Multivariate Bayesian Linear Regression

MLAI Lecture 11

Neil D. Lawrence

Department of Computer Science
Sheffield University

21st October 2012

Outline

Univariate Bayesian Linear Regression

Multivariate Bayesian Linear Regression

Prior Distribution

- Bayesian inference requires a prior on the parameters.
- The prior represents your belief *before* you see the data of the likely value of the parameters.
- For linear regression, consider a Gaussian prior on the intercept:

$$c \sim \mathcal{N}(0, \alpha_1)$$

Gaussian Noise

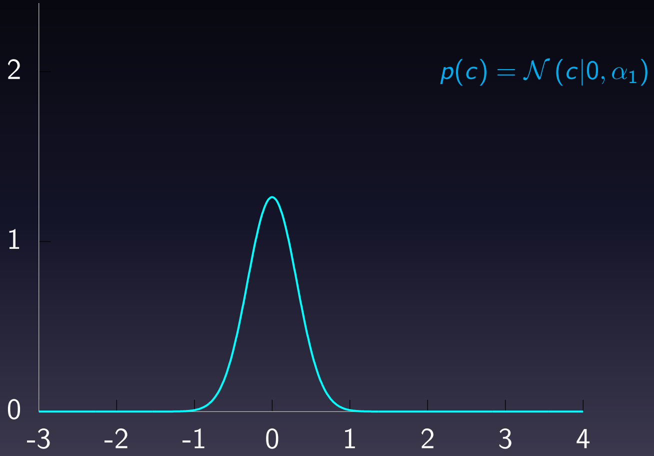


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Gaussian Noise

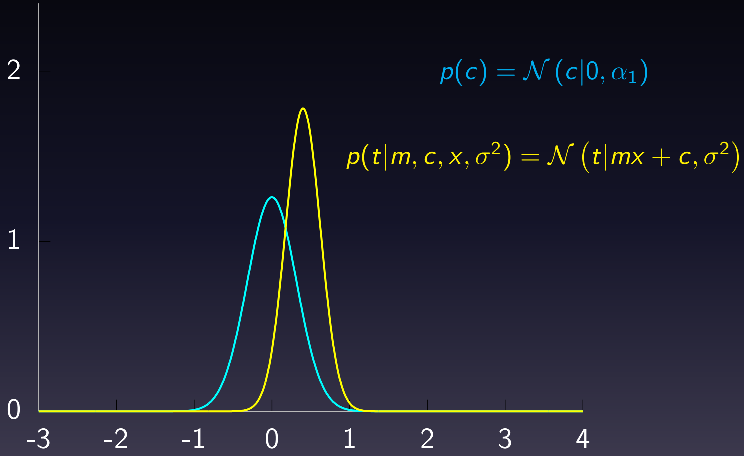


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Gaussian Noise

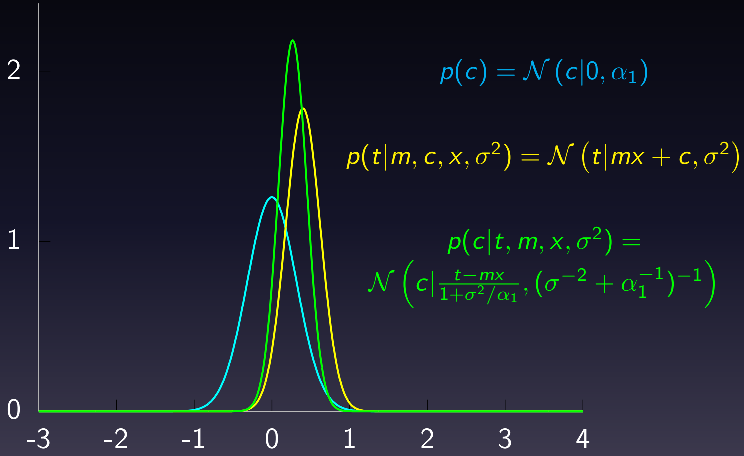


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Stages to Derivation of the Posterior

- Multiply likelihood by prior
  - they are "exponentiated quadratics", the answer is always also an exponentiated quadratic because $\exp(a^2)\exp(b^2) = \exp(a^2 + b^2)$.
- Complete the square to get the resulting density in the form of a Gaussian.
- Recognise the mean and (co)variance of the Gaussian. This is the estimate of the posterior.

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1} c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (t_i - m x_i - c)^2\right)$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2)p(c)}{p(\mathbf{t}|\mathbf{x}, m, \sigma^2)}$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2)p(c)}{\int p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2)p(c)\mathrm{d}c}$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1} c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (t_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) \propto p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) p(c)$$

$$
\begin{aligned}
\log p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - c - mx_i)^2 - \frac{1}{2\alpha_1}c^2 + \text{const} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - mx_i)^2 - \left(\frac{N}{2\sigma^2} + \frac{1}{2\alpha_1}\right)c^2 \\
&\quad + c\frac{\sum_{i=1}^{N}(t_i - mx_i)}{\sigma^2},
\end{aligned}
$$

complete the square of the quadratic form to obtain

$$
\log p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) = -\frac{1}{2\tau^2}(c - \mu)^2 + \text{const},
$$

where $\tau^2 = \left(N\sigma^{-2} + \alpha_1^{-1}\right)^{-1}$ and $\mu = \frac{\tau^2}{\sigma^2}\sum_{n=1}^{N}(t_i - mx_i)$.

The Joint Density

- Really want to know the *joint* posterior density over the parameters c *and* m.
- Could now integrate out over m, but it's easier to consider the multivariate case.

