Introductory Notes on Probability and Statistical Inference

Lea Duncker

September 2017

This set of notes is intended to give a quick introduction to basic probability theory and statistical inference. The goal is to convey key concepts and intuition, without an overly technical treatment.

1 Basic Concepts and Definitions

In this section we will briefly introduce some key concepts you will come across all the time. If you feel unfamiliar with these you should do some extra reading. The introductory material in Bishop's *Pattern Recognition and Machine Learning* [1] or MacKay's *Information Theory, Inference, and Learning Algorithms* [2], or Murphy's *Machine Learning: A probabilistic perspective* [3] are good starting points. These books are also useful resources more generally.

1.1 Probabilities and Random Variables

Probabilites and random variables are the most fundamental concept in probability theory. Intuitively, a random variable is a variable that takes on some values, but we are unsure about what those values might be. Thus, we assign a probability to each of the possible values the random variable could take. We might express this as

$$X \sim p(X)$$

which means the random variable X is distributed according to $p(X)$, where $p(X)$ is a probability distribution. Depending on whether X takes values in the discrete or continuous domain $p(X)$ is called a probability mass function (pmf) probability, or density function (pdf), respectively.

For example, we could consider the random variable $H \in \{0, 1\}$ denoting whether the outcome of a coin toss is heads.

If the coin is fair, we would expect that

$$\lim_{N \to \infty} \frac{\#\text{Heads}}{N} = 0.5$$

as we repeat many coin tosses, where N is the total number of coin tosses. This would be a definition of probability in terms of frequencies: a probability is the relative occurrence of an event over a number of repeated experiments in the limit of infinite repetitions.

Thus, this gives us a pmf for the random variable H:

$H = h$	0	1
$p(H = h)$	0.5	0.5

This frequency-based definition makes sense when thinking about coin tosses, but what about more abstract events? We can't repeat things like a natural disaster infinitely many times. However, we might still want to attach a probability to the occurrence of such an event. Thus, another way to think of probabilities is as *beliefs*. You could view these beliefs as a state of knowledge, personal or objective, which you can use to reason logically and coherently about the world. This interpretation is known as the *Bayesian* view of probability, while the former is known as the *Frequentist* view.

In order for our beliefs to be consistent, we require them to satisfy certain properties. These are also known as the three axioms of probability:

1. Probabilities are non-negative and real: $P(X) \geq 0, \in \mathbb{R}^+$
2. Probabilities are normalised: $\int P(X)dX = 1$
3. Probabilities of disjoint (mutually exclusive) sets (e.g. alternatives) add: $P(A \cup B) = P(A) + P(B)$, if $P(A \cap B) = 0$.

1.2 Marginals, Joints and Conditionals

So far we have looked at a distribution over a single variable. This is also called the *marginal distribution*. However, we might also be interested in asking questions about multiple random variables. For example, I might have a belief about how tall a person is. How is that related to that person's weight? Knowing that someone is really tall also changes my belief about how much they weigh. Similarly, knowing someone's weight also gives me some idea about how tall they might be. We can formalise this intuiting using *joint* and *conditional* distributions.

A joint distribution is a distribution over two events co-occuring, i.e. the probablity that X *and* Y occur. We write this as $P(X, Y)$. The conditional distribution of X given Y is defined as

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

and allows us to make statements about our belief about one event given that we know the outcome of another one.

From this definition follows the *product rule* which rearranges the about to show that the joint distribution can be factorised into

$$P(X, Y) = P(X|Y)P(Y)$$

Another important concept is that of *marginalisation*. This means that if we sum/integrate over all possible outcomes of one variable in the joint distribution, we obtain the marginal distribution of the other variable:

$$P(X) = \int P(X, Y)dY$$

1.3 Independence

Independence is an important concept in probability theory. Formally, two random variables are independent if

$$P(X, Y) = P(X)P(Y)$$

From this and the product rule, we also know that this means

$$P(X|Y) = P(X)$$

So knowing Y doesn't tell me anything about X and vice versa – the two variables are independent. Independence means the joint distribution *factorises*[1].

