
(a) A 25 dimensional correlated random variable (values ploted against index)



(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

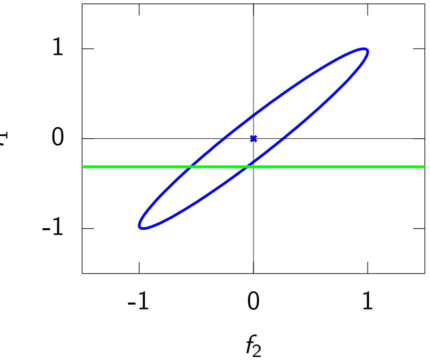
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2|f_1 = -0.313)$.

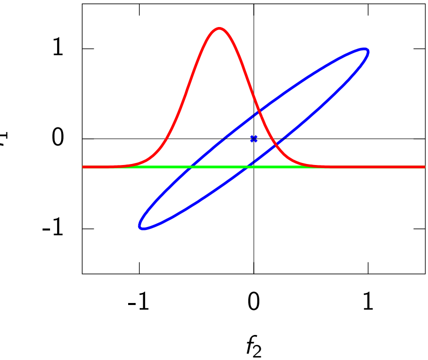
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2|f_1 = -0.313)$.

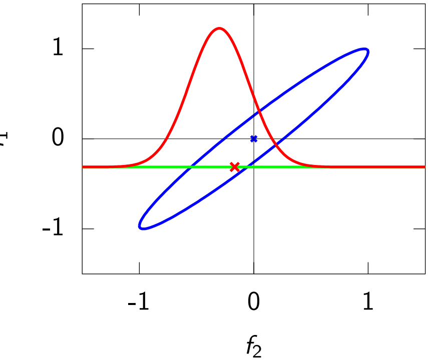
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2|f_1 = -0.313)$.

Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2|f_1 = -0.313)$.

Prediction with Correlated Gaussians

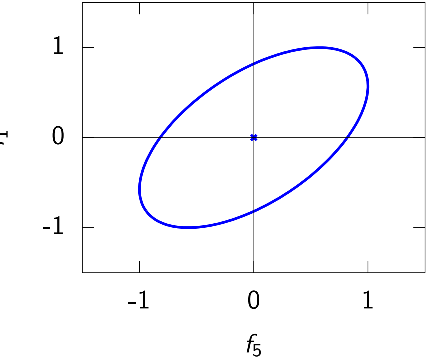
- Prediction of f_2 from f_1 requires *conditional density*.
- Conditional density is *also* Gaussian.

$$p(f_2|f_1) = \mathcal{N}\left(f_2|\frac{k_{1,2}}{k_{1,1}}f_1, k_{2,2} - \frac{k_{1,2}^2}{k_{1,1}}\right)$$

where covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} k_{1,1} & k_{1,2} \\ k_{2,1} & k_{2,2} \end{bmatrix}$$

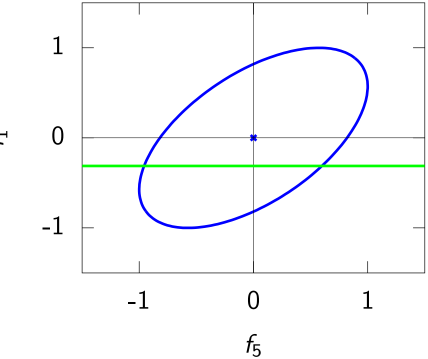
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5 | f_1 = -0.313)$.

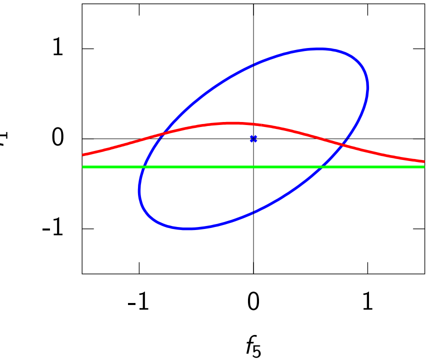
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$.

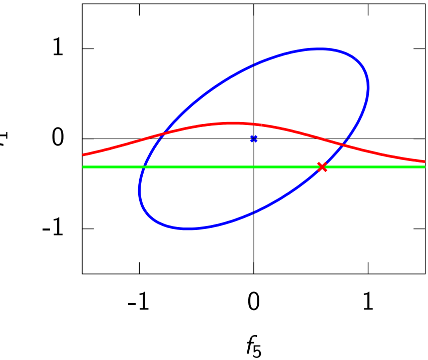
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5 | f_1 = -0.313)$.

Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5 | f_1 = -0.313)$.

Prediction with Correlated Gaussians

- Prediction of $\mathbf{f}_*$ from $\mathbf{f}$ requires multivariate *conditional density*.
- Multivariate conditional density is *also* Gaussian.

$$p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}\left(\mathbf{f}_*|\mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f}, \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}\right)$$

- Here covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{*,\mathbf{f}} \\ \mathbf{K}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix}$$

Prediction with Correlated Gaussians

- Prediction of $\mathbf{f}_*$ from $\mathbf{f}$ requires multivariate *conditional density*.
- Multivariate conditional density is *also* Gaussian.

$$p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}\left(\mathbf{f}_*|\boldsymbol{\mu}, \Sigma\right)$$

$$\boldsymbol{\mu} = \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f}$$

$$\Sigma = \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}$$

- Here covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{*,\mathbf{f}} \\ \mathbf{K}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix}$$

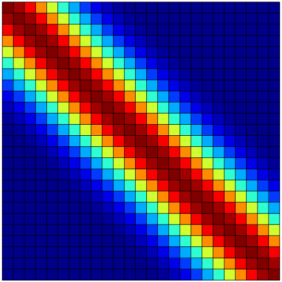
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



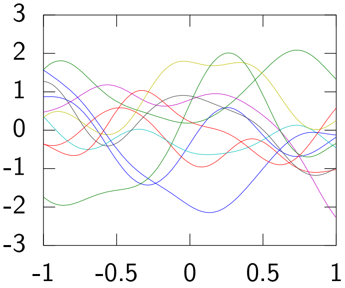
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



Gaussian Process Interpolation

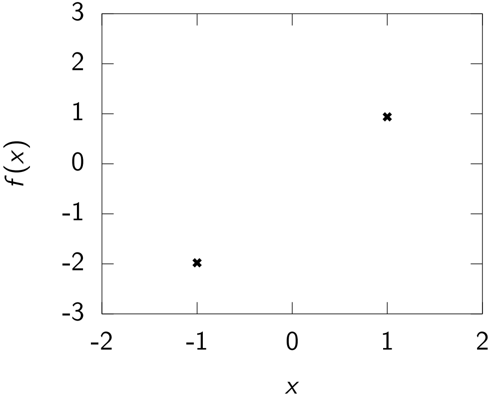


Figure: Real example: BACCO (see *e.g.* (**?**)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

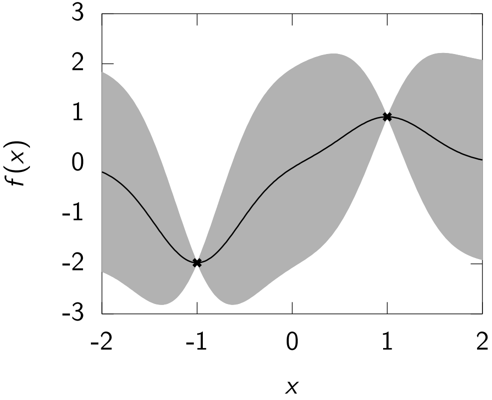


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

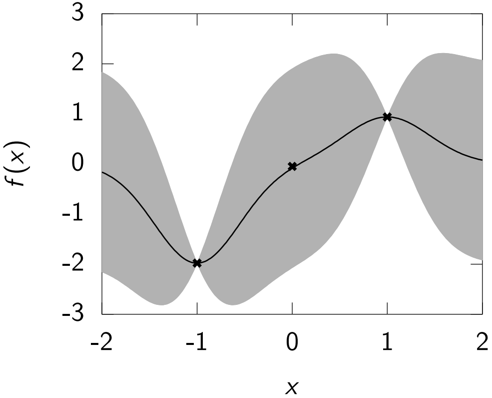


Figure: Real example: BACCO (see *e.g.* (**?**)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

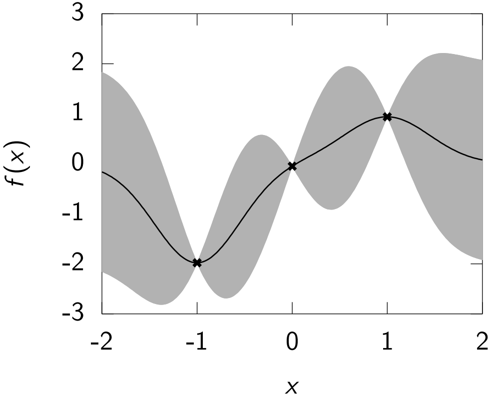


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

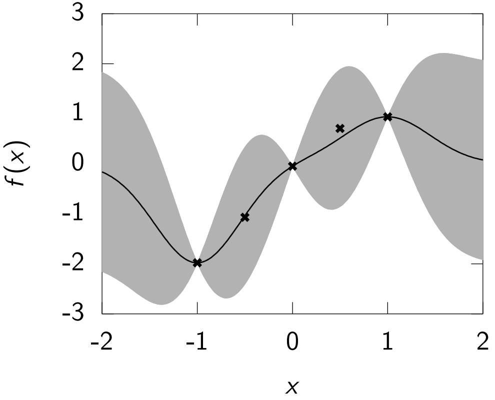


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

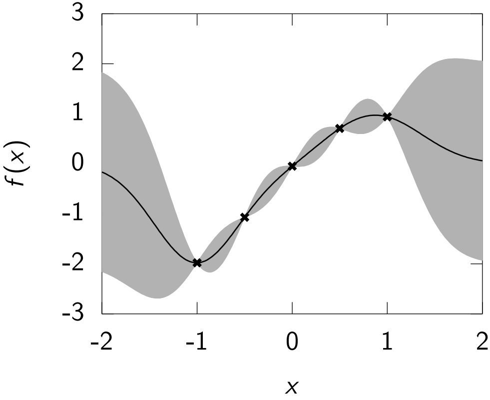


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

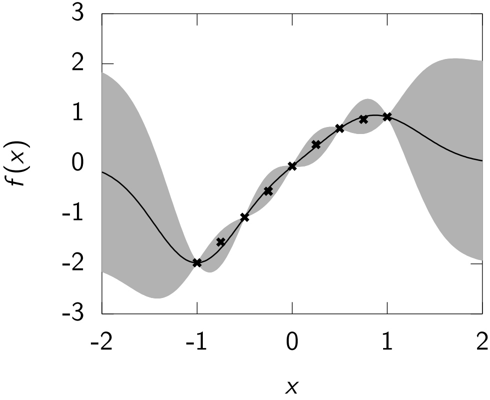


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

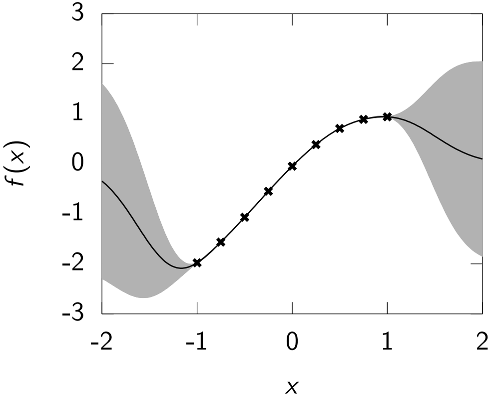


Figure: Real example: BACCO (see *e.g.* (**?**)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Noise Models

Graph of a GP

- Relates input variables, $\mathbf{X}$, to vector, $\mathbf{y}$, through $\mathbf{f}$ given kernel parameters $\boldsymbol{\theta}$.
- Plate notation indicates independence of $y_i|f_i$.
- Noise model, $p(y_i|f_i)$ can take several forms.
- Simplest is Gaussian noise.

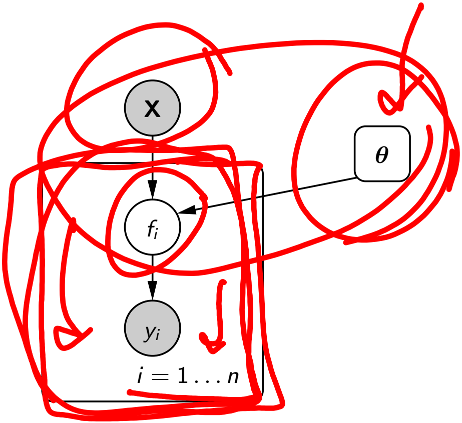


Figure: The Gaussian process depicted graphically.

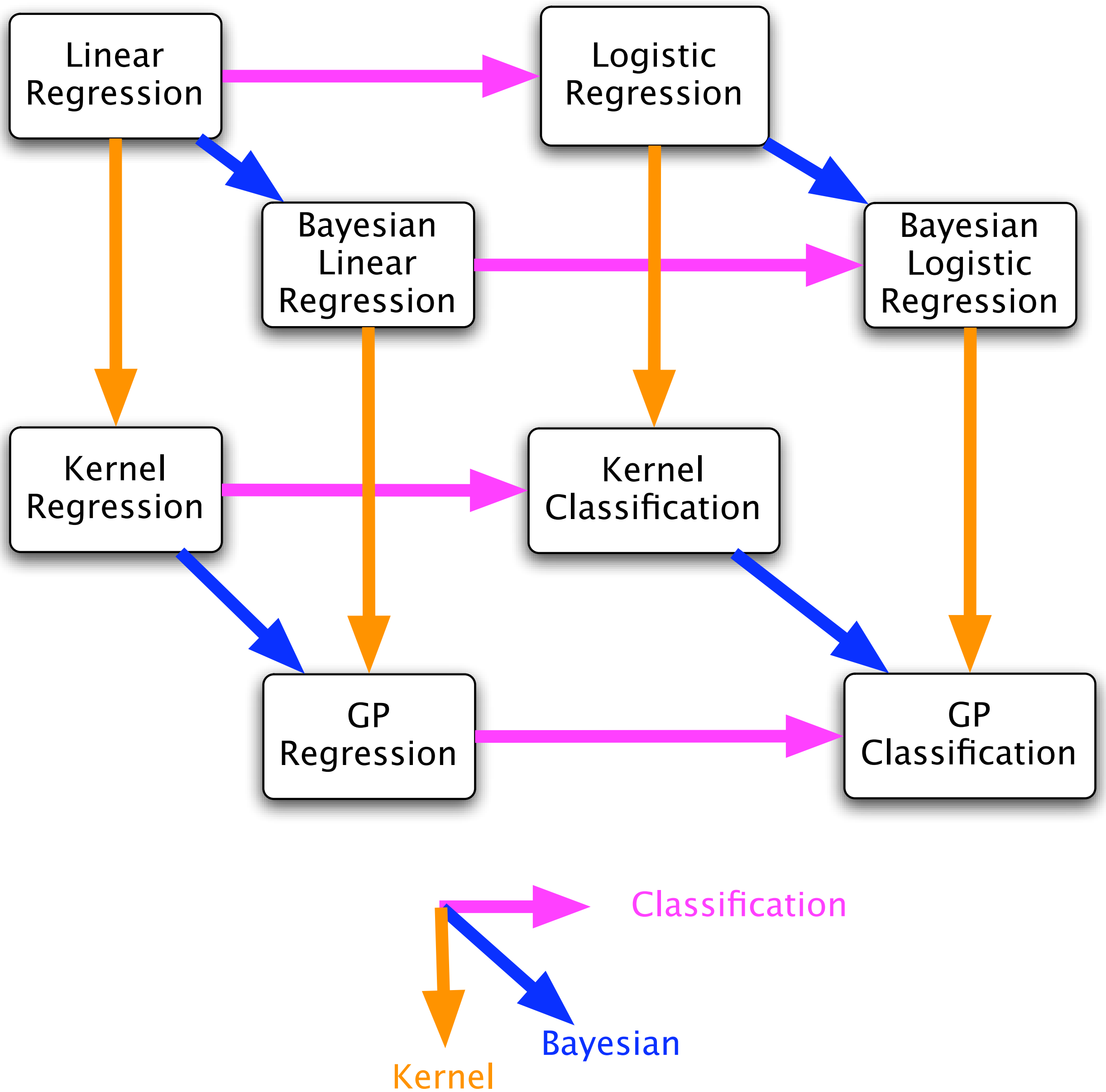
Limitations of Gaussian Processes

- Inference is $O(n^3)$ due to matrix inverse (in practice use Cholesky).
- Gaussian processes don't deal well with discontinuities (financial crises, phosphorylation, collisions, edges in images).
- Widely used exponentiated quadratic covariance (RBF) can be too smooth in practice (but there are many alternatives!!).

Summary

- Broad introduction to Gaussian processes.
  - Started with Gaussian distribution.
  - Motivated Gaussian processes through the multivariate density.
- Emphasized the role of the covariance (not the mean).
- Performs nonlinear regression with error bars.
- Parameters of the covariance function (kernel) are easily optimized with maximum likelihood.

A picture



Linear Regression
Logistic Regression
Bayesian Linear Regression
Bayesian Logistic Regression
Kernel Regression
Kernel Classification
GP Regression
GP Classification
Classification
Bayesian
Kernel