Two Dimensional Gaussian

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

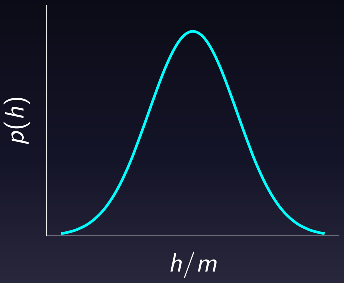
$$p(h) \sim \mathcal{N}(1.7, 0.0225)$$

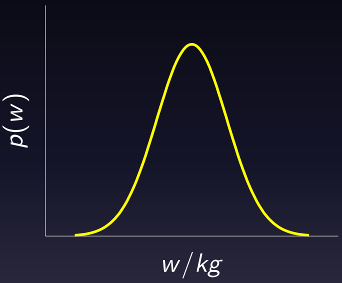
- And similarly weight:

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

Height and Weight Models

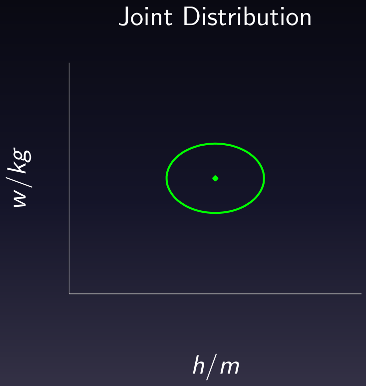
Marginal Distributions





Gaussian distributions for height and weight.

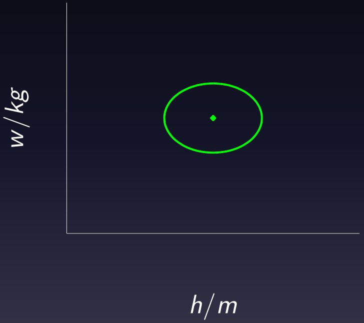
Sampling Two Dimensional Variables



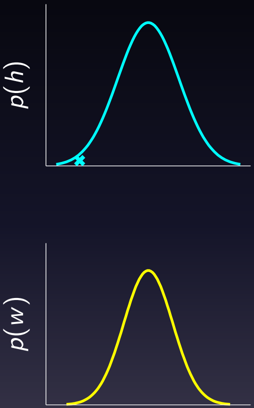
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables

Joint Distribution



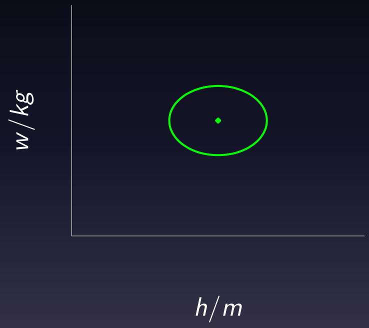
Marginal Distributions



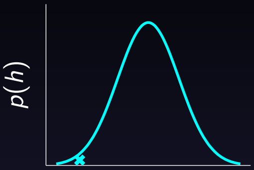
Sample height and weight one after the other and plot against each other.

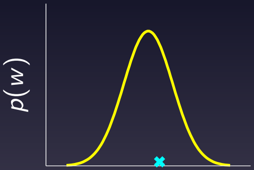
Sampling Two Dimensional Variables

Joint Distribution



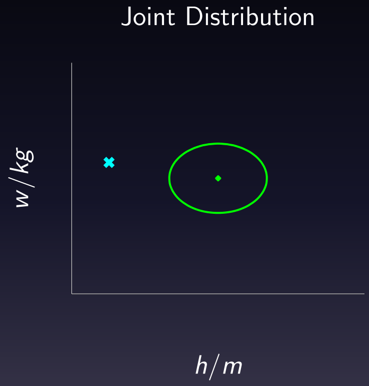
Marginal Distributions

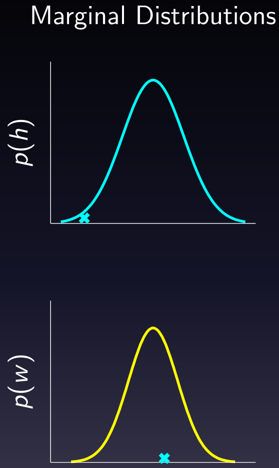




Sample height and weight one after the other and plot against each other.

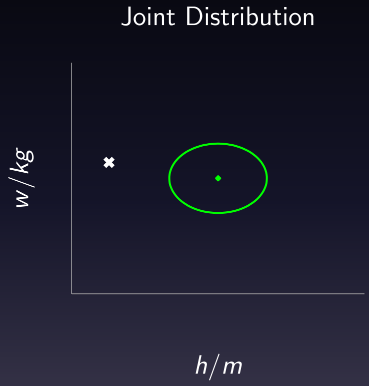
Sampling Two Dimensional Variables

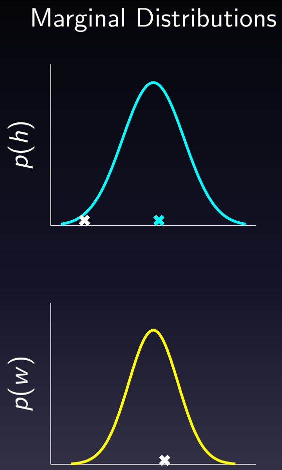




Sample height and weight one after the other and plot against each other.

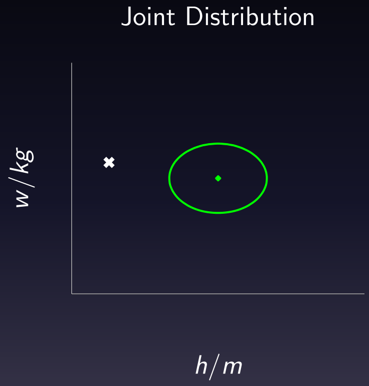
Sampling Two Dimensional Variables

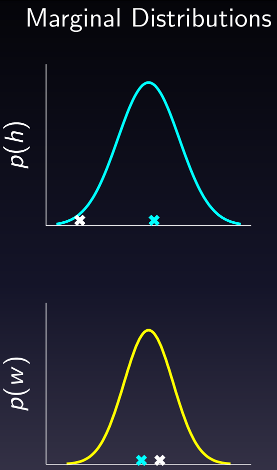




Sample height and weight one after the other and plot against each other.