1.4 Expected Value

The expected value is an important concept. Its formal definition for a random variable $X \sim P(X)$ is

$$\mathbb{E}_{p(X)}[X] = \int xp(x)dx$$

or

$$\mathbb{E}_{p(X)}[X] = \sum_{x_i} x_i p(x_i)$$

for continuous and discrete variables, respectively. This gives us the average value that X would take under the density $P(X)$. For discrete variables this is really intuitive: we just compute a weighted sum of all possible events, where the weight is determined by the probability of that event occurring. We can also take expected values of functions of X:

$$\mathbb{E}_{p(X)}[f(X)] = \int f(x)p(x)dx$$

For $f(X) = (X - \mathbb{E}_{p(X)}[X])^2$ this gives us the variance of X under $p(X)$. There are another few useful results relating to expected values, which are called the *law of total expectation* or *law of iterated expectation*.

$$\mathbb{E}_{p(X)}[f(X)] = \mathbb{E}_{p(Y)}[\mathbb{E}_{p(X|Y)}[f(X)]]$$

[1] Don't think about independence in terms of correlations: uncorrelated variables are not necessarily independent!

To see why this is the case, we can use our previous results about marginal,joint and conditional distributions:

$$\begin{aligned}\mathbb{E}_{p(X)}[f(X)] &= \int f(x)p(x)dx = \iint f(x)p(x,y)dxdy \\ &= \iint f(x)p(x|y)p(y)dxdy = \int \big(\int f(x)p(x|y)dx\big)p(y)dy \\ &= \mathbb{E}_{p(Y)}[\mathbb{E}_{p(X|Y)}[f(X)]]\end{aligned}$$

We can do something similar for computing the marginal variance, which is called the *law of total variance*.

1.5 Bayes Theorem

Bayes Theorem is

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Bayes theorem basically tells us that we can learn something about an underlying, unknown cause by recording data that was generated by that cause, or in other words, $X|Y$ tells us something about $Y|X$. Bayes Theorem is really fundamental to the field of Bayesian inference and you should know it by heart. We will get back to it later when talking about Bayesian Posterior Inference in section 2.2.

1.6 Example: HIV tests

Let us apply Bayes theorem to a simple problem. Suppose you are interested in HIV tests. You know a few things:

- 95% of people who have HIV test positive
- 98% of people who do not have HIV test negative
- the probability of HIV is one in thousand[2]

Suppose you know that someone tested positive. What is your belief about that person actually having HIV?[3]

We can use Bayesian reasoning to answer this, so let us start by translating the information above into numbers. Let $X = 1$ denote that someone has HIV and let $Y = 1$ denote a positive test outcome. We have been given the information

$$\begin{aligned}P(Y = 1|X = 1) &= 0.95 \\ P(Y = 0|X = 0) &= 0.98 \\ P(X = 1) &= 0.001\end{aligned}$$

[2]this is probably not very accurate but it's just an example

[3]This example is adapted from Jinghao Xue's UCL course on Applied Bayesian Methods

The quantity we are interested in is

$$P(X = 1|Y = 1) = \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1)}$$

We have everything we need to calculate the numerator, but we need to work a bit harder to get an expression for the denominator. Luckily, we know about marginalisation, so we can write

$$P(Y = 1) = P(Y = 1, X = 1) + P(Y = 1, X = 0)$$

Applying the product rule, we can write

$$P(Y = 1) = P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0)$$

To get expressions for the probabilities in the second term, we need to remember that probabilities sum to 1. Thus

$$P(X = 0) = 1 - P(X = 1) = 0.999$$
$$P(Y = 1|X = 0) = 1 - P(Y = 0|X = 0) = 0.02$$

Plugging everything back into the initial expression we get

$$\begin{aligned} P(X = 1|Y = 1) &= \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0)} \\ &= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 * 0.999} = 0.0454 \end{aligned}$$

2 Statistical Inference

The previous section has introduced a lot of terms and definitions, which we will now need to talk about Statistical Inference. Statistical Inference is really important and a key principle in science. In essence, it allows us to ask questions about what we can learn from the data we record.

