Statistics and Machine Learning 1

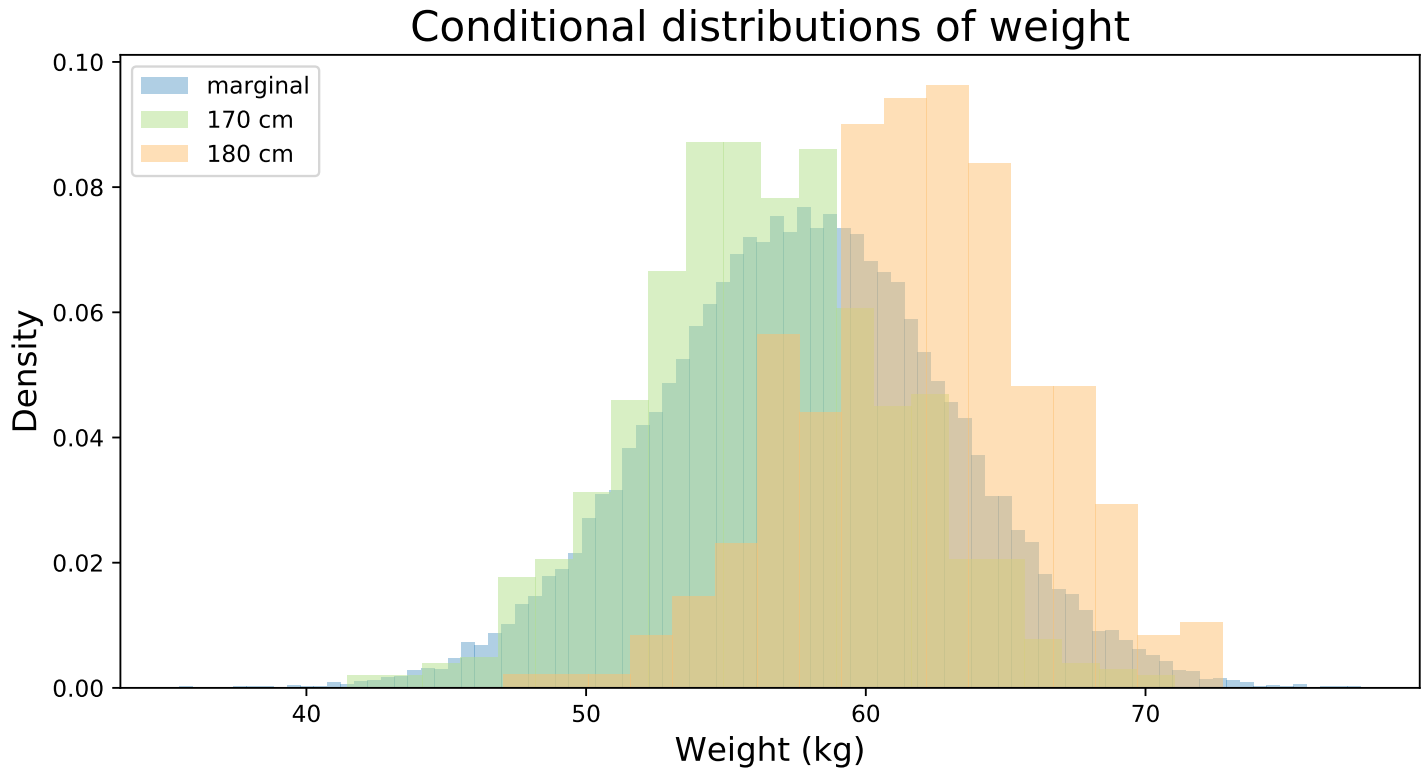
Lecture 7B: Modelling Joint and Conditional Densities Directly

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 7

Why should we care about conditional distributions?

Conditioning on height tells us something about weight.