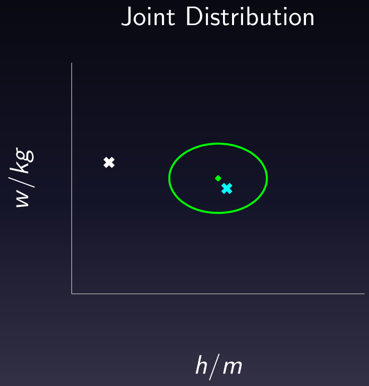
Sampling Two Dimensional Variables

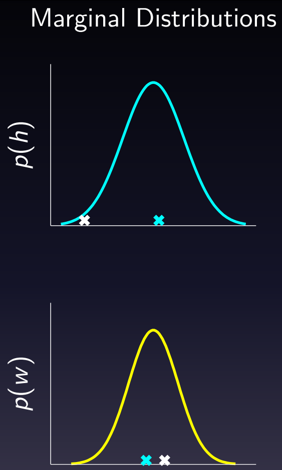




Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables

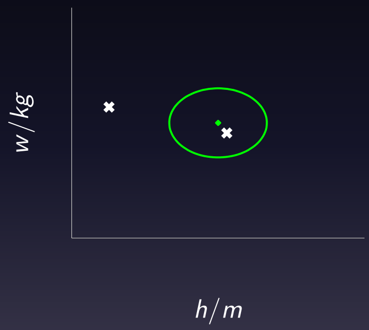




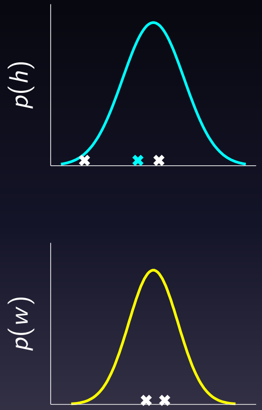
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables

Joint Distribution



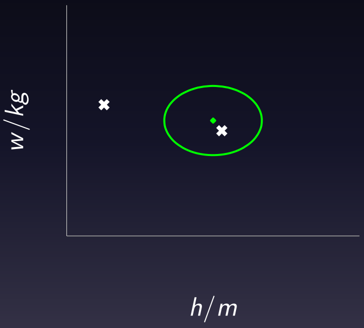
Marginal Distributions



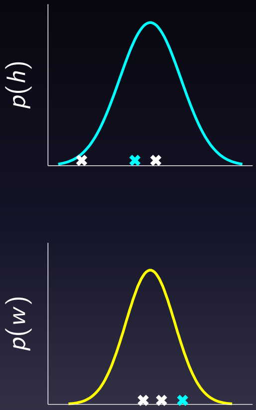
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables

Joint Distribution

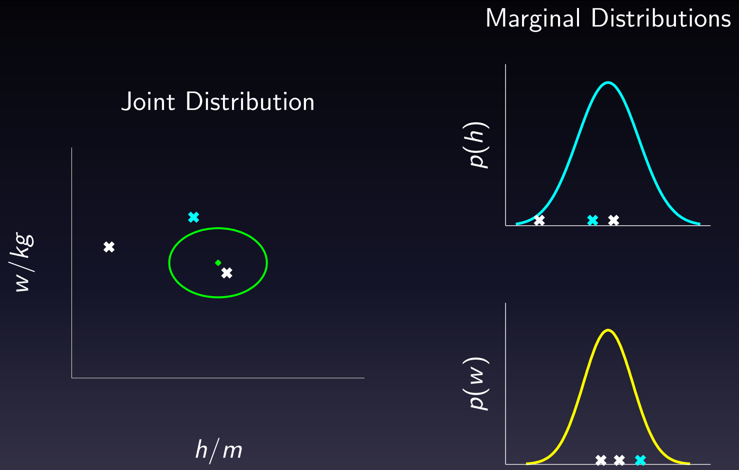


Marginal Distributions



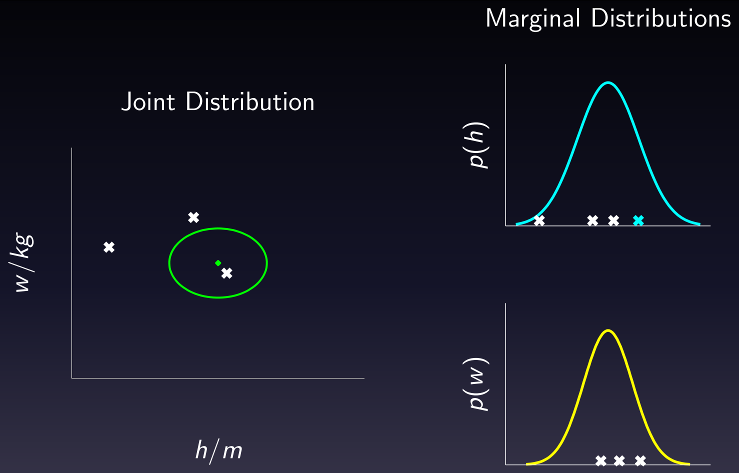
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



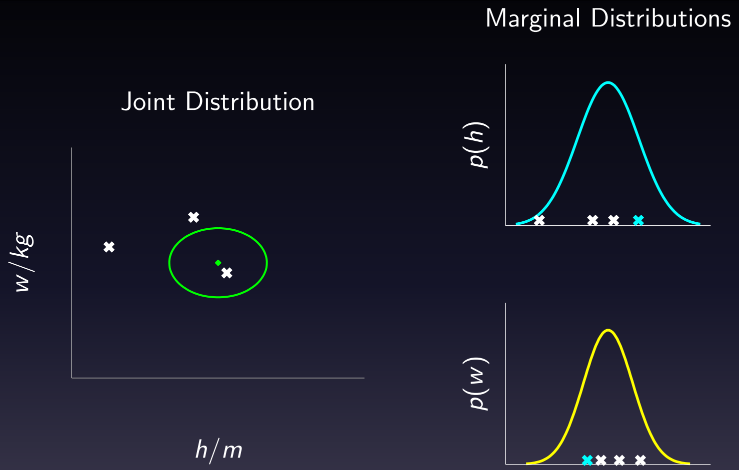
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