Let us first introduce some notation which we will use throughout this section:

- θ is a parameter of interest.
- $y|\theta \sim P(y|\theta)$ is some observed value of a random variable (i.e. data) whose distribution depends on θ. This is called a *likelihood*.

2.1 Point estimation

Point estimation means that we want to find a particular estimate of θ, which we will denote as $\hat{\theta}$. We could think as our data $\{y_1, \ldots, y_N\}$ as being a finite sample from a population Y. The population might have a true underlying mean, which we don't know but would like to estimate using the finite sample we have recorded. For example, if $\theta = \mathbb{E}_{p(Y|\theta)}[Y]$, an estimate of the mean is the average:

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

How can you find such estimates? A good idea is to optimise an objective function, like, for example, the log-likelihood. Maximising the log-likelihood gives you the *maximum likelihood* (ML) estimate:

$$\hat{\theta}^{ML} = \arg\max_{\theta} \sum_{i=1}^{N} \log p(y_i|\theta)$$

There exists a lot of theory about what makes an estimator (i.e. the function of the data that gives you an estimate) a good one or a bad one. Is it biased? Does it have high variance? Is your estimator the best you can do? Can you find a better estimator from the one you already have? (Check out the bias-variance trade-off, the Cramer-Rao lower bound, and the Rao-Blackwell theorem if you are interested in this stuff.)

There is also a big literature on hypothesis testing, confidence intervals and p-values related to this. I won't touch on it here, but if these words sound unfamiliar read about it!

2.2 Bayesian Posterior Inference

Bayesian Inference and point estimation have similar goals: we observe some data and we want to learn about our unknown parameter θ. The key difference, however, is that Bayesian Inference gives us the ability to argue about *uncertainty*. I have seen some data, I have a point estimate $\hat{\theta}$, but how uncertain am I about this value? What is my *belief* about it now after having seen some data? It turns out we can formalise this quite easily using Bayes theorem:

$$P(\theta|y_1, \ldots, y_N) = \frac{P(y_1, \ldots, y_N|\theta)P(\theta)}{P(y_1, \ldots, y_N)}$$

$P(\theta|y_1, \ldots, y_N)$ is called the *posterior* distribution of θ given the data $_1, \ldots, y_N$. It encodes my belief about θ after having made observations. $P(\theta)$ is called the *prior* distribution of θ. It encodes my belief about θ before I've made these observations.

So, in short, Bayesian Inference boils down to computing posterior distributions (though a lot of the times this is easier said than done!). Once we have the posterior distribution we can compute point estimates from it. Common choices for this are the posterior mean, or the posterior mode (this is often called the *maximum a posteriori*, or MAP estimate).

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(\theta|y_1, \ldots, y_N)$$

You can also look at how much spread there is in the distribution around $\hat{\theta}^{MAP}$ which gives you a sense of the uncertainty about this estimate. You can also ask things like: What is the probability that θ is between 3 and 5 given the data I have seen.

However, the choice of prior distribution really matters for this, since our posterior distribution depends on it. If you look back at the HIV example in section 1.6, and you change $P(X = 1)$ from 0.001 to 0.5 you will see that the posterior probability of $P(X = 1|Y = 1)$ also changes. $P(X = 1)$, the marginal probability of HIV can be viewed as our prior belief that anyone might have HIV. For our reasoning to make sense and for us to be taken seriously we want our prior to be a good one. If you are interested in the problem of choosing good priors you can look up Jeffrey's prior and Empirical Bayes, but this exceeds the scope of these notes.

