01.001 - Introduction to Probability and Statistics

Alexander Binder

Week 11: First lecture on 03-Apr-2017, Second Lecture on 05-Apr-2017

Summary of key ideas from week's first lecture

Quick recap on Independence, and Conditional distributions

For n **variables being independent from each other**

$$f(x_1, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdot \ldots \cdot f_n(x_n)$$

Independence for three sets – $f_{X_{set2}}$ is the marginal density over all variables $X \in X_{set2}$:

$$f(x_1, \ldots, x_n) = f_{X_{set1}}(x_{set1}) f_{X_{set2}}(x_{set2}) f_{X_{set3}}(x_{set3})$$

Conditional distribution for variables in X_{Set1} - must sum up to 1 when integrated over X_{Set1}, and be the corresponding joint/marginal over all variables $f_{X_{Set1 \cup Set2}}$ times another term.

$$f_{X_{Set1}|X_{Set2}}(x_{Set1 \cup Set2}) = \frac{f_{X_{Set1 \cup Set2}}(x_{Set1 \cup Set2})}{f_{X_{Set2}}(x_{Set2})}$$

Independence for conditional distributions is:

$$f_{X_{Set1}|X_{Set2}}(x_{Set1} \mid x_{Set2}) = f_{X_{Set1}}(x_{Set1})$$
$$f_{X_{Set2}|X_{Set1}}(x_{Set2} \mid x_{Set1}) = f_{X_{Set2}}(x_{Set2})$$

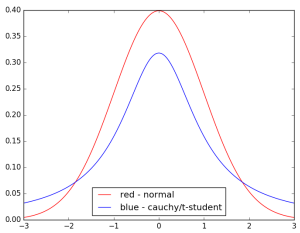
Quick Recap on modelling with conditional probabilities

First thing to note is that conditional probabilities and conditional densities can be naturally used to capture relationships between variables.

Example: Suppose one wants to model the ordering of some goods Y in an online store - for purposes of keeping the right amount of

stocks, as a function of month X_1 weekday X_2, and the a vector denoting distance to public holidays X_3 . There is no measuring noise in this problem, however clearly the problem is of a probabilistic nature. There exists no deterministic function $Y = g(X_1, X_2, X_3)$. In the occurrence of orderings there are many unaccounted effects (advertisement, new items coming out rendering this one obsolete) which are not modeled by above three variables. All those **unaccounted effects** induce random fluctuations, which can be modelled as a **noise random variable**.

Imagine $Y = g(X_1, X_2)$ is the power consumption of a computer. That might depend on the CPU load X_1, the GPU load X_2 but also many other variables. One can imagine to model this as $f(X_1, X_2) + Z$ where Z is a zero-mean gaussian random variable for the random fluctuations caused by all other effects in a running computer.



In this case $f(Y|X_1, X_2) = g(X_1, X_2) + N(0, \sigma^2)$

$g(X_1, X_2) + N(0, \sigma^2) \sim N(\mu = g(X_1, X_2), \sigma^2)$. Out of exams: In the simplest case, $f(X_1, X_2)$ could be a weighted sum of inputs, a linear/affine model:

$$Y = f(X_1, X_2) = w_1 X_1 + w_2 X_2 + b$$

Such models are used indeed for small-training data problems, see Gaussian process regression. (For Gaussian process regression usually a non-linear kernel is used.) This is one way to think about conditional distributions/ densities.

The second thought introduced was the averaging out of conditioning variables.

Suppose one has fitted a model with many causes

$$f(\text{awake after drink}|\ \text{caffeine}, \text{awake before drink}, age)$$

. One needs to use the model where in a setup where one conditioning variable X_1 cannot be measured (here: the level of awakeness before drinking). The idea is to average it out – by a den-

sity/distribution over the variable:

$$f(Y) = \int_{x_1} f(Y|X_1 = x_1) f_{X_1}(x_1) dx \qquad \text{density-modeled } X_1$$
$$f(Y) = \sum_{x_1} f(Y|X_1 = x_1) P_{X_1}(x_1) \qquad \text{discrete } X_1$$

If there are more conditioning variables, then the same equations hold when applying the conditioning variables X_2, X_3 to all terms:

$$f(Y|X_2, X_3) = \int_{x_1} f(Y|X_1 = x_1, X_2, X_3) f_{X_1|X_2,X_3}(x_1, x_2, x_3) dx$$
$$f(Y|X_2, X_3) = \sum_{x_1} f(Y|X_1 = x_1, X_2, X_3) P_{X_1|X_2,X_3}(x_1, x_2, x_3)$$

This can be applied to the above:

$$f(\text{awake after drink}|\ \text{caffeine}, age) =$$
$$\int_{x_1} f(\text{awake after drink}|\ \text{caffeine}, X_1 = \text{awake before drink} = x_1, age)$$
$$\cdot f_{X_1|\ \text{caffeine},age}(x_1,\ \text{caffeine}, age) dx_1$$

Another use case of conditional densities was to model hidden effects - like the factory task time T and the worker experience level Exp

$$f(T) = \sum_k f(T|Exp = k) P(Exp = k)$$

Here one can adapt the learned model that described the density of task time T conditional to the experience level Exp to different distributions of experience $P(Exp = k)$ in different factories.