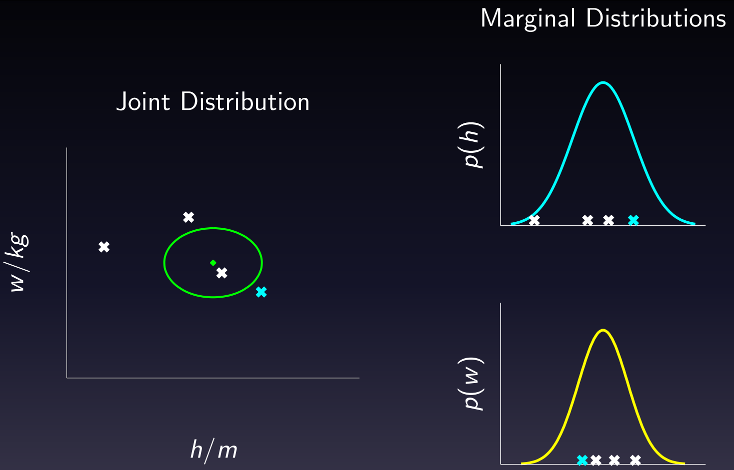
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



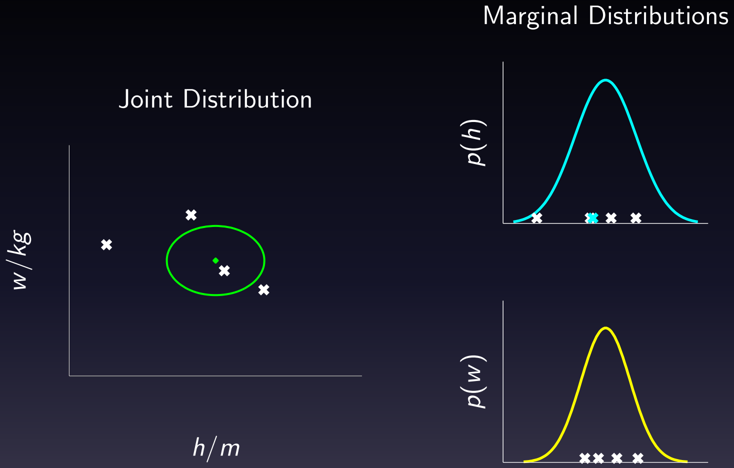
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



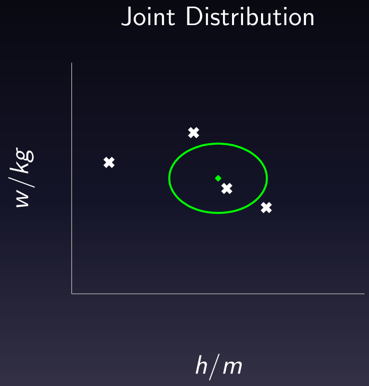
Sample height and weight one after the other and plot against each other.

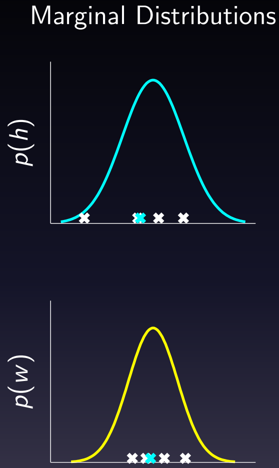
Sampling Two Dimensional Variables



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

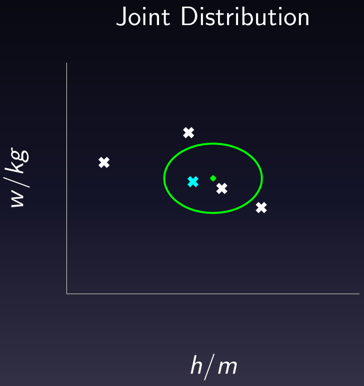
Sampling Two Dimensional Variables





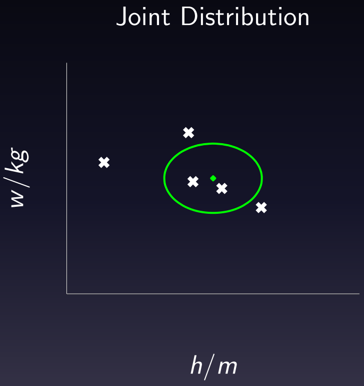
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



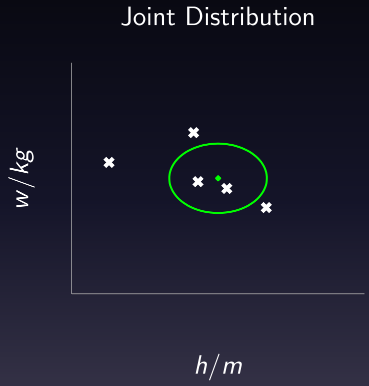
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



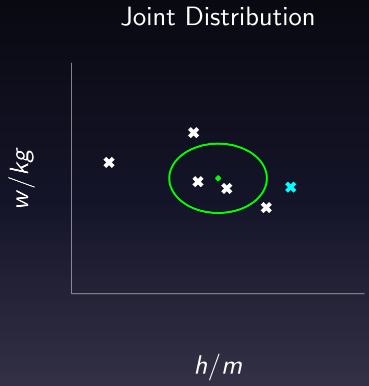
Sample height and weight one after the other and plot against each other.