2.3 Example: The mean of a Gaussian distribution

Suppose we are interested in the mean height of people in the UK. We have some prior belief about average heights based on the people we have seen so far, which we can incorporate into a prior distribution. Let θ denote the mean height and y_i denote the height of one of N persons you got to measure. Let $Y = \{y_1, \ldots, y_N\}$ denote the collection of all these observed measurements. We will assume that both our prior belief about the mean height, and the variability in height can be formulated as Gaussian distribution. Let us also assume that we know the variance around the mean height to keep things simple for now.

$$\begin{aligned} \theta &\sim \mathcal{N}(\theta|\theta_0, \nu_0^2) \\ y_i|\theta &\sim \mathcal{N}(y|\theta, \sigma^2) \qquad\qquad \text{for } i = 1, \ldots, N \end{aligned}$$

We say that the data are *independent and identically distributed* (iid), which means that all y_i follow the same distribution and observing a particular value y_i does not affect the probability of observing a value $y_j, j \neq i$.

2.3.1 Maximum likelihood estimate

We can first find the ML estimate given our data. The log-likelihood is

$$\ell(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \theta)^2 - \frac{1}{2}\log(2\pi\sigma^2)$$

To find the maximum we differentiate the above with respect to θ and set to zero:

$$\begin{aligned} \frac{\partial}{\partial\theta}\ell(\theta) &= \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \theta) \stackrel{!}{=} 0 \\ 0 &= \sum_{i=1}^{N} y_i - N\hat{\theta} \\ \Rightarrow \hat{\theta} &= \frac{1}{N}\sum_{i=1}^{N} y_i \end{aligned}$$

So our ML estimate is just the sample mean.

2.3.2 Posterior Inference

In order to compute the posterior distribution we can recall Bayes Theorem:

$$\begin{aligned} p(\theta|Y) &= \frac{p(Y|\theta)p(\theta)}{p(Y)} \\ &\propto \prod_{i=1}^{N} p(y_i|\theta)p(\theta) \end{aligned}$$

Note that the likelihood factorises over data points because we have assumed that they are iid. The proportionality sign comes from the fact that $p(Y)$ is just a constant with respect to θ, so the functional form of the posterior distribution does not depend on it. This is an important concept that makes life a lot easier when computing posteriors: any constant terms are part of the *normaliser* – the constant that ensures that the proability density is normalised, i.e. integrates to 1. Thus, if we can recognise the functional form of the distribution and know that $\int p(\theta|Y)d\theta = 1$, we get the normaliser for free.

$$\begin{aligned} p(\theta|Y) &\propto \prod_{i=1}^{N} p(y_i|\theta)p(\theta) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\theta)^2\right)\exp\left(-\frac{1}{2\nu^2}(\theta-\theta_0)^2\right) \end{aligned}$$

We have again absorbed all constants with respect to μ into the proportionality sign. We can now expand out the squares and rearrange, again dropping any constants we don't care about because they don't depend on θ:

$$\begin{aligned} p(\theta|Y) &\propto \exp\left(-\frac{1}{2\sigma^2}\left(N\theta^2 - 2\theta\sum_{i=1}^{N}y_i\right) - \frac{1}{2\nu^2}\left(\theta^2 - 2\theta\theta_0\right)\right) \\ &\propto \exp(-\frac{1}{2}(\frac{N}{\sigma^2}\theta^2 - 2\frac{1}{\sigma^2}\theta\sum_{i=1}^{N}y_i) - \frac{1}{2}(\frac{1}{\nu^2}\theta^2 - 2\frac{1}{\nu^2}\theta\theta_0)) \\ &\propto \exp\left(-\frac{1}{2}\left(\left(\frac{N}{\sigma^2}+\frac{1}{\nu^2}\right)\theta^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}y_i + \frac{1}{\nu^2}\theta_0\right)\theta\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\nu^2}\right)\left(\theta^2 - 2\frac{\frac{1}{\sigma^2}\sum_{i=1}^{N}y_i + \frac{1}{\nu^2}\theta_0}{\frac{N}{\sigma^2}+\frac{1}{\nu^2}}\theta\right)\right) \end{aligned}$$

