Statistics and Machine Learning 1

Lecture 3C: Multivariate Distributions

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

Week 3

Random vectors (for the mathematically minded!)

- For two length-p vectors $\mathbf{u}$ and $\mathbf{v}$, we write $\mathbf{u} \leq \mathbf{v}$ only if $u_i \leq v_i$ for each i from 1 to p.
- A random variable can be defined through its cumulative distribution function; we can do the same for a random vector $\mathbf{X}$

$$F(\mathbf{x}) = \mathrm{P}(\mathbf{X} \leq \mathbf{x}). \tag{1}$$

- We then define the probability density function for continuous variables through

$$F(\mathbf{b}) - F(\mathbf{a}) = \int_{x_1=a_1}^{b_1} \cdots \int_{x_p=a_p}^{b_p} f(\mathbf{x}) \, \mathrm{d}x_1 \cdots \mathrm{d}x_p. \tag{2}$$

Expectations and the Multivariate Normal

- Expectations are then

$$\mathbb{E}[g(\mathbf{X})] = \int_{x_1=a_1}^{b_1} \cdots \int_{x_p=a_p}^{b_p} g(\mathbf{x}) f(\mathbf{x}) \, \mathrm{d}x_1 \cdots \mathrm{d}x_p, \tag{3}$$

- An important example is the *multivariate normal*,

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{(2\pi)^{p/2}(\det(\mathbf{\Sigma}))^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \\ &\equiv \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}). \end{aligned} \tag{4}$$

- Note that the generalisation of variance is a covariance, with $\operatorname{cov}(\mathbf{X})$ being a matrix with (a, b)-th element $\mathbb{E}[X_a X_b] - \mathbb{E}[X_a]\mathbb{E}[X_b]$. For the multivariate normal, the covariance matrix is $\mathbf{\Sigma}$.

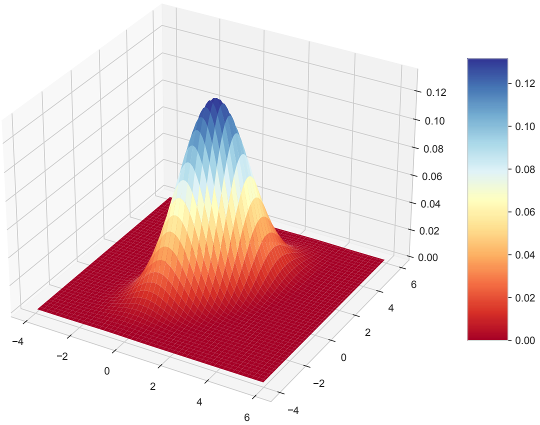
Multivariate normal contours

- The multivariate normal distribution defines a surface that has *ellipsoidal* contours centred on the mean.
- Consider a multivariate normal with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 2 \end{pmatrix}.$$

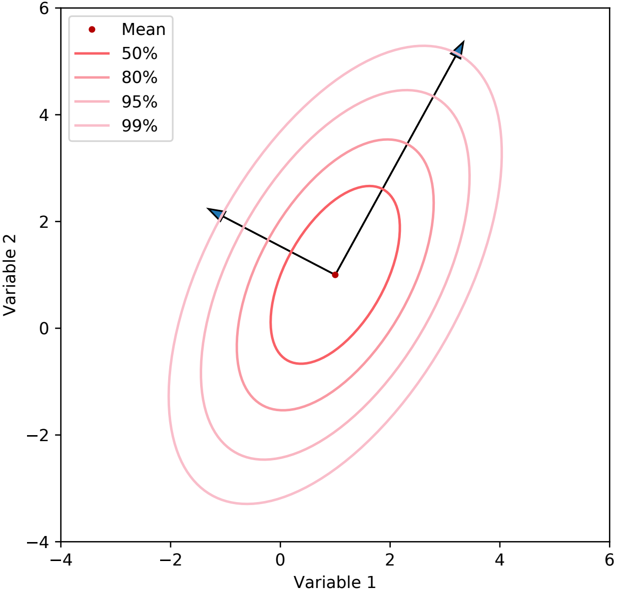
- defines a probability distribution over the plane.

Multivariate normal as a surface

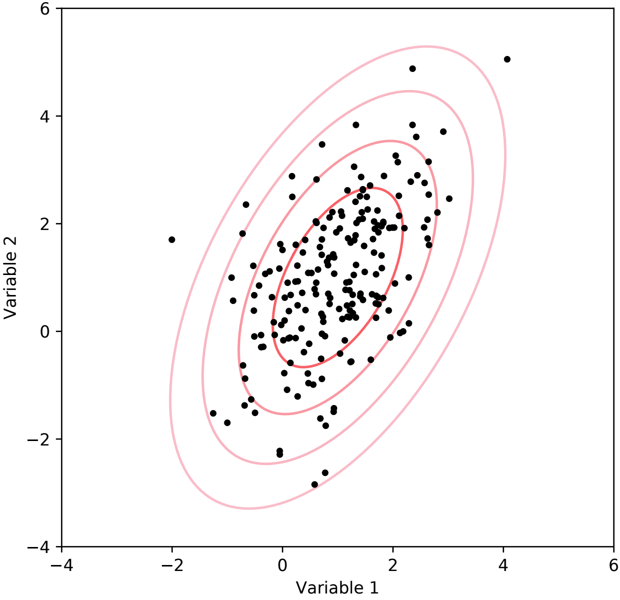


The probability of a region in the plane is given by the *volume* of the region beneath a surface, rather than the area beneath a curve.