The Multivariate Normal Density

The one-dimensional normal distribution with parameters μ, σ^2 generates real numbers $x \in \mathbb{R}^1$ according to the density

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

The multi-variate normal generates **vectors** of real numbers $x \in \mathbb{R}^n$ according to the density

$$f(x) = \frac{1}{(2\pi)^{n/2} det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

The density has as parameters

1. mean μ - a vector $\mu \in \mathbb{R}^{n\times 1}$
2. Covariance matrix Σ - a matrix $\Sigma \in \mathbb{R}^{n\times n}$. This matrix must satisfy two properties
   (a) Σ is symmetric
   (b) Σ is positive definite.

$det(\Sigma)$ is the Determinant of a matrix which is equal to the product of all its eigenvalues.

We make here the convention, that x is a column vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

We denote this case as $x \in \mathbb{R}^{n\times 1}$ – n rows and 1 column. A row vector, such as its transpose x^T would have shape $1 \times n$.

Recap on Vectors and Matrices

Let A be a $n \times n$ matrix. We denote by A_{ik} the entry in the matrix at row i and column k.

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix}$$

A is a **symmetric matrix** if $A_{ik} = A_{ki}$. In terms of the matrix transpose: $A^T = A$.

A is a **positive definite matrix**, if all its eigenvalues are positive.

Eigenvalues and Eigenvectors:
$\lambda \in \mathbb{R}^1$ is an **eigenvalue** of a $n \times n$-matrix A, if there exists a vector $v \neq 0, v \in \mathbb{R}^{n \times 1}$ such that

$$Av = \lambda v$$

The vector v to that is called **eigenvector**.

What is an eigenvector? One can imagine that a multiplying a vector with a matrix will rotate a vector and stretch or shrink it. An eigenvector is a direction in space, such that the vectors will be only stretched or shrunk, without any rotation.

Example: The eigenvectors of a diagonal matrix $\begin{pmatrix} s & 0 \\ 0 & t \end{pmatrix}$ are $v_1(a) = \begin{pmatrix} a \\ 0 \end{pmatrix}$ and $v_2(a) = \begin{pmatrix} 0 \\ a \end{pmatrix}$. So a diagonal matrix has eigenvectors which are along the coordinate axes.

Matrix-vector Multiplication:
one can multiply a matrix $A \in \mathbb{R}^{c \times n}$ from the right with a vector $x \in \mathbb{R}^{n \times 1}$. The result will be a vector: $(c, n) \cdot (n, 1) \to (c, 1)$. This product is defined as

$$(Ax)_i = \sum_{k=1}^{n} A_{ik} x_k$$

where $(Ax)_i$ is the i-th component of the vector ($i = 1, \ldots, c$). This formula says that the i-th entry of the vector Ax is the muliplication of the i-th row of matrix A with the vector x, as depicted in the graphics below

$$(Ax)_2 = A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + \ldots + A_{2n}x_n$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

$$(Ax)_3 = A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + \ldots + A_{3n}x_n$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (Ax)_1 \\ (Ax)_2 \\ (Ax)_3 \\ \vdots \\ (Ax)_n \end{pmatrix} = Ax$$

Similarly one can multiply a (row) vector $x \in \mathbb{R}^{1\times n}$ from the left with a matrix $A \in \mathbb{R}^{n\times c}$. The result vA will be a vector with format $\mathbb{R}^{1\times c}$.

Matrix Transpose:

The matrix transpose mirrors a matrix along its diagonal. Columns and Rows become swapped

$$(A^T)_{ik} = A_{ki}$$

$$\begin{pmatrix} a & B \\ C & d \end{pmatrix}^T = \begin{pmatrix} a & C \\ B & d \end{pmatrix}$$

Same one can apply for a vector. A transpose of a vector turns a $1 \times n$ vector into a $n \times 1$ vector and vice versa.

$$\begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix}^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

the identity Matrix I:

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$I_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{else} \end{cases}$$

The identity matrix has the following property under multiplication

$$AI = IA = A$$

Matrix Inverse:

A $n \times n$ matrix A can have an inverse matrix. The inverse does not exist for all matrices, though. A $n \times n$ matrix A^{-1} is an inverse matrix of matrix A if

$$A^{-1}A = AA^{-1} = I$$

where I is the identity matrix.