We almost have something that looks like a Gaussian in θ, but we still need to complete the square:

$$\begin{aligned} p(\theta|Y) &\propto \exp\left(-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\nu^2}\right)\left(\theta^2 - 2\frac{\frac{1}{\sigma^2}\sum_{i=1}^{N}y_i + \frac{1}{\nu^2}\theta_0}{\frac{N}{\sigma^2}+\frac{1}{\nu^2}}\theta\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\nu^2}\right)\left(\theta - \frac{\frac{1}{\sigma^2}\sum_{i=1}^{N}y_i + \frac{1}{\nu^2}\theta_0}{\frac{N}{\sigma^2}+\frac{1}{\nu^2}}\right)^2\right) \end{aligned}$$

We can recognise the functional form to be Gaussian – all that is missing is the normalising constant which we have neglected to begin with. Thus, we have

$$p(\theta|Y) = \mathcal{N}\left(\theta\middle|\frac{\frac{1}{\sigma^2}\sum_{i=1}^{N}y_i + \frac{1}{\nu^2}\theta_0}{\frac{N}{\sigma^2}+\frac{1}{\nu^2}}, \frac{1}{\frac{N}{\sigma^2}+\frac{1}{\nu^2}}\right)$$

The posterior mean is a weighted combination of the prior mean and the sample mean of the data, where the weights scale with the inverse variance. This is again intuitive: we want to combine our knowledge from the prior with the data we have observed, but we also want to take into account how uncertain we are about this knowledge.

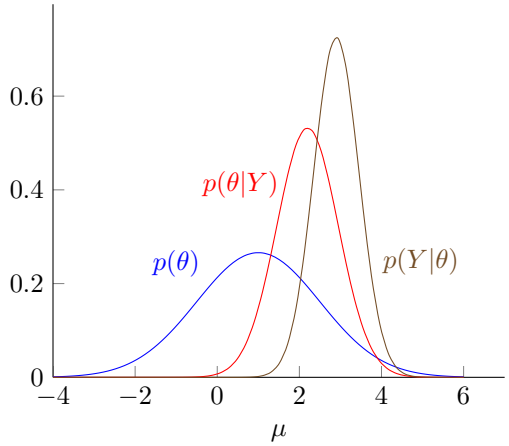

Figure 1: Illustration of the prior and likelihood, which are combined to compute the posterior distribution of the mean parameter θ, given the observed data Y.

For example, if I have not seen many people in the UK I might expect they are as tall as me. So my prior could be centred around 1.70m. But since I haven't seen that many people and really don't know how much taller or shorter other people are, I would be pretty uncertain about this statement and my prior variance, ν^2 would be large. If I now observe many people who are all between 1.55m and 1.75m, the variance in my likelihood (expressed as a function of θ) would not be that large. My posterior would then combine both the likelihood and the prior to give the posterior, which can be viewed as an updated belief. If I observe more height measurements in the future, I can use it as a starting point (prior) and sequentially update my distribution over θ. Figure 1 illustrates the difference between the shape of the prior, likelihood and posterior distribution.

3 Multivariate Distributions

A lot of standard distributions are univariate, e.g. Binomial, Poisson, Gaussian, Gamma, etc. (if you don't know any of these, do look them up!) but they can usually be generalised to account for vector- or matrix-valued random variables.

We have already seen an example of a multivariate distribution when we looked at the joint distribution of two random variables, X and Y. Let $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ then $P(X, Y) = P(Z)$ – a distribution over a vector valued random variable. In this section, we will look at the multivariate Gaussian distribution.

3.1 Bivariate Normal Distribution

We can construct a bivariate distribution of two independent Gaussian random variables

$$x \sim \mathcal{N}(\mu_x, \sigma_x^2)$$
$$y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

Since x and y are independent we can write

$$\begin{aligned} p(x, y) &= p(x)p(y) \\ &= \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2\sigma_x^2}(x-\mu_x)^2\right) \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{1}{2\sigma_y^2}(y-\mu_y)^2\right) \end{aligned}$$

we can re-write the above in matrix form:

$$\begin{aligned} p(x, y) &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^\mathsf{T} \begin{bmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}\right) \\ &= \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_z)^\mathsf{T}\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)\right) = p(\boldsymbol{z}) \end{aligned}$$

Convince yourself that multiplying out the quadratic form in the exponent yields the previous expression. The final expression above is the general form of a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

In our example $\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$ is diagonal and x and y were independent. More generally, Σ is of the form

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

where ρ is a value between -1 and 1 that indicates the correlation between x and y.