Multivariate normal contours



Multivariate normal contours, plus samples



Centering

- When we centre data we subtract the sample mean from each row.
- One natural transform on data is to subtract the mean from each row.
- This can be achieved through use of a *centering matrix*,

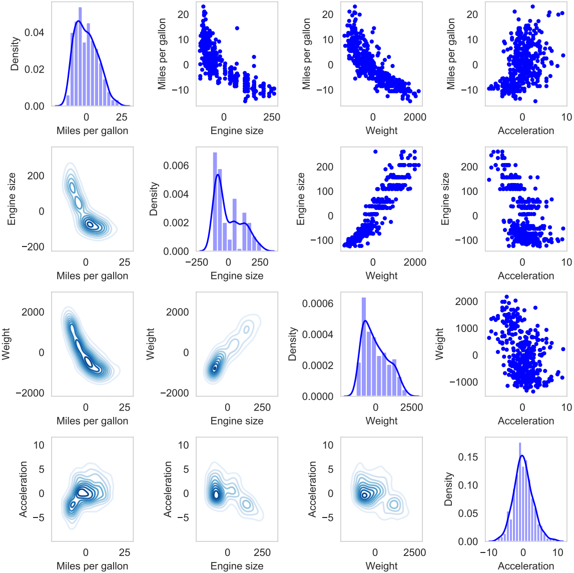
$$\boldsymbol{H} = \left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right). \tag{5}$$

- Then a transformed data matrix

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} \tag{6}$$

  will have mean $\langle \mathbf{y} \rangle = \mathbf{0}$.
- We visualise $\boldsymbol{Y}$ next.

Centred auto data



Standardisation

- Consider also the $p \times p$ matrix $\boldsymbol{D} = [D_{ab}]_{a=1,\ldots,p}^{b=1,\ldots,p}$ such that

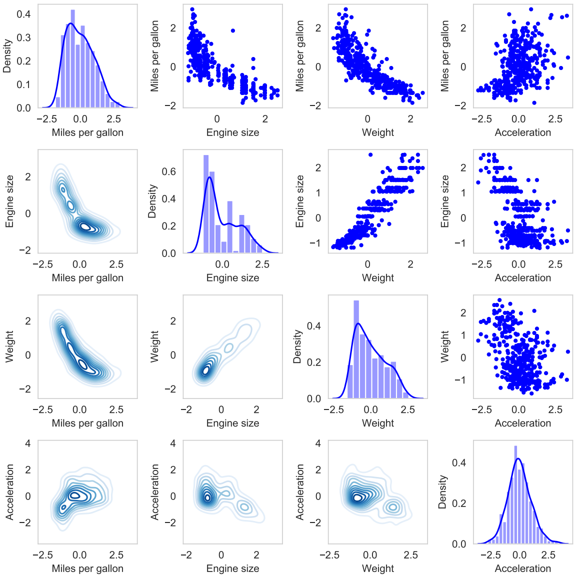
$$D_{ab} = \begin{cases} S_{aa}^{-1/2} & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

- Then a transformed data matrix

$$\boldsymbol{Z} = \boldsymbol{H}\boldsymbol{X}\boldsymbol{D} = \boldsymbol{Y}\boldsymbol{D} \tag{8}$$

  will have mean $\langle \mathbf{z} \rangle = \mathbf{0}$ and variance-covariance matrix equal to $\boldsymbol{X}$'s correlation matrix.
- We visualise $\boldsymbol{Z}$ next.