Euclidean norm of a vector:

Let $v = (v_1, v_2, \ldots, v_n)$ be a vector in n dimensions, then

$$\|v\| = \sqrt{\sum_{i=1}^{n} v_i^2}$$

is the euclidean length of a vector. For 2 dimensions this reduces to the well known vector length formula: $\|(v_1, v_2)\| = \sqrt{v_1^2 + v_2^2}$

Normal Distribution and its expectation

$$f(x) = \frac{1}{(2\pi)^{n/2} det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Property:

the expectation is equal to the mean vector μ

$$E[X] = \mu$$

Note that here

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_N \end{pmatrix}$$

is a vector. So the expectation is also a vector, defined for every component of the random variable vector X:

$$E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ E[X_3] \\ \vdots \\ E[X_N] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_N \end{pmatrix} = \mu$$

and its k-th component is

$$E[X_k] = \int x_k f(x) dx = \int x_k f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n$$

Normal Distribution and its Covariance

We will analyze the meaning of covariance next week. The covariance of two variables is defined as: $Cov(X_k, X_l) = E[(X_k - \mu_k)(X_l - \mu_l)]$. The normal distribution has also a covariance matrix Σ.

Property:

The Covariance of a normal distribution is related as follows to its covariance matrix parameter:

$$Cov(X_k, X_l) = E[(X_k - \mu_k)(X_l - \mu_l)] = \Sigma_{kl} = \Sigma_{lk}$$

One can write above relationship in matrix form:

$$\begin{pmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & Cov(X_2, X_3) & \cdots & Cov(X_2, X_n) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Cov(X_3, X_3) & \cdots & Cov(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & Cov(X_n, X_3) & \cdots & Cov(X_n, X_n) \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \cdots & \Sigma_{2n} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \cdots & \Sigma_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \Sigma_{n2} & \Sigma_{n3} & \cdots & \Sigma_{nn} \end{pmatrix} = \Sigma$$

Normal Distribution and its Marginals

This is nicely simple, too.

Suppose we model random variables $(X_1, X_2, \ldots, X_n)$ by a multivariate normal with parameters μ and Covariance matrix Σ. Let Σ be

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \Sigma_{n2} & \cdots & \Sigma_{nn} \end{pmatrix}$$

and μ be

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

Theorem:

Then the marginal distribution of any subset of $k < n$ variables from $(X_1, X_2, \ldots, X_n)$ is given by those rows of the mean vector μ, which is about the variables in the subset. Similarly the Covariance matrix of the marginal is given by those rows and columns of Σ which is about the variables in the subset. One needs to drop simply the irrelevant variables from μ and Σ.

This is best demonstrated by a few examples:

The marginal for $(X_2, \ldots, X_n)$ is given by parameters – column 1, row 1 gets deleted from Σ

$$\begin{pmatrix} \Sigma_{22} & \Sigma_{23} & \cdots & \Sigma_{2n} \\ \Sigma_{32} & \Sigma_{33} & \cdots & \Sigma_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n2} & \Sigma_{n3} & \cdots & \Sigma_{nn} \end{pmatrix}, \begin{pmatrix} \mu_2 \\ \mu_3 \\ \vdots \\ \mu_n \end{pmatrix}$$

The marginal for $(X_1, X_3, \ldots, X_n)$ is given by parameters – column 2, row 2 gets deleted from Σ

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{13} & \cdots & \Sigma_{1n} \\ \Sigma_{31} & \Sigma_{33} & \cdots & \Sigma_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \Sigma_{n3} & \cdots & \Sigma_{nn} \end{pmatrix}, \begin{pmatrix} \mu_1 \\ \mu_3 \\ \vdots \\ \mu_n \end{pmatrix}$$

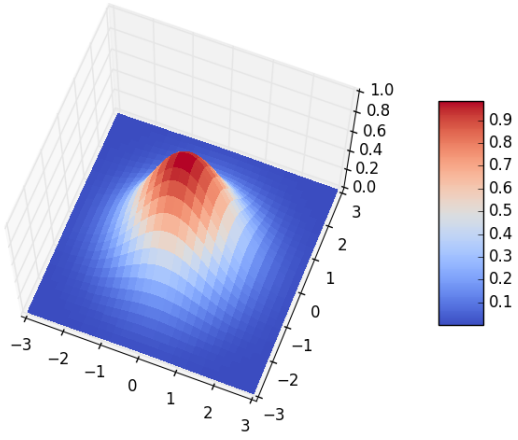
The marginal for $(X_1, X_3, X_5, X_6, \ldots, X_n)$ is given by parameters – columns $2, 4$, rows $2, 4$ gets deleted from Σ

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{13} & \Sigma_{15} & \Sigma_{16} & \cdots & \Sigma_{1n} \\ \Sigma_{31} & \Sigma_{33} & \Sigma_{35} & \Sigma_{36} & \cdots & \Sigma_{3n} \\ \Sigma_{51} & \Sigma_{53} & \Sigma_{55} & \Sigma_{56} & \cdots & \Sigma_{5n} \\ \Sigma_{61} & \Sigma_{63} & \Sigma_{65} & \Sigma_{66} & \cdots & \Sigma_{6n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \Sigma_{n3} & \Sigma_{n5} & \Sigma_{n6} & \cdots & \Sigma_{nn} \end{pmatrix}, \begin{pmatrix} \mu_1 \\ \mu_3 \\ \mu_5 \\ \mu_6 \\ \vdots \\ \mu_n \end{pmatrix}$$