Distrib.	Type	Mean (kg)	Std. dev. (kg)
$f(w)$	marginal	57.8	5.30
$f(w \mid h = 170)$	conditional	56.4	4.64
$f(w \mid h = 180)$	conditional	62.0	4.32

Regression models yield conditional distributions

Recall, from Areks's lectures, that the goal of a regression model is to

> $\cdots$ *characterise* $f(Y \mid X)$*, the conditional probability distribution of* Y *for different levels of* X*.*

A linear regression model for weight as a function of height would say

$$w_j = \beta_0 + \beta_1 h_j + \varepsilon_j$$

with $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$.

That is, it would predict that $f(w \mid h_\star)$ is a normal distribution with mean $E[w \mid h_\star] = \beta_0 + \beta_1 h_\star$ and variance $\sigma^2 = \mathrm{Var}(\varepsilon)$ that's independent of h.

Thinking more probabilistically I

An alternative to linear regression is to model the joint distribution $f(X, Y)$. That is, we think of the pairs

$$(h_1, w_1),\ (h_2, w_2), \ldots, (h_N, w_N)$$

as being drawn from some parametric family of two-dimensional probability distributions $f_\theta(h, w)$ and estimate the parameters.

Differences from the regression view:

- We have made a new aspect of the problem probabilistic: we now model the distribution of the h's as well as $f(w \mid h)$.
- Depending on the model we use for the joint distribution, the conditional distributions $f(w \mid h)$ may turn out to be easy to do computations with.
- The $f(w \mid h)$ may have variances that depend on h.

From points to densities with a multivariate normal

Given a collection $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of independent samples of a multivariate random variable $X \in \mathbb{R}^n$, we can approximate its density $f(\mathbf{x})$ with a *multivariate normal distribution (MVN)*

$$\hat{f}_{MVN}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\hat{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \hat{\mu})^\intercal \hat{\Sigma}^{-1}(\mathbf{x} - \hat{\mu})\right]$$

where

- $\hat{\mu}$ is the sample's *mean* and
- $\hat{\Sigma}$ is the sample's *covariance matrix*. Its determinant $|\hat{\Sigma}| = \det(\hat{\Sigma})$ appears in the normalisation constant and its inverse, $\hat{\Sigma}^{-1}$, appears in the argument of the exponential function.

A brief word about notation

We'll write vector quantities such as $\mathbf{x}_j \in \mathbb{R}^n$ as $n \times 1$ column vectors, so $\mathbf{x}_j$ and $\hat{\mu}$ on the previous slide would be

$$\mathbf{x}_j = \begin{bmatrix} x_{j1} \\ \vdots \\ x_{jn} \end{bmatrix} \qquad \text{and} \qquad \hat{\mu} = \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_n \end{bmatrix} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j .$$

The *transpose* $\mathbf{a}^\mathsf{T}$ of a column vector $\mathbf{a}$ is the $1 \times n$ row vector $\mathbf{a}^\mathsf{T} = [a_1, \ldots, a_n]$.

Similarly, the transpose of an $n \times m$ matrix $\mathbf{A}$ is an $m \times n$ matrix $\mathbf{A}^\mathsf{T}$ with entries $A^\mathsf{T}_{ij} = A_{ji}$. For example,

$$\text{if } \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \text{ then } \mathbf{A}^\mathsf{T} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

The covariance matrix: formula and code

The covariance matrix $\hat{\Sigma}$ that appears in the multivariate normal $\hat{f}_{MVN}$ has entries given by

$$\hat{\Sigma}_{rs} = \frac{1}{N-1} \sum_{j=1}^{n} (x_{jr} - \hat{\mu}_r)(x_{js} - \hat{\mu}_s).$$

Note that $\hat{\Sigma}$ is, by construction, symmetric: $\hat{\Sigma}^\intercal = \hat{\Sigma}$.