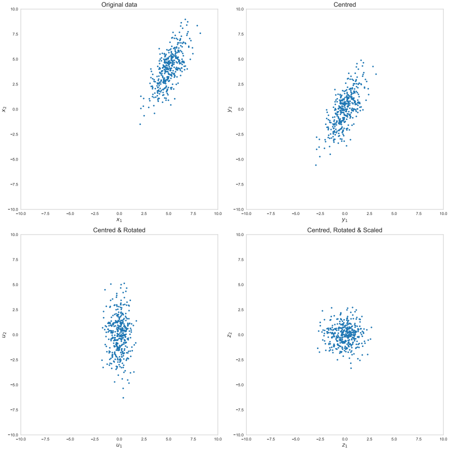
Standardised auto data



The Mahalanobis Transform for Data

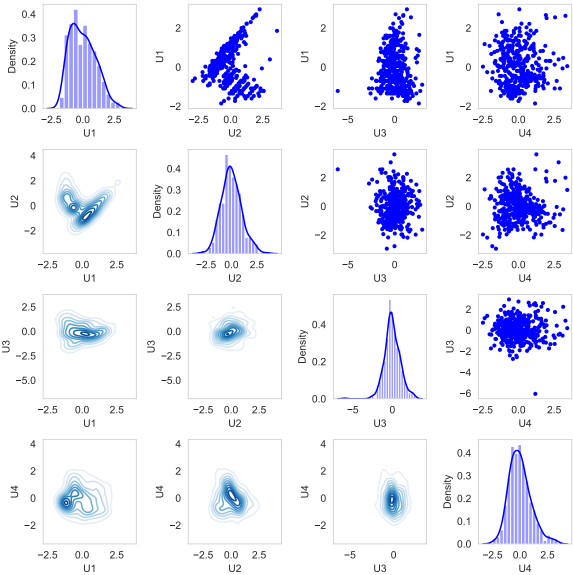
- The matrix version of the square root is called the *Cholesky decomposition*,
$$\boldsymbol{S} = \boldsymbol{C}^\top \boldsymbol{C}. \tag{9}$$
- We can use this to remove the correlations in data through the Mahalanobis transform:
$$\boldsymbol{U} = \boldsymbol{Y}(\boldsymbol{C}^{-1}). \tag{10}$$
- This yields a dataset with no correlations between variables, making comparison with a multivariate normal easier.
- We visualise such a $\boldsymbol{U}$ next, but note that the axes now show combinations of variables, rather than the original variables.

The Mahalanobis Transform: visually



This figure was inspired by a similar one in an `article` by Richard G. Brereton that relates the Mahalanobis Trasform to a topic we'll study early in the second semester: *Principal Components Analysis* (PCA).

Mahalanobis-Transformed Auto Data

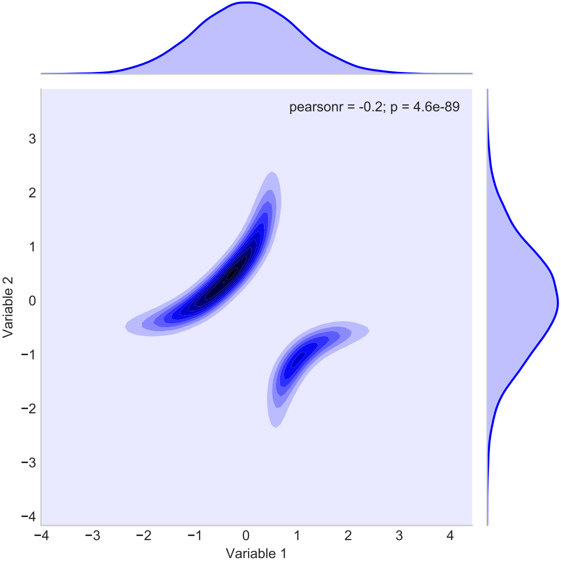


Spotting Deviations from Normality

- We saw that we can observe deviations from univariate normality with summary statistics such as skewness, kurtosis or through the detection pf multiple modes.
- These measures also generalise to the multivariate case.
- Here we show three cases where the multivariate distribution is far from multivariate normal, even if the marginal (i.e. univariate) datasets are very close to (or even indistinguishable from) univariate normal.
- We will show these using a visualisation technique known as a *joint plot*, where a bivariate visualisation is shown in the main panel and univariate visualisations are shown along the axes. These also show the correlation and associated p-value.

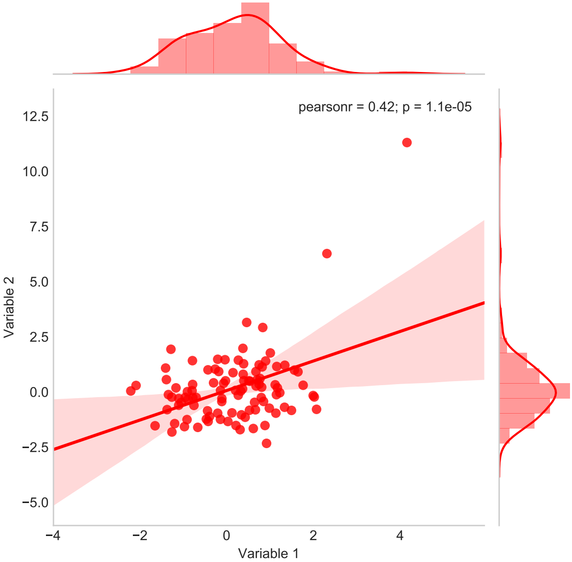
Multivariate multimodality

The data may not be centred around a unique mode. This is possible in complex ways in the multivariate case that can be far from obvious from lower dimensional analysis.



Outliers

Some datapoints may be very far from the mode. As well as strongly affecting estimates of means and variances, this can affect estimates of *correlations* between variables in the multivariate case.



Restricted Support

The data may only take positive values, or only values in some restricted region of space. For the multivariate case, such restrictions can be more complex than the univariate case.