What do the mean and covariance tell us about the shape of the 2-dimensional Gaussian distribution? Figure 2 illustrates bi-variate Gaussian densities with different parameter settings. As in the univariate case, the mean parameter μ denotes the centre of the distribution. The covariance parameters σ_x, σ_y and ρ change the scaling. When $\rho = 0$ the ellipsoids on the contour plot are axis aligned, when ρ is non-zero they are not. This means that the probability of observing a particular point along the y-axis depends on the location along the x-axis. For example, if x and y represent height and weight of a person, I would expect y to be large for large x. This sort of information is captured in the bi-variate density.

3.2 Multivariate Normal Distribution

To denote a Gaussian distribution over a random variable $\boldsymbol{x} \in \mathbb{R}^N$ we write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$. The density generalises from our previous expression in the bi-variate case.

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

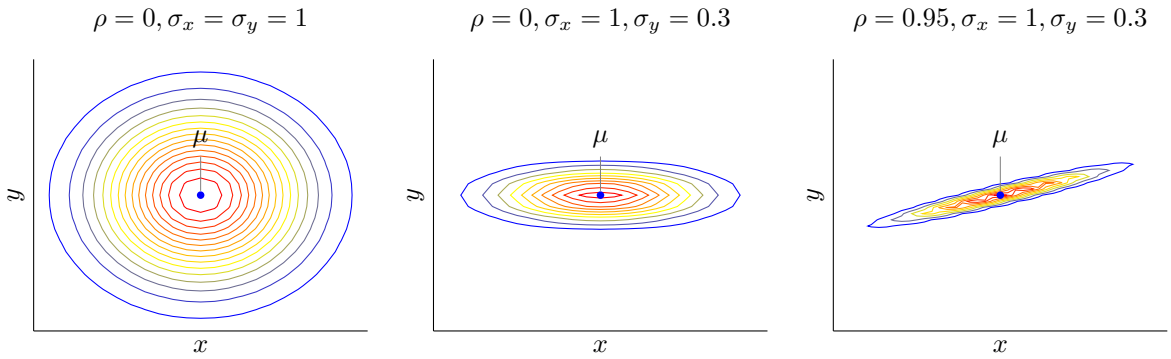

Figure 2: Contour plots of bi-variate Gaussian densities with different covariance parameter settings.

3.3 Example: The mean of a multivariate Gaussian

We can walk through a similar example to the one we had earlier in the univariate case in section 2.3. Perhaps assuming that all height measurements are independent was a bad assumption, because we have actually been given measurements of height in groups, namely those of two parents and their fully-grown child. There might be some dependence between the height of parents and children, so we can't assume that these are iid, but perhaps it is still fair to assume that the measurements across different families are iid. We can again formalise this in a model:

$$\begin{aligned} \boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \Lambda) \\ \boldsymbol{y}_i|\boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{\mu}, \Sigma) \qquad\qquad \text{for } i = 1, \ldots, N \end{aligned}$$

We can follow a similar procedure to the univariate case in section 2.3 to derive maximum likelihood estimates and compute the posterior distribution.