In practice, we'll compute neither $\hat{\mu}$ nor $\hat{\Sigma}$ from their definitions as the following snippet does the job:

```
# Estimate the parameters of a multivariate normal
hwCovMat = np.cov( hw_df['Height'], hw_df['Weight'] )
hwMeanVec =[np.mean(hw_df['Height']), np.mean(hw_df['Weight'])]
```

It yields

$$\hat{\mu}^\intercal \approx [172.70, 57.76] \text{ and } \hat{\Sigma} \approx \begin{bmatrix} 23.33 & 12.87 \\ 12.87 & 28.09 \end{bmatrix}.$$

Multivariate normals in python

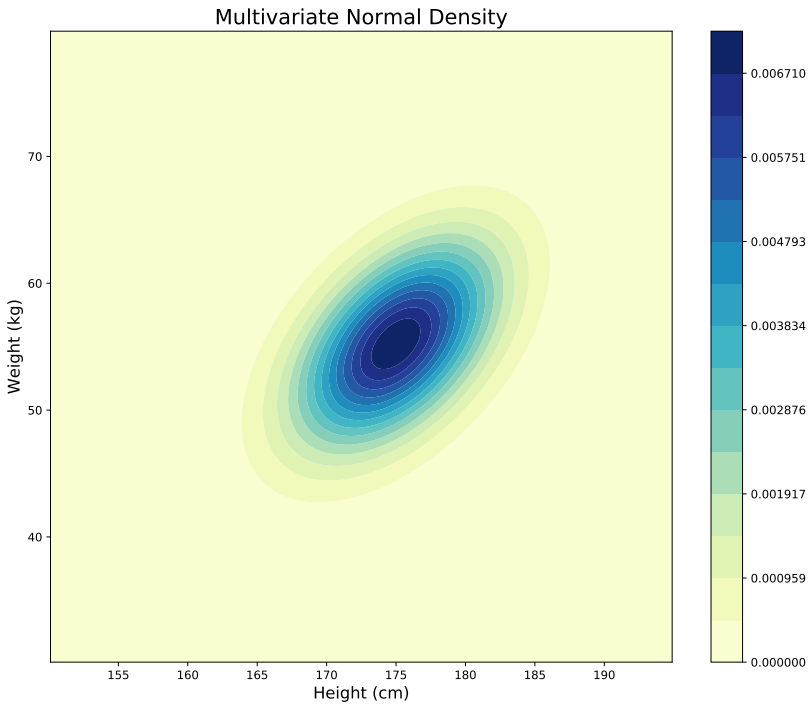
It's equally straightforward to evaluate the resulting density:

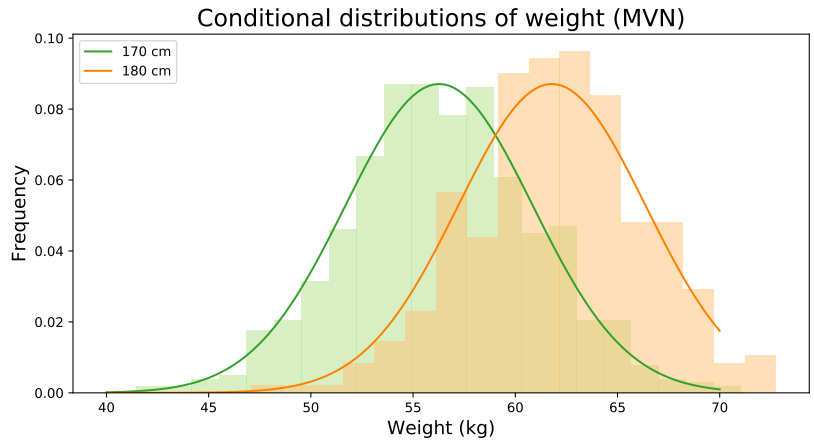
```
# Construct a multivariate normal object
from scipy.stats import multivariate_normal

hwCovMat = np.cov( hw_df['Height'], hw_df['Weight'] )
hwMeanVec =[np.mean(hw_df['Height']), np.mean(hw_df['Weight'])]
hw_mvn = multivariate_normal( mean=hwMeanVec, cov=hwCovMat )

# Evaluate it at a few points
hw_points = [[175, 58], [175, 59], [175, 60] ]
density_val = hw_mvn.pdf( hw_points )
```

produces

```
density_vals = array([0.00626202, 0.0064225 , 0.00628063]).
```

MVNs for the Hong Kong data





At left: a two-dimensional MVN for the Hong Kong data.