The shape of multivariate normal density in 2 dimensions

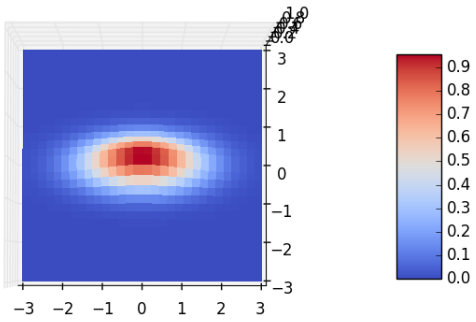
So far, the multivariate Normal has nice properties, its parameters μ and Σ are equal to the expectation and the covariance. Marginal distributions are nice. How does it look like?

Lets visualize how the density function looks in two dimensions for different choices of Σ, when $\mu = 0$ in two dimensions.



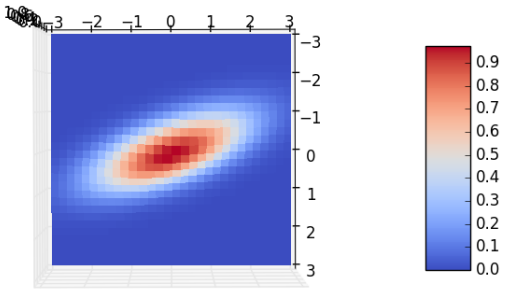
For the case: Covariance Matrix is a scaled Identity Matrix

$$\Sigma = \begin{pmatrix} s^2 & 0 & \cdots & 0 \\ 0 & s^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s^2 \end{pmatrix} = s^2 I$$



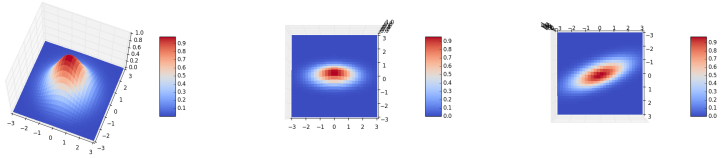
For the case: Covariance Matrix being a diagonal matrix

$$\Sigma = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^2 \end{pmatrix}$$



For the case: Covariance Matrix being general symmetric positive definite matrix $\Sigma = A^T \cdot A + \epsilon \cdot I$

plotting these three against each other shows the differences:



One can see:

1. if the covariance is a scaled identity, then the density looks like a surface made of circles.
2. if the covariance is a diagonal matrix, then the density looks like a surface made of ellipsoids. An ellipsoid looks like aa deformed circle, that got flattened in one dimension. The ellipsoids are parallel to the coordinate axes.
3. For the general case it looks like a surface made of ellipsoids which is rotated against the coordinate axes and not parallel to them. We will explain why the shapes are like that later on.

How is this related to the shape of Σ?

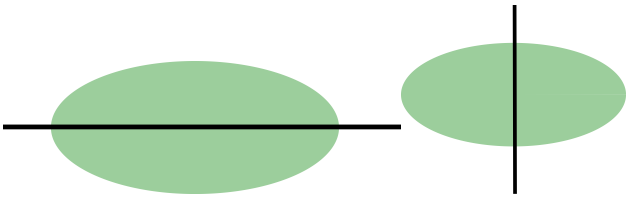
The shape of the ellipsoid and properties of Σ

The first question here is: how to define the circles or ellipsoids that we have seen in the plot above? One can define them by choosing a

positive constant $c > 0$ and consider the set of all those vectors x such that the density $f(x) = c$ is equal to that constant – we define them as contour lines for level c.

We will see here, that the shape of the ellipsoid is related to eigenvalues and eigenvectors of the covariance Σ.

Every ellipse in 2 dimensions has 2 symmetry axes – it can be mirrored along each of its 2 symmetry axes.



This extends directly to n dimensions, every ellipse in n dimensions has n symmetry axes. You can draw an ellipse in 3 dimensions in mplot3d of the matplotlib package to see the 3 symmetry axes.

Theorem -1:

The directions of the symmetry axes are the eigenvectors of the covariance Σ, that is, those vectors v for which a real number λ exists, such that $\Sigma v = \lambda v$.

An ellipse in n dimensions has n symmetry axes, and a symmetric $n \times n$ matrix has n pairs (v_i, λ_i) of eigenvector and eigenvalue – that fits.

Theorem 0:

The distance of from the center point of the ellipse to its boundary along the direction of the eigenvector v is proportional to the square root of the corresponding eigenvalue λ which solves $\Sigma v = \lambda v$.

Both results hold for n dimensions.

In class coding task:

See it yourself.