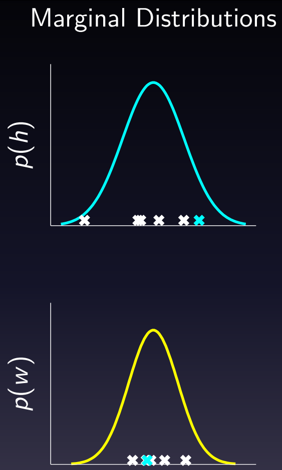
Sampling Two Dimensional Variables



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

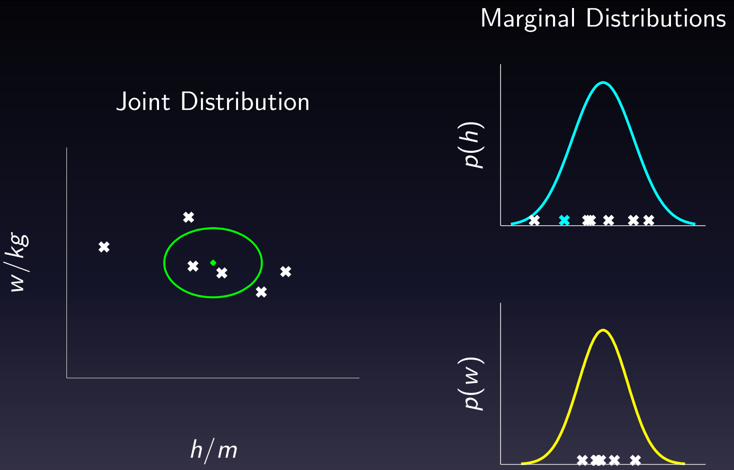
Sampling Two Dimensional Variables





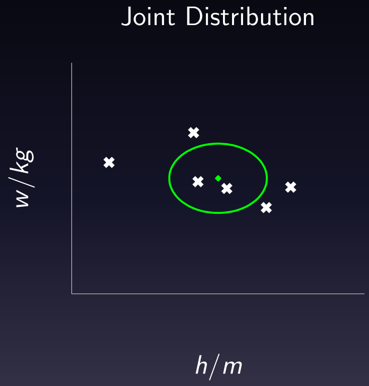
Sample height and weight one after the other and plot against each other.

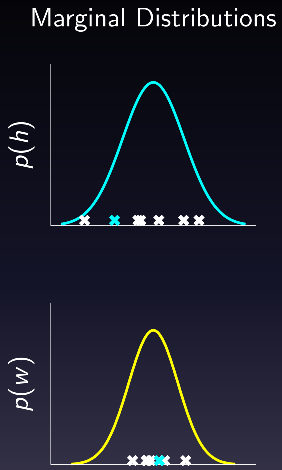
Sampling Two Dimensional Variables



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

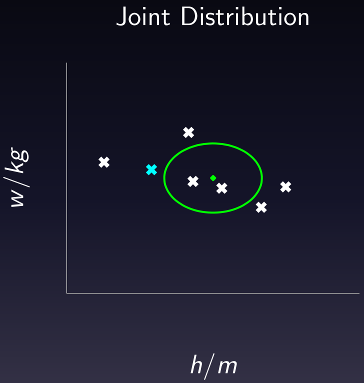
Sampling Two Dimensional Variables

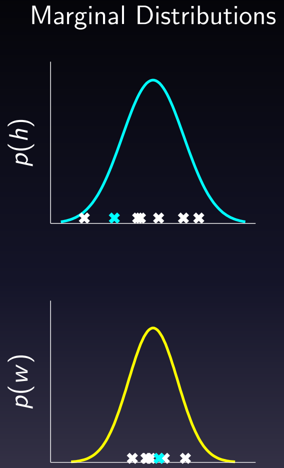




Sample height and weight one after the other and plot against each other.

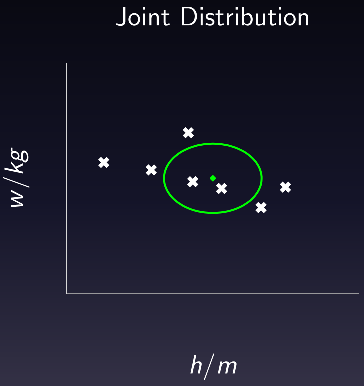
Sampling Two Dimensional Variables



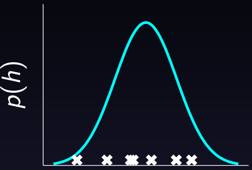


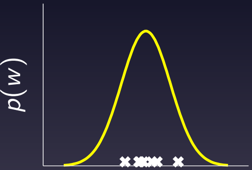
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



Marginal Distributions





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

Independence Assumption

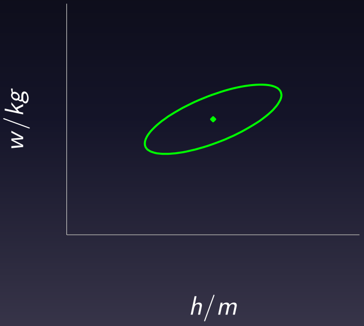
- This assumes height and weight are independent.

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

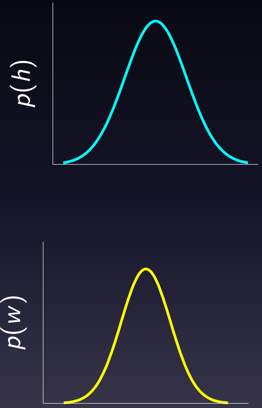
- In reality they are dependent (body mass index) $= \frac{w}{h^2}$.