At right above: MVN and histogram estimates for the conditional distributions $f(w \mid h_\star)$ with $h_\star = 170$ (green) and $h_\star = 180$ (orange). These conditional distributions are also Gaussian and their means and variances can be computed exactly in terms of $h_\star$, $\hat{\mu}$ and $\hat{\Sigma}$.

Technical aside: conditionals and marginals of an MVN

The marginal distributions obtained by integrating-out some variables in a multivariate normal are . . .

- also normal (though perhaps still multivariate) and
- have means and covariance matrices that are simply related to those of the original distribution.

The same is true of the conditional distributions obtained by holding some subset of the MVN's components constant.

In what follows, we'll write

$$X \sim \mathcal{N}(\mu, \Sigma)$$

to mean that the random variable X has a multivariate normal distribution with mean μ and covariance matrix Σ.

MVN: mean and covariance of a marginal distribution

If $\mathbf{x}^\mathsf{T} = [x_1, \ldots, x_n]$ has an MVN distribution $\mathcal{N}(\mu, \Sigma)$ and we want the marginal distribution produced by integrating-out some subset of the components of $\mathbf{x}$, then the resulting marginal distribution is also MVN, with mean μ' and covariance matrix Σ' obtained by simply picking out the relevant entries of μ and Σ.

Example

Suppose $\mathbf{x}^\mathsf{T} = [x_1, x_2, x_3]$ has an multivariate normal distribution with mean and covariance

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}.$$

Then the marginal distribution $f(x_1, x_3)$ obtained by integrating-out x_2 is also a multivariate normal, with mean and covariance

$$\mu' = \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix} \quad \text{and} \quad \Sigma' = \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix}.$$

MVN: mean and covariance of a conditional distribution

Suppose $\mathbf{x}^\mathsf{T} = [x_1, \ldots, x_n]$ has an MVN distribution $\mathcal{N}(\mu, \Sigma)$ and that we have split its components into two disjoint groups, a and b, so $\mathbf{x}^\mathsf{T} = [\mathbf{x}_a, \mathbf{x}_b]$ and that

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}.$$

Note that that various parts of Σ are in general *matrices* and that two of them—Σ_{ab} and Σ_{ba}—need not even be square, though the symmetry of Σ means that $\Sigma_{ab}^\mathsf{T} = \Sigma_{ba}$.

If we condition on the variables in set a, then the conditional density $f(\mathbf{x}_b \mid \mathbf{x}_a = \mathbf{x}_a^\star)$ is $\mathcal{N}(\mu', \Sigma')$ with

$$\mu' = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\mathbf{x}_a^\star - \mu_a) \qquad \text{and} \qquad \Sigma' = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}.$$

Further reading

- If you have forgotten (or never knew) about matrices and their applications, the first chapter of

  S. Rogers and M. Girolami (2017), *A First Course in Machine Learning*, 2nd edition, Chapman & Hall/CRC. ISBN: 978-1-4987-3848-4.

  provides a rapid introduction to the basic ideas in the context of regression problems. It's available `online` through the University Library.
- An alternative, brief treatment of the same material is available via Section 7 of the HELM (Helping Engineers Learn Mathematics) workbooks hosted at
  `https://www.mub.eps.manchester.ac.uk/helm/`
- The formulae for the marginal and conditional distributions of a multivariate normal appear in Chapter 8 of Rogers and Girolami.