1. Take the file `plotcov_studentversion.py`
2. generate a random matrix Σ in the following way:

$$\Sigma = A^T \cdot A + 0.1 \cdot I,\ A \in \mathbb{R}^{2\times 2},\ A_{ij} \sim N(0,1)$$

The name of this random matrix Σ in the code should be `cov`. $N(0,1)$ is the one-dimensional standard normal distribution. `numpy.random` helps here. Note here: $A^T \cdot A$ is a symmetric matrix, but it can have zero eigenvectors, for example if A is the zero

matrix, then $A^T A$ is also the zero matrix. Adding $0.1 \cdot I$ ensures that it is positive definite, and also that both axes of the ellipse are not too thin for plotting nice visuals.

3. You will need to compute $\exp(-\frac{1}{2}x^T\Sigma^{-1}x)$, so you need the inverse of Σ. Invert it (scipy) and let the name of the inverse be `siginv`.

4. Compute the eigenvectors and eigenvalues of Σ. It is a symmetric matrix, so use `scipy.linalg.eigh`. These results should be stored it `eigenvecs,eigenvals`

5. validate that the solutions you got satisfy $Ax = \lambda x$. Validate that you really access the two pairs of eigenvector and eigenvalue correctly. You can do this by printing the vector norm of the difference between Ax and λx.

6. compute `z[index_1,index_2]` $= \exp(-\frac{1}{2}x^T\Sigma^{-1}x)$. The vector x is given in the double for-loop

7. run the script to see the eigenvector directions

8. modify the script: multiply each eigenvector with the corresponding eigenvalue. Rerun the script to see that eigenvector lengths now correspond to the ellipsoid axes lengths.

9. modify the script: try a simpler matrix type for `cov` - a diagonal matrix

$$\Sigma = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^2 \end{pmatrix}$$

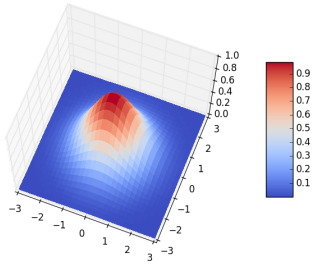
$s_i \sim N(0,1)$

The moral of it is: the shape of the normal density is an ellipsoid and it is defined by its eigenvectors and eigenvalues. The eigenvectors give the directions of the symmetry axes of the ellipse, and the eigenvalues tell how much the ellipse is stretched in this dimension.

We have seen the shapes and how they are related to eigenvectors and eigenvalues. Here we will deduce why in the case of $\Sigma = \lambda I$ the shape is like a circle and why in the case of Σ being a diagonal matrix, the shape is an ellipsoid.

Understanding the shape - case: Σ is a scaled Identity Matrix

The tool to analyze it will be to look at the set of vectors x such that $f(x) = c$ – the density f is constant for those vectors. These sets will be circles/ellipsoids.



For the case: Covariance Matrix is a scaled Identity Matrix

$$\Sigma = \begin{pmatrix} s^2 & 0 & \cdots & 0 \\ 0 & s^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s^2 \end{pmatrix} = s^2 I$$

Theorem 1: why the shape $x : f(x) = const$ is a circle:

The set of vectors x such that $f(x) = const$ under a gaussian density with covariance being a scaled identity $\Sigma = s^2 I$ is given by

$$\|x - \mu\|^2 = c^2$$

for some constant c.

Its interpretation is: the set of all vectors that have constant euclidean distance from the mean μ. This is a circle in 2 dimensions and a ball/sphere with radius c in n dimensions.

It can be seen easily for 2 dimensions and setting $\mu = 0$: The resulting equation

$$\|x - 0\|^2 = c^2 \Leftrightarrow x_1^2 + x_2^2 = c^2$$

is solved by

$$x_1(t) = c\cos(t), x_2(t) = c\sin(t)$$
$$x_1^2(t) + x_2^2(t) = c^2\left(\sin^2(t) + \cos^2(t)\right) = c^2 \cdot 1 = c^2$$

With an additional vector $\mu \neq 0$ the solution simply changes by adding μ_1, μ_2 to the components of the solution.

Theorem 2: the circle in 2 dimensions:

For 2 dimensions, the set of vectors x such that

$$\|x\|^2 = x_1^2 + x_2^2 = c^2$$

are given by the 1-dimensional curve $(x_1(t), x_2(t)) = (c\cos(t), c\sin(t))$ ... because $\cos^2(t) + \sin^2(t) = 1$, and

$$\|x - \mu\|^2 = (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 = c^2$$

are given by the 1-dimensional curve $(x_1(t), x_2(t))$, which describes a circle with radius c

$$(x_1(t), x_2(t)) = (\mu_1 + c\cos(t), \mu_2 + c\sin(t)),\ t \in [0, 2\pi)$$

Proof of Theorem 1:

The idea is to show that for the set of x such that the density is constant

$$\begin{aligned}
x:&\ \frac{1}{(2\pi)^{n/2}|det(\Sigma)|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right) = const_1 \\
\Leftrightarrow x:&\ \exp\left(-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right) = const_2 \\
\Leftrightarrow x:&\ (x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu) = const_3 \\
\Leftrightarrow x:&\ \|x-\mu\|^2 = const_4
\end{aligned}$$

The complication is to compute $(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)$ that is is equal to $\|x-\mu\|^2$

This matrix has a simple inverse: $\Sigma^{-1} = s^{-2}I$, because $s^{-2}I \cdot s^2 I = I \cdot I = I$.