3.3.1 Maximum Likelihood learning

The log-likelihood of the multivariate Gaussian is

$$\ell(\boldsymbol{\mu}, \Sigma) = -\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})$$

To differentiate this expression with respect to $\boldsymbol{\mu}$ we will need matrix derivatives. You can find some useful identities and further references on this in section 4.1. For now, just take for granted

that the gradient is

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}\ell(\boldsymbol{\mu}, \Sigma) &= \sum_{i=1}^{N} \Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) \stackrel{!}{=} 0 \\ 0 &= \Sigma^{-1}\sum_{i=1}^{N} \boldsymbol{y}_i - N\Sigma^{-1}\hat{\boldsymbol{\mu}} \\ \Rightarrow \hat{\boldsymbol{\mu}} &= \frac{1}{N}\boldsymbol{y}_i\end{aligned}$$

3.3.2 Posterior inference

$$\begin{aligned}p(\boldsymbol{\mu}|Y) &\propto p(Y|\boldsymbol{\mu})p(\boldsymbol{\mu}) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\mathsf{T}\Lambda^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu}^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu} - 2\mu^\mathsf{T}\Sigma^{-1}\sum_{i=1}^{N}\boldsymbol{y}_i) - \frac{1}{2}(\boldsymbol{\mu}^\mathsf{T}\Lambda^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\mathsf{T}\Lambda\boldsymbol{\mu}_0)\right) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu}^\mathsf{T}(\Sigma^{-1} + \Lambda^{-1})\boldsymbol{\mu} - 2\mu^\mathsf{T}(\Sigma^{-1}\sum_{i=1}^{N}\boldsymbol{y}_i + \Lambda\boldsymbol{\mu}_0))\right)\end{aligned}$$

We can again recognise the Gaussian functional form and complete the square to obtain our posterior distribution:

$$p(\boldsymbol{\mu}|Y) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - (\Sigma^{-1} + \Lambda^{-1})^{-1}(\Sigma^{-1}\sum_{i=1}^{N}\boldsymbol{y}_i + \Lambda\boldsymbol{\mu}_0))^\mathsf{T}(\Sigma^{-1} + \Lambda^{-1})(\boldsymbol{\mu} - (\Sigma^{-1} + \Lambda^{-1})^{-1}(\Sigma^{-1}\sum_{i=1}^{N}\boldsymbol{y}_i + \Lambda\boldsymbol{\mu}_0)\right)$$

which gives

$$p(\boldsymbol{\mu}|Y) = \mathcal{N}\left(\boldsymbol{\mu}\middle|\left(\Sigma^{-1} + \Lambda^{-1}\right)^{-1}\left(\Sigma^{-1}\sum_{i=1}^{N}\boldsymbol{y}_i + \Lambda\boldsymbol{\mu}_0\right), \left(\Sigma^{-1} + \Lambda^{-1}\right)^{-1}\right)$$

You can recognise the similarities of this solution to the univariate result.

4 Other useful things

4.1 Matrix derivatives

If you work with multivariate densities you will need to get comfortable with manipulating matrices and vectors, and that includes differentiating expressions with respect to matrices and vectors. Here, I include a few identities which I refer back to earlier in the notes or which are generally useful

to have at hand. An extremely useful resource for checking your work is the *Matrix Cookbook* [4].

$$\frac{\partial}{\partial X}\mathsf{Tr}\left[X^{\mathsf{T}}A\right] = A$$
$$\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{a} = \boldsymbol{a}$$
$$\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x} = 2A\boldsymbol{x}$$
$$\frac{\partial}{\partial X}\mathsf{Tr}\left[f(X)^{\top}g(X)\right] = \frac{\partial}{\partial X}\mathsf{Tr}\left[f(X)^{\top}g(Z) + g(X)^{\top}f(Z)\right] \qquad \text{product rule}$$
$$\frac{\partial}{\partial x}f(A(x)) = \mathsf{Tr}\left[\frac{\partial f(A)}{\partial A}^{\top}\frac{\partial A}{\partial x}\right] \qquad \text{chain rule}$$

4.2 Transformation of Variables

It is useful to know how to get the density of a random variable $y = f(x)$, given that you know the distribution of x,$p_x(x)$. This is called transformation of variables.

$$p_y(y) = p_x(f^{-1}(y))\left|\frac{dx}{dy}\right|$$

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[3] Kevin P Murphy. *Machine learning: a probabilistic perspective*. 2012.

[4] Kaare Brandt Petersen et al. The matrix cookbook.