Sampling Two Dimensional Variables

Joint Distribution

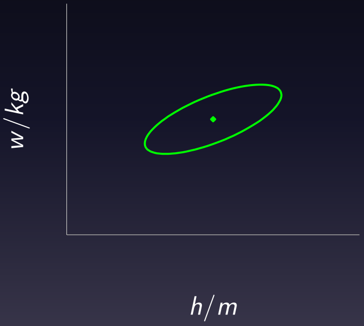


Marginal Distributions

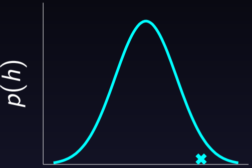


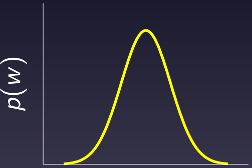
Sampling Two Dimensional Variables

Joint Distribution

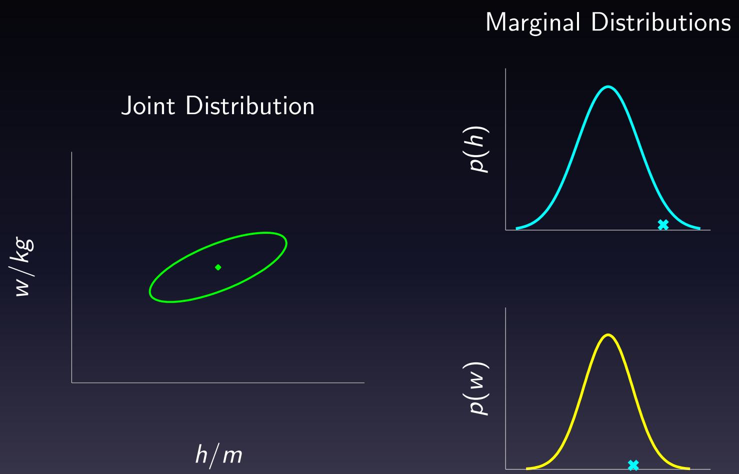


Marginal Distributions

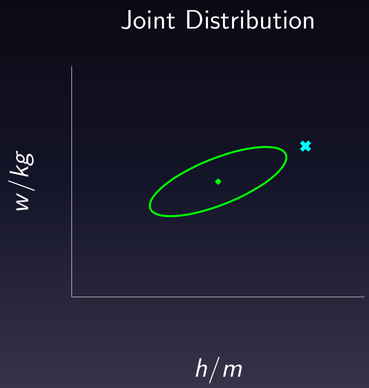




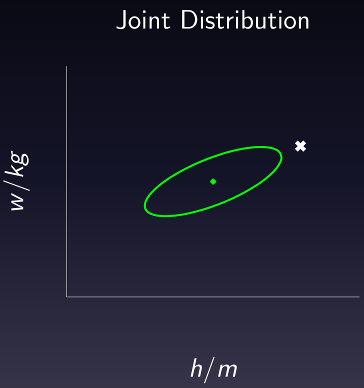
Sampling Two Dimensional Variables



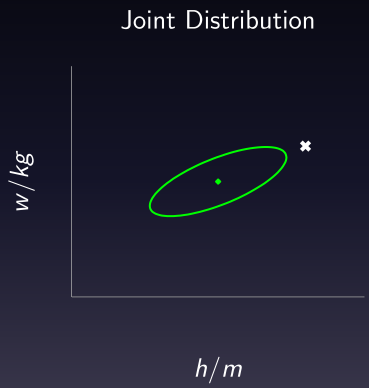
Sampling Two Dimensional Variables

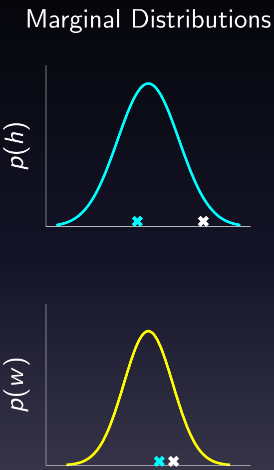


Sampling Two Dimensional Variables

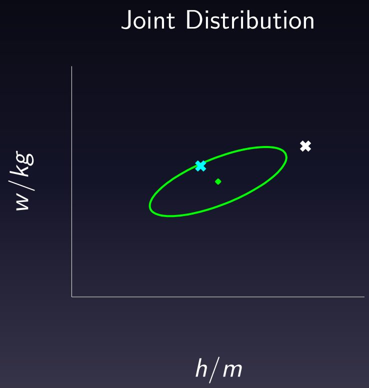


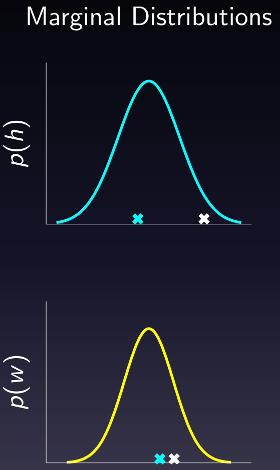
Sampling Two Dimensional Variables



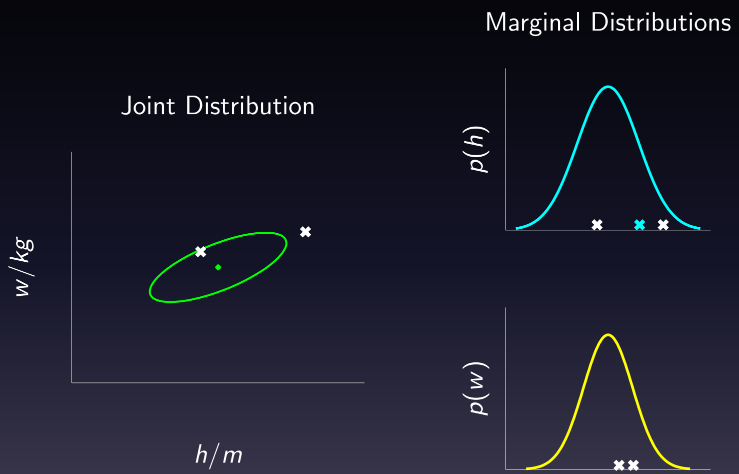


Sampling Two Dimensional Variables

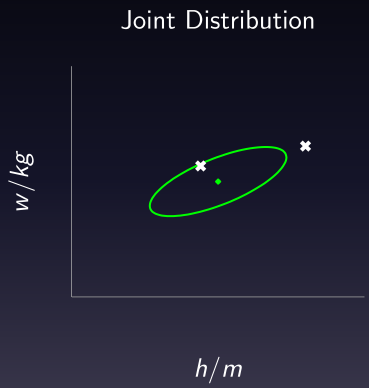




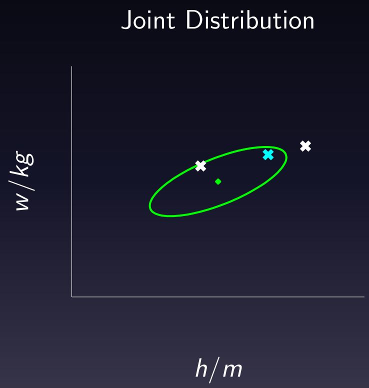
Sampling Two Dimensional Variables

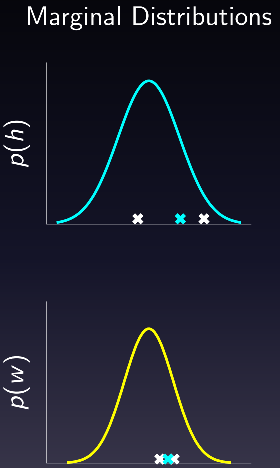


Sampling Two Dimensional Variables



Sampling Two Dimensional Variables

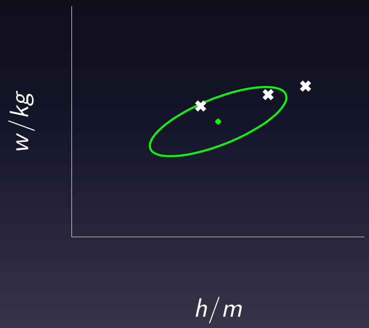




Sampling Two Dimensional Variables

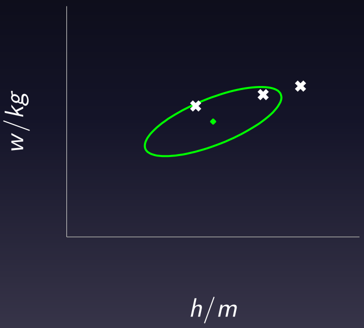
Marginal Distributions

Joint Distribution



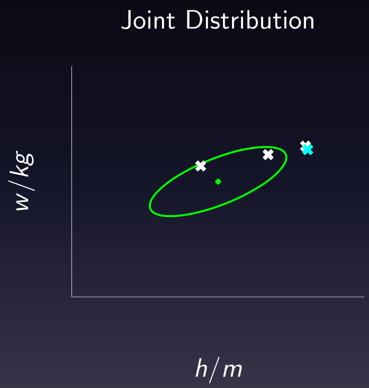
Sampling Two Dimensional Variables

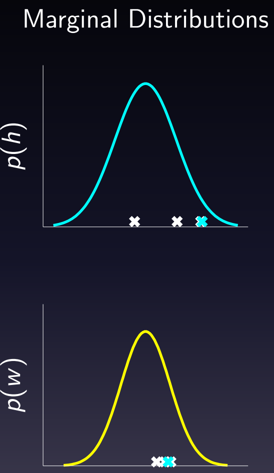
Joint Distribution



Marginal Distributions

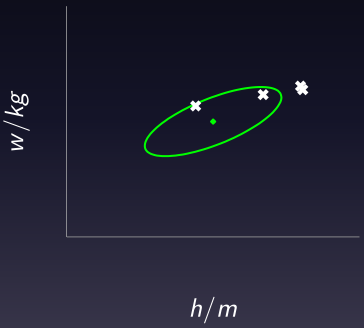
Sampling Two Dimensional Variables





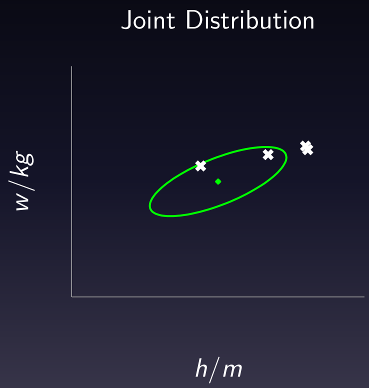
Sampling Two Dimensional Variables

Joint Distribution

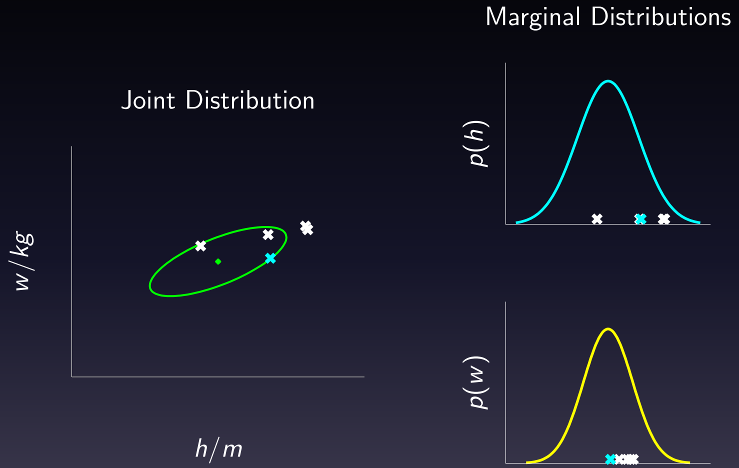


Marginal Distributions

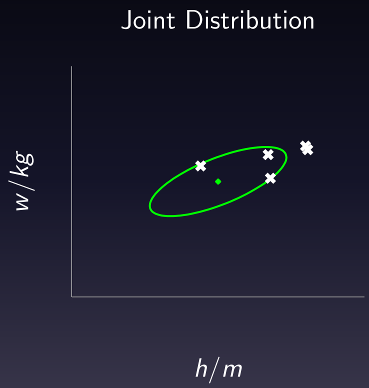
Sampling Two Dimensional Variables

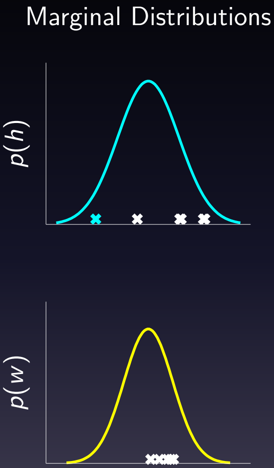


Sampling Two Dimensional Variables