Using this we arrive at:

$$\begin{aligned}
&(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu) \\
=&(x-\mu)^T \cdot (s^{-2}I) \cdot (x-\mu) \\
=&s^{-2}(x-\mu)^T \cdot (x-\mu) \\
=&s^{-2}\sum_{l=1}^{n}(x_l - \mu_l)(x_l - \mu_l) \\
=&s^{-2}\sum_{l=1}^{n}(x_l - \mu_l)^2 \\
=&s^{-2}\|x-\mu\|^2
\end{aligned}$$

$\|x-\mu\| = \sqrt{\sum_{l=1}^{n}(x_l - \mu_l)^2}$ is the euclidean norm of the difference vector $x - \mu$.

Theorem 3 – Eigenvectors and Eigenvalues:

In the case $\Sigma = s^2 I$ any vector $v \neq 0$ is an eigenvector, and all eigenvalues are equal to s^2.

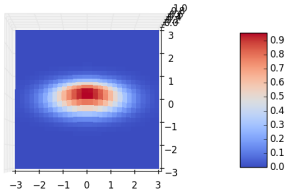
This is clear because $(s^2 I) \cdot v = s^2 v$.

In class coding task:
Use matplotlib or whatever you like to plot for various values of c

$$(x_1(t), x_2(t)) = (\mu_1 + c\cos(t), \mu_2 + c\sin(t)),\ t \in [0, 2\pi)$$

Understanding the shape - case: Σ is a Diagonal Matrix

We have seen from the coding task, that the shape looks like an ellipse, which is aligned to the coordinate axes. Here we explain why is that so. The tool to analyze it will be again to look at the set of vectors x such that $f(x) = c$ – the density f is constant for those vectors. This set will be the ellipsoid.



For the case: Covariance Matrix being a diagonal matrix

$$\Sigma = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^2 \end{pmatrix}$$

Theorem 4: why the shape $x : f(x) = const$ is an ellipsoid:

For

$$\Sigma = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^2 \end{pmatrix}$$

the set of vectors x such that $f(x) = const$ is given by

$$\sum_{l=1}^{n} s_l^{-2}(x_l - \mu_l)^2 = const$$

This is an ellipsoid in n dimensions.

Before we prove it, why are the x such that $\sum_{l=1}^{n} s_l^{-2}(x_l - \mu_l)^2 = const$ an ellipsoid?

It is very similar to the equation for the circle, but with a weight s_l^{-2} for every dimension. The equation for 2 dimensions

$$s_1^{-2}(x_1 - \mu_1)^2 + s_2^{-2}(x_2 - \mu_2)^2 = c^2$$

is very similar to the equation for a circle

$$(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 = c^2$$

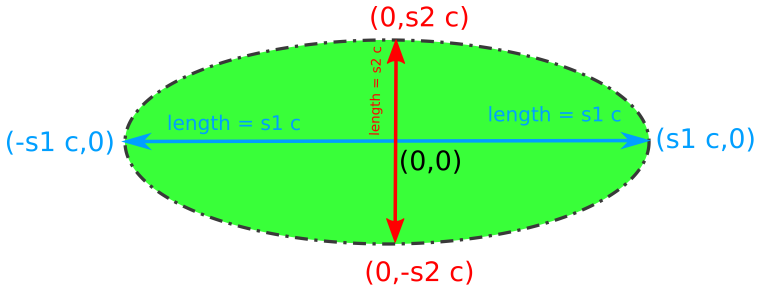
except that every dimension $i = 1, 2$ gets stretched/shrunk by a factor s_i^{-2}.

Insight: From a circle to an ellipsoid

If a pair (x, y) solves the circle equation $x^2 + y^2 = c^2$ around the mean $\mu = 0$ being the zero vector $(0, 0)$, then ... the pair $(\hat{x}, \hat{y}) = (s_1 x, s_2 y)$ solves the equation

$$\frac{\hat{x}^2}{s_1^2} + \frac{\hat{y}^2}{s_2^2} = s_1^{-2}(s_1 x)^2 + s_2^{-2}(s_2 y)^2 = c^2$$

That means: if (x, y) lie on a circle with radius around $(0, 0)$, then $(s_1 x, s_2 y)$ solve the ellipsoid equation – but $(s_1 x, s_2 y)$ is a circle that is deformed by a factor s_1 in the first dimension, and a factor s_2 in the second dimension – and this is exactly the shape of an ellipsoid.



(side note out of class: if (x, y) solves for $(x - \mu_1)^2 + (y - \mu_2)^2 = c^2$ – the equation with a mean $\mu \neq 0$, then $(s_1 x, s_2 y)$ is NOT enough. Instead then one needs $(\mu_1 + s_1(x - \mu_1), \mu_2 + s_2(y - \mu_2)) = (s_1 x + \mu_1(1 - s_1), s_2 y + \mu_2(1 - s_2))$)

Theorem 5: The set of x such in 2 dimensions that

$$s_1^{-2}(x_1 - \mu_1)^2 + s_2^{-2}(x_2 - \mu_2)^2 = c^2$$

is satisfied by the ellipse with half-axes length $s_1 c$ and $s_2 c$

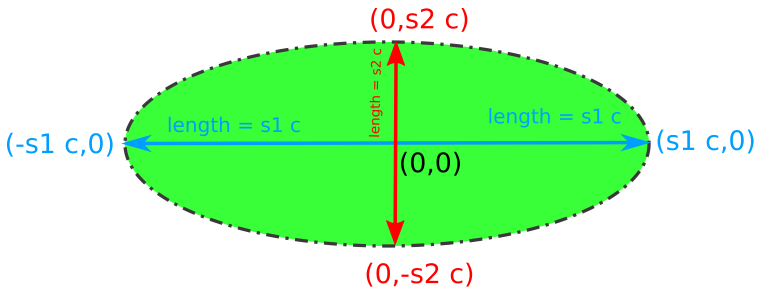
$$(x_1(t), x_2(t)) = (\mu_1 + s_1 c \cos(t), \mu_2 + s_2 c \sin(t))$$

This curve

$$(x_1(t), x_2(t)) = (0 + s_1 c\cos(t), 0 + s_2 c\sin(t))$$

goes for $(\mu_1, \mu_2) = (0,0)$ through the points

$$\text{at } t = \pi: \ (-s_1\, c, 0), \ \text{at } t = 0: \ (s_1\, c, 0)$$
$$\text{at } t = 1.5\pi: \ (0, -s_2\, c), \ \text{at } t = 0.5\pi: \ (0, s_2\, c)$$



We know that $(\mu_1 + 1 \cdot c\cos(t), \mu_2 + 1 \cdot c\sin(t))$ would be a circle with radius c around the center (μ_1, μ_2).

Above curve for $\mu \neq 0$ is a *deformed* circle around the center (μ_1, μ_2). It is deformed because it stretches between $[-s_1\, c, s_1\, c]$ along the x-axis and between $[-s_2\, c, s_2\, c]$ along the y-axis. And this is precisely an ellipse in 2 dimensions, as you can see from the graphic.

Regarding theorem 5, we can show by plugging it in

$$\begin{aligned}
&s_1^{-2}(x_1(t) - \mu_1)^2 + s_2^{-2}(x_2(t) - \mu_2)^2 \\
=&s_1^{-2}(s_1 c\cos(t))^2 + s_2^{-2}(s_2 c\sin(t))^2 \\
=&s_1^{-2} s_1^2 c^2\cos^2(t) + s_2^{-2} s_2^2 c^2\sin^2(t) \\
=&c^2(\cos^2(t) + \sin^2(t)) = c^2 \cdot 1
\end{aligned}$$

For n dimensions the argument of stretching out in each dimension k between $[-s_k\, c, s_k\, c]$ still holds. Consider for dimension k the vector

$$x = \mu + r \cdot e_k$$

, where e_k is the vector that is zero except in dimension k, and solve for the value of r

when plugging in this vector into the equation $f(x) = const$:

$$e_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \text{dimension } k \rightarrow \ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$x = \mu + r \cdot e_k \rightarrow$$

$$\sum_{l=1}^{n} s_l^{-2}(x_l - \mu_l)^2 = c^2 \Rightarrow s_k^{-2}(r \cdot 1)^2 + \sum_{l \neq k} 0 = c^2$$

$s_k^{-2} r^2 = c^2$ solves for $r = -s_k\, c$ and $r = +s_k\, c$. This shows that the solution stretches out between $[-s_k\, c, s_k\, c]$ along dimension / axis k. This is an ellipse in n dimensions[1]

[1] One can derive a formulation similar to theorem 5 for n dimensions using polar coordinates for n dimensions, and multiplying each dimension with $s_l c$.

Proof of Theorem 4:

The first part is same as for theorem 1 - without changes: the set of x such that the density is constant

$$\begin{aligned} x: \ & \frac{1}{(2\pi)^{n/2}|det(\Sigma)|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right) = const_1 \\ \Leftrightarrow x: \ & \exp\left(-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right) = const_2 \\ \Leftrightarrow x: \ & (x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu) = const_3 \end{aligned}$$

The complication is to compute $(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)$ and show that it is equal to what is claimed above.

The good news is that Σ^{-1} is a diagonal matrix, and $(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)$ is easy to compute for diagonal matrices.

First of all the inverse matrix is the diagonal matrix with diagonal entries inversed.

$$\Sigma^{-1} = \begin{pmatrix} s_1^{-2} & 0 & \cdots & 0 \\ 0 & s_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^{-2} \end{pmatrix}$$

We can compute $(x-\mu)^T \Sigma^{-1} (x-\mu)$ by using the matrix multipli-

cation formula

$$A \in \mathbb{R}^{n\times n}, v \in \mathbb{R}^{n\times 1} \Rightarrow Av \text{ is a vector } \in \mathbb{R}^{n\times 1} \text{ and}$$
$$(Av)_l = \sum_{r=1}^{n} A_{lr} v_r$$
$$\Rightarrow v^T Av = \sum_l \sum_r v_l A_{lr} v_r$$

Note here: Σ^{-1}_{lr} is zero when $l \neq r$ (because $l \neq r$ are the terms outside of the diagonal), so plugging in $v = x - \mu$, and $A = \Sigma^{-1}$ yields:

$$\begin{aligned}
&(x-\mu)^T\Sigma^{-1}(x-\mu) = \\
=&\sum_l \sum_r v_l \Sigma^{-1}_{lr} v_r \\
=&\sum_l v_l \Sigma^{-1}_{ll} v_l + \sum_l \sum_{r\neq l} 0 \\
=&\sum_l (x_l-\mu_l) s_l^{-2} (x_l-\mu_l) \\
=&\sum_{l=1}^{n} s_l^{-2}(x_l-\mu_l)^2
\end{aligned}$$

So $(x-\mu)^T\Sigma^{-1}(x-\mu) = const$ translates into $\sum_{l=1}^{n} s_l^{-2}(x_l-\mu_l)^2 = const$

In class coding task:

Use matplotlib or whatever you like to plot for $c = 1$ and various values of $s_1 > 0, s_2 > 0$

$$(x_1(t), x_2(t)) = (0 + s_1\, c\cos(t), 0 + s_2\, c\sin(t)),\ t \in [0, 2\pi)$$

Theorem 6 – Eigenvectors and Eigenvalues: :

For

$$\Sigma = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^2 \end{pmatrix}$$

the eigenvectors are given by the unit vectors e_k (see below theorem 5 for its definition), and the corresponding eigenvalues are s_k^2.

This theorem explains why for this case the visual showed an ellipse aligned to the coordinate system axis.

The proof is simple: $\Sigma \cdot e_k$ returns by the rules of matrix to vector multiplication (see above) the k-th column of Σ, which is $s_k^2 e_k$.

No exam/extra reading: Understanding the shape - case: Σ is a general symmetric positive definite matrix.

To understand why we have a *rotated* ellipse here, we can use a not so trivial result from linear algebra:

Theorem:

Every symmetric $n \times n$ matrix Σ can be decomposed as:

$$\Sigma = R^T D R$$

where D is a diagonal matrix, and R is a rotation matrix, that is its inverse is equal to its transpose: $R^T R = R R^T = I$.

We can show that the i-th eigenvector of Σ is a rotated unit vector $R^T e_i$, and the eigenvalue is the i-th diagonal entry $d_{ii} \in \mathbb{R}^1$ of the diagonal matrix D:

$$\Sigma \cdot (R^T e_i) = R^T D R \cdot R^T e_i = R^T D e_i = R^T d_{ii} e_i = d_{ii}(R^T e_i)$$

This rotation R explains why the eigenvectors in the code example were rotated against the coordinate axes.

Σ is – in a rotated coordinate system – a diagonal matrix. Therefore we have an ellipse which is rotated. The rotation of axes is done by R^T.[2]

[2] R is a rotation matrix because:1. it preserves the length of any vector v: $\|Rv\|^2 = \|v\|^2$. 2. it maps the zero vector onto itself (which a translation would not).

Worked-out examples related to week's first lecture

We had in class coding instead ... but

1.

$$f(x,y) = c\frac{y}{\sqrt{1+y^2}}x^5\sin(x^6)$$

Domain of integration is: $[0,2] \times [3,4]$

(a) What is f_X, f_Y

(b) What is $f_{Y|X}$?

Solution:

$$f_X(x) = \int_y f(x,y)dy = \int_y f(x,y)c\frac{y}{\sqrt{1+y^2}}x^5\sin(x^6)dy$$
$$= \sqrt{1+y^2}\Big|_3^4 \cdot x^5\sin(x^6)dx = (\sqrt{17}-\sqrt{10})cx^5\sin(x^6)$$

$$f_Y(x) = \int_x f(x,y)dx = \int_y f(x,y)c\frac{y}{\sqrt{1+y^2}}x^5\sin(x^6)dx$$
$$= \frac{y}{\sqrt{1+y^2}}\cdot(-\frac{1}{6})\cdot cos(x^6)\big|_0^2 dx = \frac{y}{\sqrt{1+y^2}}\cdot(-\frac{1}{6})\cdot(\cos(64)-1)$$

$$f_{Y|X}(x,y) = \frac{f(x,y)}{f_X(x)} = \frac{c\frac{y}{\sqrt{1+y^2}}x^5\sin(x^6)}{(\sqrt{17}-\sqrt{10})cx^5\sin(x^6)}$$
$$\frac{y}{\sqrt{1+y^2}}\frac{1}{\sqrt{17}-\sqrt{10}}$$

This does not depend on x anymore. The random variables are independent!