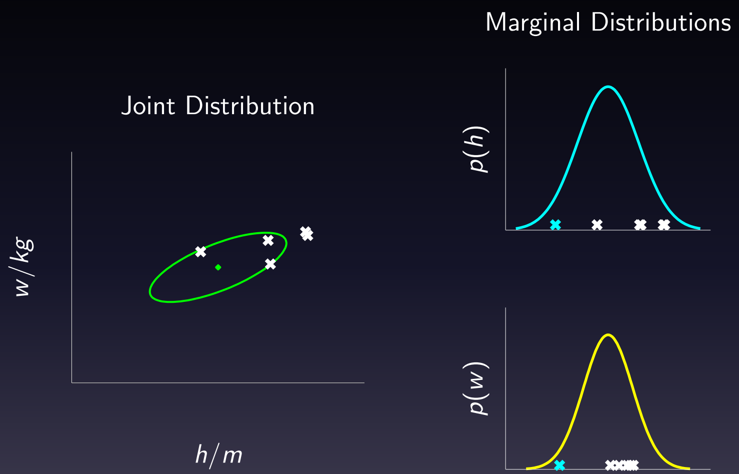


Sampling Two Dimensional Variables

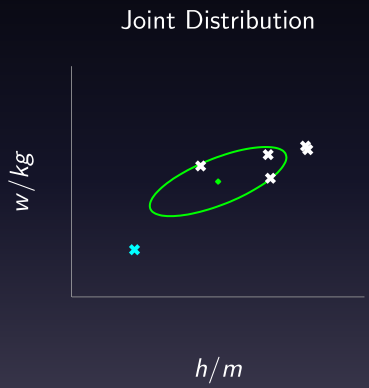




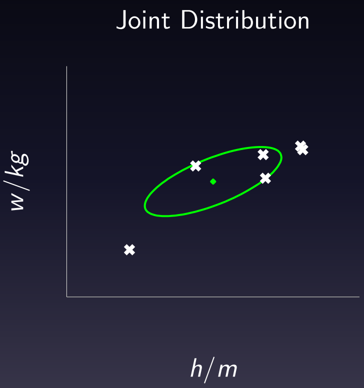
Sampling Two Dimensional Variables



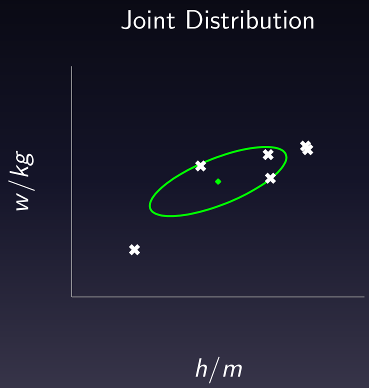
Sampling Two Dimensional Variables

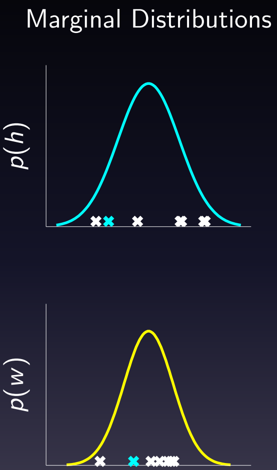


Sampling Two Dimensional Variables

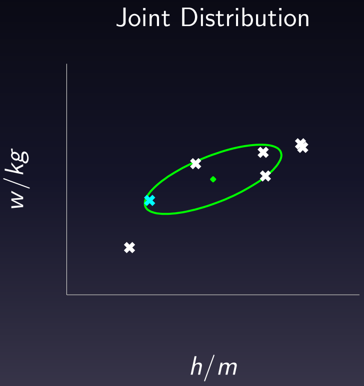


Sampling Two Dimensional Variables

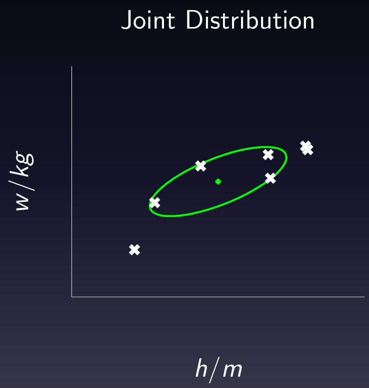




Sampling Two Dimensional Variables



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

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

Independent Gaussians

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

Correlated Gaussian

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

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

Correlated Gaussian

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

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

Correlated Gaussian

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

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

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 $\mathbf{R}$.

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

this gives a covariance matrix:

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

Outline

Univariate Bayesian Linear Regression

Multivariate Bayesian Linear Regression

Multivariate Regression Likelihood

- Recall multivariate regression likelihood:

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

- 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:

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

- Now use a multivariate Gaussian prior:

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

Posterior Density

- Once again we want to know the posterior:

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

- And we can compute by completing the square.

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

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

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

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

Posterior Density

- Once again we want to know the posterior:

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

- And we can compute by completing the square.

$$\begin{aligned}\log p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = &-\frac{1}{2\sigma^2}\sum_{i=1}^{N} t_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{N} t_i \mathbf{x}_{i,:}^\top \mathbf{w} \\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{N} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const.}\end{aligned}$$

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

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

Bayesian vs Maximum Likelihood

- Note the similarity between posterior mean

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

- and Maximum likelihood solution

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

Marginal Likelihood is Computed as Normalizer

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X})p(\mathbf{t}|\mathbf{X}) = p(\mathbf{t}|\mathbf{w}, \mathbf{X})p(\mathbf{w})$$

Marginal Likelihood

- Can compute the marginal likelihood as:

$$p(\mathbf{t}|\mathbf{X}, \alpha, \sigma) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \alpha\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)$$

Reading

- Section 2.3 of Bishop up to top of pg 85 (multivariate Gaussians).
- Section 3.3 of Bishop up to 159 (pg 152–159).

References I

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .