DTU Compute
Department of Applied Mathematics and Computer Science

Cognitive Modeling Lecture Notes

v. 1.23.13

DTU

Contents

CHAPTER 1

Signal Detection Theory and Psychophysics

1.1 Introduction

Perception is the information processing of sensory data in the nervous system that leads to a conscious experience reflecting the outside world. We cannot, of course, measure conscious experience directly but we can define a perceptual task and record the observers response to the task. The response can take many forms: it can be a verbal response, it can be in the form of reaching one's arm towards an object or it could be an eye-movement. However, in this chapter, we will consider only categorical responses, where the observer selects a discrete response from a set of response options. A single pair of a stimulus and a response is called an experimental *trial*.

The perceptual process is noisy. This may seem surprising as our perceptual experience is often clear and vivid. When you look at the letters on this page you see them clearly and perception does not seem noisy. That is because the signal-to-noise ratio of these letters are far above your *perceptual threshold*, the minimal signal-to-noise ratio or stimulus intensity, at which you can just only identify the letters. If you saw the letters at low contrast or only for a very brief amount of time then the signal-to-noise ratio could be near your perceptual threshold. In that case, perception would seem noisy, you would often be unsure of how to respond, and you would often make errors. These errors are stochastic. Even when the stimulus is carefully controlled so that it is identical across multiple trials, the observer's response may not be the same.

We want to quantify an observer's performance in a perceptual task. We will start by calculating the response proportion $\frac{n_r}{N}$ where n_r is the number of trials where the observer chose response category r and N is the total number of trials for a particular stimulus. The response proportion for the correct response option is called the *proportion correct*. At first, we might think that the proportion correct is a reasonable measure of the observer's perceptual sensitivity. After all, we will expect that an observer that performs well will have a high perceptual sensitivity. However, even this simple measure is not just influenced by perceptual sensitivity but also by the observer's *response bias*.

Pure tone audiometry will serve as an example to illustrate the problem in using the proportion correct as a measure of perceptual sensitivity. In this test sinusoidal (pure) tones at various frequencies are played at a certain sound intensity level. You have probably experienced a similar test as it is used as a screening method for frequency specific hearing loss. Every time a sound is played the observer can answer 'yes' to indicate that she could hear the sound or 'no' to indicate that she could not hear the sound. Now assume that an observer answers 'yes' in 100% of the trials in which a tone was played at a certain sound intensity. We may think that this means that the observer has a high perceptual sensitivity but we might well be mistaken: perhaps the observer did not really care to perform well and just responded 'yes' all the time. In order to check whether this was the case we should look at trials in which no sound was played: if the observer also responds 'yes' to all of those trials then we know that the observer was strongly biased towards 'yes'-responses and we have learned nothing about the observers perceptual sensitivity. In order to estimate the observer's perceptual sensitivity we need, in other words, to compare the rate of *true negative* (TN) responses (responding 'no' when no sound was played) *and* the rate of *true positive* (TP) responses (responding 'yes' when a sound was played).

1.2 Signal detection theory

Signal detection theory is a theoretical framework commonly used in cognitive science to separate the cognitive processes of perception and response selection. Perception is modelled as the *encoding* of the stimulus onto a scalar *internal representation*, x. Response selection is modeled as decoding the internal representation, x, to a response category, r.

The encoding process is noisy so that noise, typically assumed to be Gaussian, is added to the encoding x of the stimulus s, so that

$$p(x \mid s) = f(x \mid \mu_s, \sigma^2) = \phi\left(\frac{x - \mu_s}{\sigma}\right) \quad (1.1)$$

where s denotes the stimulus, f denotes the normal probability density and ϕ denotes the *standard* normal probability function for which $\mu = 0$ and $\sigma = 1$.

Importantly, the noise is also present when the stimulus, or signal, is not presented so that

$$p(x \mid s_0) = f(x \mid \mu_0 = 0, \sigma_0 = 1) = \phi(x) \quad (1.2)$$

where s_0 denotes no stimulus, and $\mu_0 = 0$ and $\sigma_0 = 1$ to ensure that the model is identifiable. This is necessary because we need to define the origin (zero) and scale of the internal representation, x. Choosing $\mu_0 = 0$ as the origin seems reasonable: It means that when no stimulus, or signal, is presented the distribution of x is, logically, centered around zero. Setting the $\sigma_0 = 1$ sets the scale of the model, so that, one might say, μ_s is measured in units of σ_0.

Decoding the internal representation, x, to a response option, is modeled by a threshold function, so that observers respond 'no' if $x < c$ and 'yes' if $x > c$ where c is the threshold, or, *response criterion*. We can thus calculate the response probability, $P_{FN} = P(r = no \mid s)$, of a false negative (FN) response as the probability mass of $p(x \mid s)$ that lies below the criterion c as

$$P_{FN} = P(r = no \mid s) = P(x < c \mid s) = \int_{-\infty}^{c} p(x \mid s)\, dx = \Phi\left(\frac{c - \mu_s}{\sigma}\right) \quad (1.3)$$

where Φ denotes the standard normal cumulative distribution function.

Likewise, we can calculate the response probability, $P_{TN} = P(r = no \mid s_0)$, of a true negative response as the probability mass of $p(x \mid s_0)$ that lies below the criterion c as

$$P_{TN} = P(r = no \mid s_0) = P(x < c \mid s_0) = \int_{-\infty}^{c} p(x \mid s_0) = \Phi(c) \quad (1.4)$$

The TN and FN responses are both negative, 'no'-responses but whereas the TN response are correct 'no'-responses when no stimulus was presented, the FN responses are incorrect 'no'-responses when a stimulus ***was*** presented. We can, similarly, also calculate the positive, 'yes'-response probabilities by integration but it is simpler to note that if 'yes' and 'no' are the only response options then

$$P_{FN} + P_{TP} = P(r = no \mid s) + P(r = yes \mid s) = 1$$
$$P_{TN} + P_{FP} = P(r = no \mid s_0) + P(r = yes \mid s_0) = 1$$

From this we can derive the following expressions for response probabilities for positive responses

$$P_{TP} = P(r = yes \mid s) = 1 - P(r = no \mid s) = 1 - \Phi\left(\frac{c - \mu_s}{\sigma}\right) = \Phi\left(\frac{\mu_s - c}{\sigma}\right) \quad (1.5)$$

$$P_{FP} = P(r = yes \mid s_0) = 1 - P(r = no \mid s_0) = 1 - \Phi(c) = \Phi(-c) \quad (1.6)$$

where we have used the identity $1 - \Phi(x) = \Phi(-x)$.

1.2.1 The equal variance model

An additional simplifying assumption commonly used in signal detection theory is the *equal variance model*, which assumes that $\sigma = \sigma_0 = 1$. This model is illustrated in Figure 1.1 where the overlap between the two Gaussian probability distributions $p(x \mid s_0)$ and $p(x \mid s)$ have been filled in two shades of grey. The dark grey area indicates the probability, $P_{FN} = P(r = no \mid s)$, of a false negative response and the lighter grey area indicates the probability, $P_{FP} = P(r = yes \mid s_0)$, of a false positive response.

In the model illustrated in the top panel of Figure 1.1 the criterion, c, is set to $c = \frac{\mu_s}{2} = 0.6$ mid between to two distributions. This is the optimal criterion in terms of maximising the *accuracy*, $P_T = P_{TN} + P_{TP}$, which is the probability of a true response, *if* the probability that a trial contains a stimulus is equal to the probability that a trial does not contain a stimulus $P(s) = P(s_0) = 0.5$. This can be shown by calculating the accuracy, P_T, as

$$P_T = P(r = yes \mid s) P(s) + P(r = no \mid s_0) P(s_0) = \Phi(\mu_s - c) + \Phi(c)$$

and finding the global maximum

$$\arg\max_c (P_T) = \frac{\mu_s}{2}$$

using differentiation.

We can also see that $c = \frac{\mu_s}{2}$ is the optimal criterion by inspecting the lower panel of Figure 1.1, which illustrates the effect of shifting the criterion to a value lower than $c = \frac{\mu_s}{2}$. This increases the probability of a 'yes'-response, which increases the probability, P_{FP}, of a false positive error. It also decreases the probability, P_{FN}, of a false negative error but this decrease is less than the increase in the probability of a false positive, P_{FP}, because $p(x \mid s_0) > p(x \mid s)$ when $x < \frac{\mu_s}{2}$.

For the unbiased observer model in the top panel of Fig. 1, the amount of overlap of the two distributions, and hence the probability of making an error, depends only on the distance μ_s between the distributions and this distance is thus a measures of *perceptual sensitivity*. Therefore the distance μ_s is often denoted as $d' = \mu_s$ only for the equal variance model. Using d' as a measure of perceptual sensitivity is commonly used in the cognitive science literature.

To further illustrate the role of d' as a measure of perceptual sensitivity we will look at some extreme cases. In the extreme case when $d' = \mu_s = 0$, the two distributions overlap completely, so that $p(x \mid s) = p(x \mid s_0)$. In this case the probability of a true or negative response will be the same regardless of whether a stimulus is presented or not, so that $P_{FP} = P_{TP}$ and $P_{FN} = P_{TN}$. This extreme case of $d' = \mu_s = 0$ means that the observer was unable to perceive the stimulus. In the other extreme case when $d' = \mu_s >> 1$ the overlap will be small and the probability of making an error similarly small. This, confirms that $d' = \mu_s$ is a reasonable measure of sensitivity under the assumptions of the model.

The validity of using $d' = \mu_s$ as a measure of sensitivity depends on the validity of the equal variance assumption. Nevertheless, d' is often used as a measure of sensitivity in the cognitive science literature without testing whether the the equal variance assumption holds.

Note that there is an unfortunate confusion of terminology between the signal detection theory literature and the machine learning literature. In the machine learning literature sensitivity typically refers to the probability, P_{TP}, of a true positive response or an estimate thereof. It is a reasonable way to use the word since a very sensitive sensor would detect the stimulus with a high chance of success possibly at the cost of a high chance of making false alarms. Hence, sensitivity in the machine learning literature is influenced by response bias as a high probability of 'yes'-responses lead to a high sensitivity in this sense of the word. Here, I have used the term *perceptual sensitivity* to denote sensitivity in terms of signal detection theory even though the term used in the cognitive science literature is often simply 'sensitivity'.

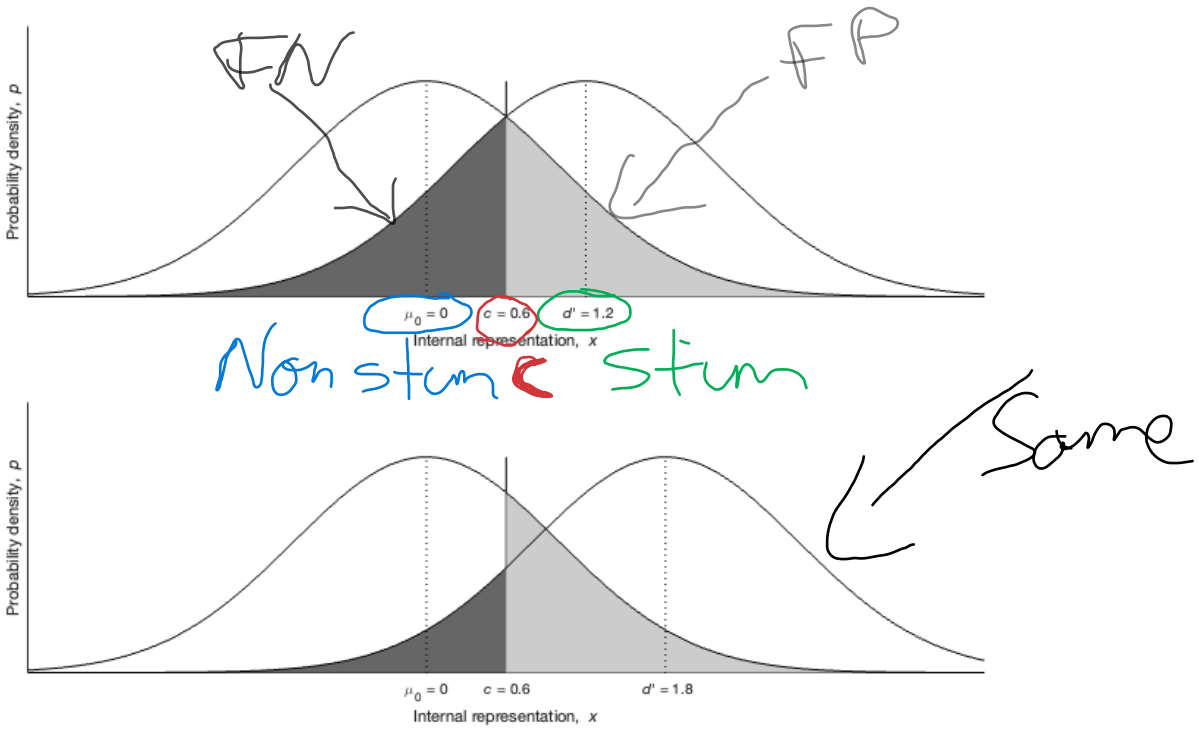


Figure 1.1: The equal variance model from signal detection theory. In the top panel the criterion of the observer, c, is set to the optimal value $c = \frac{d'}{2}$. In the lower panel the criterion, c, is set to bias the observer towards 'yes'-responses. This criterion is suboptimal if the probability that a trial contains a stimulus is equal to the probability that a trial does not contain a stimulus $P(s) = P(s_0) = 0.5$ because it increases the probability of making an error. The area shaded in dark represents the probability of making a false negative error. The area shaded in lighter grey represents the probability of making a false positive error..

1.2.2 Parameter estimation for the equal variance model

So far, we have described how to calculate response probabilities from the model parameters d' and c. In practice, we would face the inverse problem of estimating the model parameters from data. In that case we can estimate the underlying response probabilities from response proportions

$$\hat{P}_{TP} = \hat{P}(r = yes \mid s) = \frac{n_{TP}}{N_s}$$

$$\hat{P}_{FP} = \hat{P}(r = yes \mid s_0) = \frac{n_{FP}}{N_{s_0}}$$

where n_{TP} is the number of true positive responses, n_{FP} is the number of false positive responses, N_s is the number of trials in which a stimulus was presented and N_{s_0} is the number of trials in which no stimulus was presented.

We can now isolate c and μ_s from Equations 1.5 and 1.6 using the inverse standard normal cumulative distribution function, also known as the *probit* function, $\Phi^{-1}(P) = x$ for the equal variance model in which $\sigma = 1$

$$\hat{\mu_s} - \hat{c} = \Phi^{-1}\left(\hat{P}_{TP}\right) \tag{1.7}$$

$$-\hat{c} = \Phi^{-1}\left(\hat{P}_{FP}\right) \tag{1.8}$$

Multiplying Equation 1.8 by -1 gives us an expression for $\hat{c}$

$$\hat{c} = -\Phi^{-1}\left(\hat{P}_{FP}\right) \tag{1.9}$$

Subtracting Equation 1.8 from Equation 1.7 gives us the expression for d'

Perceptual Sensitivity

$$d' = \hat{\mu_s} = \Phi^{-1}\left(\hat{P}_{TP}\right) - \Phi^{-1}\left(\hat{P}_{FP}\right) \tag{1.10}$$

Equations 1.9 and 1.10 provide us with a simple way to estimate the model parameters, c and d', from response proportions. Note that there is no closed form solution to the probit function, $\Phi^{-1}(P)$ but most analysis software contain a function for calculating it.

A problem with estimating the model parameters from Equations 1.9 and 1.10 occurs when any of the response proportions, $\hat{P}_{TP}$ and $\hat{P}_{FP}$ are equal to 0 or 1. In these cases the probit function is undefined. In order to estimate the model parameters, d', and c, the observer must, in other words, give both positive and negative responses, both for trials containing a stimulus and for trials containing no stimulus. The typical problem is that the detection task is too easy so that the observer makes no errors. This means that the perceptual sensitivity, d', is large but we cannot estimate how large it is. The literature contains poor heuristic approaches where an arbitrary small number is added or subtracted to the response proportions to solve this problem but the better solutions is to discard the data. If more data is needed then the experiment should be repeated using a stimulus that has been designed to ensure that the observer use both response categories.

1.2.3 Equal variance model exercise

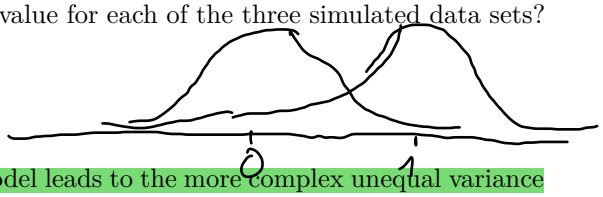
Using a random number generator, simulate responses from 100 experiments with 3 observers each completing 50 trials containing a stimulus and 50 trials containing no stimulus. All three observers behave according to the equal variance model and have a perceptual sensitivity of $d' = 1$ but they have different response criteria: One observer is strongly biased towards 'yes'-responses, one is strongly biased towards 'no'-responses, and one is not very strongly biased towards 'yes'- or 'no'-responses.

d'

Estimate the perceptual sensitivity for each observer and each experiment. Make a histogram of the 100 estimates of the perceptual sensitivity for each of the three observers. Are the distributions centered around the true underlying value for d' each of the three simulated data sets?

Do the same simulations for three observers that behave according to a model where $\mu_s = 1$ and $\sigma = 0.8$. Assume that you did not know that the data came from an observer for which the equal variance assumption $\sigma = 1$ does not hold and estimate d' using the equal variance model just as you did above. Make a histogram of the 100 estimates of the perceptual sensitivity for each of the three observers. Are the estimates centered around the same value for each of the three simulated data sets?

What are the implications of your results?



1.2.4 Unequal variance model

Releasing the constraint, $\sigma = 1$, of the equal variance model leads to the more complex unequal variance model, which is depicted in Figure 1.2. This is specified by three parameters: c, μ_s, and σ but the yes/no-detection paradigm only provides two independent equations, Equations 1.5 and 1.6, relating response probabilities to parameter values. Equation 1.6 can be solved for the response criterion, c, but we are left with one equation (1.5) for determining the two parameters, μ_s and σ, that specifies perceptual sensitivity. The unequal variance model is thus under-determined for the yes/no-paradigm and we need at least one more equation / data point to solve it. Section 1.2.5 will introduce the methods needed for estimating the parameters of the unequal variance model.

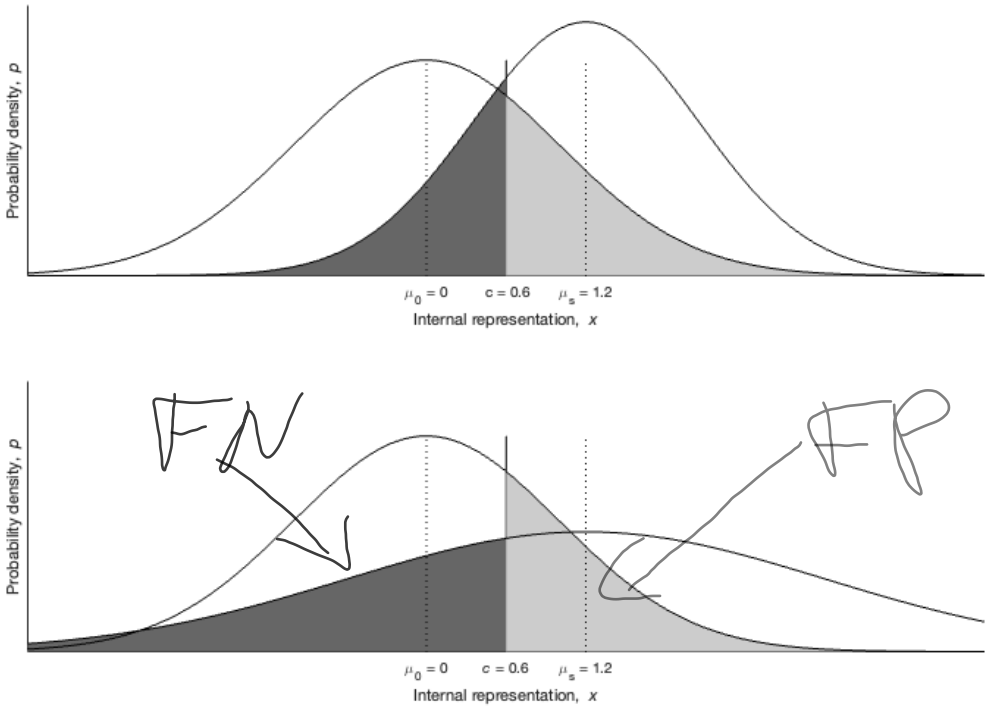


Figure 1.2: The unequal variance model from signal detection theory. The areas shaded in dark represents the probability of making a false negative error. The areas shaded in lighter grey represents the probability of making a false positive error. In the top panel the standard deviation of $p(x \mid s)$ is $\sigma = 0.85 < 1$ and therefore the probability, $p(r = no \mid s)$, of a false negative is smaller than than for the equal variance model depicted in the top panel of Figure 1.1. In the lower panel the standard deviation of $p(x \mid s)$ is $\sigma = 1.8 > 1$ and therefore the probability, $p(r = no \mid s)$, of a false negative is greater than than for the equal variance model depicted in the top panel of Figure 1.1. Note that the criterion is the same for the models depicted here and that depicted in the top panel of Figure 1.1. The probability, $p(r = yes \mid s_0)$, of false positive is therefore the same for all three models..

1.2.5 The Receiver Operating Characteristics (ROC) curve

We can view Equations 1.5-1.6 in a different way by using the probit transform on both sides and rearrange to

$$\Phi^{-1}(P_{TP}) = -\frac{1}{\sigma}c + \frac{\mu_s}{\sigma} \tag{1.11}$$

$$\Phi^{-1}(P_{FP}) = -c \tag{1.12}$$

Inserting Equation 1.12 into Equation 1.11 presents the the problem as a single linear equation

$$\Phi^{-1}(P_{TP}) = \frac{1}{\sigma}\Phi^{-1}(P_{FP}) + \frac{\mu_s}{\sigma} \tag{1.13}$$

Equation 1.13 is the probit transformed *Receiver Operating Characteristics* (ROC) curve. The ROC curve, when it is not probit transformed, is a plot of the probability, P_{TP}, of a true positive response as a function of the probability, P_{FP}, of a false positive response. The upper panels of Figure 1.3 shows ROC curves that are not probit transformed. This is the typical way of depicting ROC curves. The lower panels of Figure 1.3 show the corresponding probit transformed ROC curves.

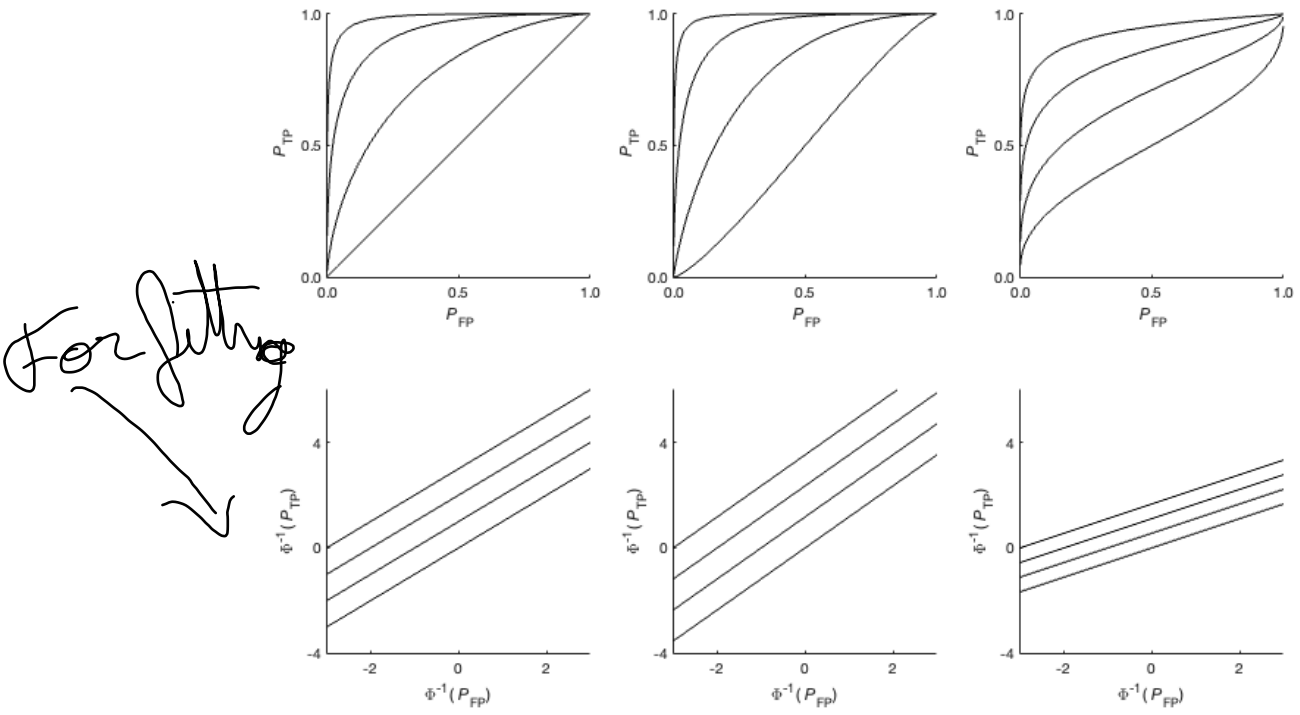


Figure 1.3: The Receiver Operating Characteristics (ROC). The top panels depicts ROC curves for values of $\sigma = 1$ (left), $\sigma = 0.85$ (middle) and $\sigma = 1.8$ (right). The lower panels show the same ROC curves for probit transformed probabilities as described in Equation 1.13. From bottom to top, each curve in every panel shows the ROC curve for values of $\mu_s = 0, 1, 2,$ and 3.

Note that the independent variable (x-axis) of ROC curves (probit transformed or not) depends only on the response criterion. Every point on the curve can thus be mapped onto a value for the

response criterion. The shape of the ROC curve, determined by the slope and the intercept for probit transformed ROC curves, which are linear, are determined by the parameters, μ_s and σ, which are related only to perceptual sensitivity.

The yes/no-paradigm has only one response criterion, c, and thus provides us with a single point on the probit transformed ROC curve, not enough to fit a line except under the equal variance assumption for which the slope is fixed to $\frac{1}{\sigma} = 1$. For the unequal variance model, we need at least one more point, i.e. another response criterion in order to fit the curve. The most efficient experimental paradigm for obtaining more points on the ROC curve is the *confidence rating* paradigm. In this paradigm the observer is offered $N_r \geqslant 2$ *ordered* response categories , so that she can indicate her confidence in her response.

To adapt the unequal variance model to the confidence rating task, we will introduce N_c response criteria that divide the internal representation into $N_r = N_c + 1$ response categories. For each response criterion, $c_i = 1, \ldots, N_c$, the probability, $P_{TP(i)}$, of a true positive response and the probability, $P_{FP(i)}$, of a false positive response can be calculated from the model parameters using Equations 1.5 and 1.6. These probabilities can then be used to calculate the points on the probit transformed ROC according to Equation 1.13.

1.2.6 Parameter estimation for the unequal variance model

In Section 1.2.2 we described how we can estimate the free parameters of the equal variance model from the observed response proportions. Similarly, we can estimate the parameters of the unequal variance model from the observed response proportions by fitting the probit transformed ROC model in Equation 1.13 to them using linear regression. In order to do this we must first calculate the true and false positive response proportions, $\hat{P}_{TP(i)}$ and $\hat{P}_{TP(i)}$. If we index the response categories, r_j, by $j = 0, \ldots, N_c$ then responses in the 0^{th} response category will be counted as negative responses for all response criteria. For the i^{th} response criterion, c_i, responses in categories $r_i, \ldots, r_{N_c}$ will be then counted as positive responses. Therefore, we can calculate the proportion, $\hat{P}_{TP(i)}$ of true positive responses as the sum

$$\hat{P}_{TP(i)} = \frac{1}{N_s} \sum_{j=i}^{N_c} n_{j|s} \tag{1.14}$$

where $n_{j|s}$ denotes the number of responses in the j^{th} response category in trials where the stimulus, s, was presented.

Similarly, we can calculate the proportion, $\hat{P}_{FP(i)}$ of false positive responses as the sum

$$\hat{P}_{FP(i)} = \frac{1}{N_{s_0}} \sum_{j=i}^{N_c} n_{j|s_0} \tag{1.15}$$

where $n_{j|s_0}$ denotes the number of responses in the j^{th} response category in trials where no stimulus, s_0, was presented.

We can use the simple confidence rating paradigm where the response options are '*yes*', '*no*' and '*maybe*' as an example. First, we must order the response categories according to the observer's confidence in a positive response: the 0^{th} response category is, r_0 ='*no*', the 1^{st} response category is, r_1 ='*maybe*', and the 2^{nd} response category is, r_2 ='*yes*'. The number of response categories will be $N_c = 2$. The 1^{st} criterion, c_1, separates the 'no'- and 'maybe'-categories and the 2^{nd}, c_2, separates the 'maybe'- and 'yes'-categories. For the 1^{st} criterion, c_1, we can calculate the proportion, $\hat{P}_{TP(i)}$ of true positive responses using Equation 1.14 as

$$\hat{P}_{TP(1)} = \frac{1}{N_s} \sum_{j=1}^{2} n_{j|s} = \frac{n_{1|s} + n_{2|s}}{N_s} = \frac{n_{maybe|s} + n_{yes|s}}{N_s} \tag{1.16}$$

Like on Mockroo

For the the 2nd criterion, c_2, the expression is simply

$$\hat{P}_{TP(2)} = \frac{1}{N_s}\sum_{j=2}^{2} n_{j|s} = \frac{n_{2|s}}{N_s} = \frac{n_{yes|s}}{N_s} \tag{1.17}$$

The false positive response proportions $\hat{P}_{FP(1)}$ and $\hat{P}_{FP(2)}$ can be calculated similarly based on trials where no stimulus, s_0, was presented.

We can then fit Equation 1.13 to the two pairs of true and false positive response proportions. Obviously, when fitting a line to two data points, the fit will be perfect.

1.2.7 Area under the curve (AUC)

For the unequal variance model, the degree of overlap between the probability densities $p(x \mid s_0)$ and $p(x \mid s)$ and hence the total probability of an error depends not only on the mean, μ_s, but also on the standard deviation, σ, of $p(x \mid s)$. This can be seen by comparing the top and bottom panels of Figure 1.2. In both panels $\mu_s = 1.2$, but the overlap is smaller in the top panel, because $p(x \mid s)$ is more narrow compared to the bottom panel where the overlap is greater because $p(x \mid s)$ is more wide. Perceptual sensitivity is therefore not specified only by $d' = \mu_s$ but also depends on the standard deviation, σ, for the unequal variance model. Note, that decoding the internal representation, x, onto a response category depends only on the criterion, c, for both the equal and unequal variance model.

The relation between the shape of the ROC curve and percpetual sensitivity is well captured by the *Area Under the Curve* (AUC) of the ROC curve when it is not probit transformed. It can be shown that, for the Gaussian models, the area, A, can be calculated as

$$A = \Phi\left(\frac{\mu_s}{\sqrt{1+\sigma_s^2}}\right) \tag{1.18}$$

It can be shown that the AUC equals the accuracy of an unbiased observer in a two-alternative forced choice task in which the observer must choose the one display, out of two, that contains the stimulus, s. This relationship is particularly clear in the equal variance ROC curve displayed in the top left panel of Figure 1.3. For $d' = \mu_s = 0$ the ROC curve is the line $P(r = yes \mid s) = P(r = yes \mid s_0)$, which divides the plane in half so that the accuracy, P_T =A=0.5. This means that the observer is at chance level of a correct answer, which is what we should expect for an unbiased observer with zero perceptual sensitivity, $d' = 0$. As d' increases the ROC curve bulges further and further towards the top left corner where $P_{FP} = 0$ and $P_{TP} = 1$, so that the AUC contains more and more of the entire plane. This means that the accuracy, P_T, increases towards $P_c = 1$, which is what we should expect for an observer with great perceptual sensitivity, $d' >> 1$.

The relationship between accuracy and AUC holds not only for the Gaussian models we have described here but also for models based on arbitrary probability densities. The AUC is therefore a better measure of perceptual sensitivity than e.g. d' because it is more general measure. Note however, that in order to estimate the AUC we need to know the full shape of the ROC, which require some assumptions on the underlying probability densities. We can, of course, interpolate the ROC curve between empirically obtained points but the choice of interpolation method translates to making assumptions on the underlying probability densities.

The AUC is often used to quantify the performance of a classifier in the machine learning literature. For some classification algorithms, it is possible to vary the bias very precisely and at low compuational cost and thereby obtain many points on the ROC. This, in turn, means that the ROC is well described empirically and the effect of the choice of interpolation method is negligible. However, in cognitive science, we cannot ask the observer to vary the bias in a very fine-grained manner. As an example, imagine if you had to rate your confidence in hearing a sound on a 1-100 scale: you would probably struggle with distinguishing between, e.g. confidence levels 77 and 78 and you might choose to use only some anchor points on the scale. Even if you could distinguish meaningfully between 100 confidence

levels, it would require many trials before you would actually use all the levels. Therefore, in the cognitive science literature, the observer is rarely offered more than seven response categories.

1.2.8 Unequal variance model exercise

In a confidence rating task, with four response categories, the observer can indicate her confidence as 'high' or 'low', in addition to answering 'yes' or 'no'. Using a random number generator, simulate responses from 100 experiments with one observer completing 50 trials containing a stimulus and 50 trials containing no stimulus. The observer behave according to the unequal variance model $\mu_s = 1$ and $\sigma = 0.8$.

Estimate the parameters of the model for each experiment. Also estimate the AUC. Plot the distribution of the parameters and AUC across experiments. Are the distributions of the parameters and the AUC centered around the true underlying values?

Compare your results to the equal variance model exercise in 1.2.3.

1.3 The psychometric function

In the signal detection task discussed in Section 1.2, only two stimuli, s and s_0, are presented to the observer. We can extend this experimental paradigm to include stimuli at multiple stimulus intensities. In their basic form, the signal detection models described in Section 1.2 do not apply to this paradigm. Instead, we can use the psychometric function, $\Psi(I) = P(r = yes \mid I)$, where I is the stimulus intensity to model an observer's performance and quantify perceptual sensitivity and response bias.

First, we must parameterise the psychometric function. Since we expect the probability, $P(r = yes \mid I)$ of positive, 'yes'-responses, to increase with stimulus intensity, I, the function should be monotonically increasing with respect to the stimulus intensity, I. Also, the psychometric function must return a probability, so it must be bounded $0 \leqslant \Psi(I) \leqslant 1$. These requirements are satisfied by cumulative distribution functions. A commonly used function is the Gaussian cumulative distribution function, Φ, which we will also use here. Then we can write the psychometric function as

$$\Psi(I) = P(r = yes \mid I) = \Phi\left(\frac{I - c_I}{\sigma_I}\right) \tag{1.19}$$

where c_1 is the 50%-threshold, since $\Psi(I) = 0.5$ when $I = c_I$, and σ_I is the standard deviation in units of stimulus intensity.

If we compare the psychometric function in 1.19 to the expression for the probability of a true positive response, P_{TP}, in Equation 1.5 from signal detection theory, we see that they are quite similar. This is unsurprising as both expressions return the probability of a positive, 'yes'-response. However, there is a very important difference between the two expressions: Whereas the input and parameters of the psychometric functions are in units of stimulus intensity, the input and parameters of the signal detection models are unitless. Recall from Equation 1.2 in Section 1.2 that fixing the standard deviation, $\sigma_0 = 1$ of the noise added in the encoding process when no stimulus, s_0, is presented sets the scale for the signal detection models. We can use the same approach for the psychometric function by rewriting it to

$$\Psi(I) = P(r = yes \mid I) = \Phi\left(\frac{I}{\sigma_I} - \frac{c_I}{\sigma_I}\right) \tag{1.20}$$

and compare this to the probability, P_{TP}, for the equal variance model, which we obtain by rewriting Equation 1.5 to

$$P_{TP} = \Phi(d' - c) \tag{1.21}$$

to see that the two models are identical if

$$d' = \frac{I}{\sigma_I} \quad \text{and} \quad c = \frac{c_I}{\sigma_I} \tag{1.22}$$

The relationship between the psychometric function and the equal variance signal detection model is illustrated in Figure 1.4. The top panel depicts a psychometric function and the lower panel depicts the corresponding internal representation scaled in units of stimulus intensity unlike the signal detection models, which are unitless. Note the graphical interpretation of the relationship: As the stimulus intensity increases, so does the probability mass of the Gaussian probability density, shaded in grey, that lies above the 50%-threshold.

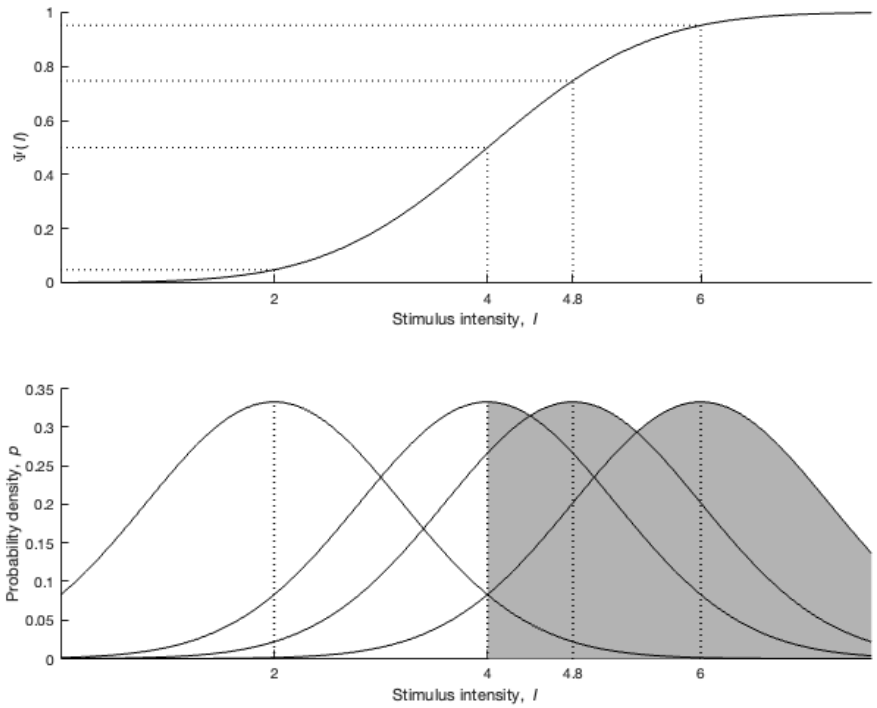


Figure 1.4: The psychometric function. The top panel depicts the cumulative Gaussian psychometric function. The lower panel depict the noisy internal representation of stimulus intensity as the underlying model for the Gaussian psychometric function. The shaded area under each probability density function represents the probability of a 'yes'-response. The 50%-threshold of the observer lies at intensity value of 4. Note that the x-axis is in units of stimulus intensity, I, in both panels..

The standard deviation of the psychometric function, σ_I, reflects the perceptual sensitivity of the observer. Low values of σ means that little noise is added in the perceptual process, so that the encoding of the stimulus intensity is very precise. This means that the psychometric function is steep: A small increase in stimulus intensity can increase the perceived stimulus intensity above the criterion so that the observer will detect the stimulus consistently. Hence the perceptual sensitivity is high when σ_I is small. This might seem different from the equal variance signal detection model. For this model, the sensitivity was quantified entirely by the distance, d', between the probability densities $p\,(r = yes \mid s_0)$ and $p\,(r = yes \mid s)$. However, this distance was measured relative to the standard deviation, σ, which

was set to $\sigma = 1$. When we rescale the signal detection model to physical units, using the expression $d' = \frac{I}{\sigma_I}$, we see that d' increases when σ_I decreases for a fixed stimulus intensity, I.

Signal detection theory, including the psychometric function described above, rests on the assumption that the stimulus must be encoded onto a one-dimensional internal representation, x. In many perceptual tasks these assumptions do not hold. Take for instance a letter identification task. The letters are high-dimensional stimuli and so is their internal representation. However, we can still use the psychometric function to model the proportion correct, P_c, as a function of stimulus intensity, I, even though the observer model illustrated in the lower panel of Figure 1.4 does not hold.

In *forced choice* paradigms, the observer must respond in one of N_r response categories. For a letter identification task, with one response category for each letter in the English alphabet, $N_r = 26$. Using the psychometric function in this task requires that we consider how the observer performs when the signal intensity, I, is near zero. The observer might not be able to perceive the letter but still has to respond and will have a probability $P_{guess} = \frac{1}{N_r}$ of guessing correctly. This effect of guessing can influence performance even at higher stimulus intensities, where the observer might not perceive the stimulus due to random fluctuations in the noisy encoding process, but still has to guess. The high threshold psychometric function in Equation 1.23 is often used to model this behavior

$$\Psi(I) = \Phi\left(\frac{I - c_I}{\sigma_I}\right) + \left(1 - \Phi\left(\frac{I - c_I}{\sigma_I}\right)\right) P_{guess} = (1 - P_{guess})\,\Phi\left(\frac{I - c_I}{\sigma_I}\right) + P_{guess} \tag{1.23}$$

The high threshold model assumes that the observer will *either* perceive the stimulus correctly with the probability given by the psychometric function in Equation 1.19 *or* not perceive it correctly and *only then* guess among the response categories. The probability, P_{guess}, of guessing correctly is sometimes assumed to be $P_{guess} = \frac{1}{N_r}$ but othertimes it is left as a free parameter in the model.

Note that, in general, $\Psi(c_I) = 0.5 + 0.5P_{guess} \neq 0.5$, for the high threshold psychometric function. Nevertheless, c_I, is still often referred to as the 50%-threshold. Sometimes this is made explicit by naming c_I the 50%-threshold *corrected for guessing*.

Another correction often made to the psychometric function is correcting for *lapsing*. Lapsing is when an observer fails to perform the task perhaps due to a lapse in attention. The problem with lapsing is that it is independent of the stimulus intensity and may thus occur even when the stimulus intensity, and hence the probability of a correct response, is high. This can have a large influence on the parameter estimates. The logic for correcting for lapsing is similar to the logic for correcting for guessing: *either* the observer lapse, with a probability of P_{lapse}, and will be forced to guess, *or*, does not lapse and responds with a probability given by the psychometric function in Equation 1.19 or 1.23. The psychometric function corrected for guessing and lapsing is given

$$\Psi(I) = (1 - P_{guess} - P_{lapse})\,\Phi\left(\frac{I - c_I}{\sigma_I}\right) + P_{guess} \tag{1.24}$$

1.3.1 Parameter estimation for the psychometric function

Theoretically, we could estimate the parameters of the psychometric function in Equation 1.19 by probit transforming the response proportions $\hat{P}(r = yes \mid I_s)$ and fitting a line to them much as we did for ROC curves. In practise, this is not a good approach since $\Phi^{-1}(P)$ is undefined for $P = 1$ and $P = 0$. Therefore, we would be unable to fit the psychometric function using this approach if the response proportion was equal to 0 or 1. Also, this approach does not work for if the probability of lapsing or the probability of guessing is a free parameter.

In stead of using the probit transform, we can use the more general approach of finding the parameter values for σ_I and c_I that maximise the likelihood $\mathcal{L}(n \mid \sigma_I, c_I)$ of the observed response counts, n_s, for a specific stimulus, s. The response counts can be the number of 'yes'-responses or the number correct responses depending on the experimental paradigm. Since we, in both cases, have two response options, the response counts follow a binomial distribution, so that the likelihood of the response counts, n_s, for a particular stimulus, s, is given by Equation 1.25.

$$\mathcal{L}(n_s \mid \sigma_I, c_I) = \binom{N_s}{n_s} P_s^{n_s} (1 - P_s)^{N_s - n_s} \tag{1.25}$$

where $P_s = \Psi(I)$ is the response probability given by the psychometric function and N_s is the number of trials for the particular stimulus, s.

We must, of course, fit the psychometric function to the response counts for *all* the stimuli, not just one specific stimulus. We do this based on the assumption that the response counts for each stimulus is independent of the response counts for the other stimuli so that the total likelihood is the product over stimuli, s, of the likelihood in Equation 1.25. This turns out to be problematic because the value of the likelihood can be very small, below machine precision. We can solve this problem by maximising the log likelihood (logarithm of the likelihood), $\mathcal{L}(n_s \mid \sigma, c)$ for a particular stimulus, s, as in Equation 1.26.

$$\log\left(\mathcal{L}\left(n_s \mid \sigma_I, c_\right)\right) = \sum_{i=1}^{N_s} \log(i) - \sum_{i=1}^{n_s} \log(i) - \sum_{i=1}^{N_s - n_s} \log(i) + n_s \log(P_s) + (N_s - n_s) \log(1 - P_s) \tag{1.26}$$

Again, we must, of course, fit the psychometric function to the response counts for all the stimuli. Since the total likelihood is the product product over stimuli, s, of the likelihood in Equation 1.25, the total log likelihood is the sum over stimuli, s, of the log likelihood in Equation 1.26.

Note that, in order to avoid very small numerical values in the calculation of the log likelihood, $\log(\mathcal{L}(n_s \mid \sigma_I, c_I))$, it is important to take the logarithm for each term as specified in Equation 1.26 rather than first calculating the likelihood $\mathcal{L}(n_s \mid \sigma_I, c_I)$ from Equation 1.25 and then taking the logarithm.

Finally, note that, the negative log likelihood $-\log(\mathcal{L}(n_s \mid \sigma_I, c_I))$ can be thought of as a *cost function*. Per convention, model fitting is often performed by minimising the cost function. Hence we may seek to minimise the negative log likelihood rather than maximising the log likelihood.

1.3.2 Psychometric function Exercise

In a 3-alternative classification task, the observer classifies speech sounds under varying sound intensities. The experiment consists of 30 experimental trials at each sound intensity. The sound intensities and the corresponding number of correct responses are shown in the table below.

Stimulus intensity, I_s (dB)	5	10	15	20	25	30
Number of correct responses	12	11	19	27	30	30

Fit the each of the three psychometric functions in Equations 1.19, 1.23 and 1.24 to the data. You can assume that the probability, P_{guess}, of guessing a correct response is $P_{guess} = \frac{1}{N_r}$ but the probability, P_{lapse}, of lapsing should be a free parameter of your model.

- Make one plot with the three psychometric functions as curves with and the response proportions as points. Estimate, by visual inspection, which of the psychometric functions fit the data better
- List the value of the negative log likelihood of the three psychometric functions. Which is lower?
- The negative log likelihood is not a good way to evaluate model fits as models with more free parameters are more flexible and will therefore provide a better fits. Akaike's Information Criterion, $AIC = 2(N_p - \log(\mathcal{L}))$, where N_p is the number of free parameters is a better measure as it corrects for the number of free parameters. Calculate the AIC for the psychometric functions.
- List the parameter values for each of the the psychometric functions. Do they give similar estimates of the parameter values?

- Repeat all of the above steps in the case that the observer lapses once at the highest stimulus intensity level, $I_s = 30$, so that the number of correct responses is $N_c = 29$, for that level.
- Discuss your findings. Does the introduction of guessing and lapsing influence your analysis of the data?

1.4 Magnitude estimation

In the previous section we described performance in detection tasks using the psychometric function $\Psi(I)$ as a function of the stimulus intensity, I, but we did not describe how to quantify, or measure, stimulus intensity. We also saw that the perceptual sensitivity is proportional to stimulus intensity in 1.22. Intuitively, we might expect that this assumption generalises to a different experimental paradigm: the magnitude estimation task in which the observer is presented with a stimulus and asked to estimate the stimulus intensity. If perceptual sensitivity is proportional to stimulus intensity, we would expect that the the estimated or *perceived* intensity, I_p should be proportional to the *physical* stimulus intensity, I_s. We will, however, learn that this relationship is, in general, more complex and that we will therefore have to transform the *physical* stimulus intensity, I_s, to a measure of *perceived* stimulus intensity, I_p. We will therefore need to replace the nondescript measure of stimulus intensity, I, used in the psychometric function by perceived stimulus intensity, I_p.

Loudness, the perceived intensity of sound, provides a good example of how perceived stimulus intensity is different from physical stimulus intensity. Sound is changes in air pressure and we can therefore measure the physical intensity of sound in the SI-unit for pressure, pascal (Pa). However, this measure aligns poorly with perceived sound intensity. The more commonly used measure of sound intensity is decibels (dB) because it aligns better with perceived sound intensity, so that for sound, intensity is measured as

$$I_p(b) = 20 \log_{10} \left(\frac{b}{b_0} \right) dB \tag{1.27}$$

where b is the sound pressure in units of pascal. The constant, b_0, is the typical hearing threshold for humans, which can be seen by noticing that $I_p(b_0) = 0$.

The logarithmic decibel scale stems back to what is some of the earliest work in cognitive modeling: Fechner's law of perceived magnitude, or, intensity. Fechner based his work on the observations of Weber who studied perceptual sensitivity in *change detection* tasks. Using an adaptive procedure in which the change in physical stimulus intensity, ΔI_s, is adjusted to a level where the observer can just notice the difference, Weber measured the change detection threshold and named it the *just noticeable difference* (JND). Weber found that the JND is proportional to the baseline stimulus intensity, I_s, across many perceptual modalities

$$\Delta I_s = k_w I_s \tag{1.28}$$

where k_w is called the Weber fraction.

Strikingly, Weber's law seems to hold in good approximation to a wide variety of perceptual phenomena. One example studied by Weber is the perceived intensity of the of the gravitational force on a weight held in one hand. Imagine you're holding a 10 gram weight in the palm of your hand. You might be able to detect if just one more gram is added to the weight. If you were holding a weight of 1 kg you would not be able to detect a change of just 1 g. It would take a change of 100 grams before you would be able to notice the change according to Weber's law. Other examples include changes in the taste of sweetness, the number of dots seen on a page and the perception of duration of time. Note, however, that the value of Weber fraction $k_w = \frac{\Delta I_s}{I_s}$ is different for for different perceptual modalities. We can thus use the Weber fraction, k_w, as a measure that compares the perceptual sensitivity across perceptual modalities. A small Weber fraction value means that a small relative change in stimulus intensity can be detected meaning that the observer is very sensitive to this modality.

Note that we can rearrange equation to

$$\frac{1}{k_w}\frac{\Delta I_s}{I_s} = 1 \tag{1.29}$$

Fechner hypothesised that Equation 1.29 could provide a unit for the change in perceived magnitude, I_p, so that, in the limit $\Delta I_s \to 0$, perceived intensity would change as

$$dI_p = \frac{1}{k_w}\frac{dI_s}{I_s} \tag{1.30}$$

perceived magnitude, I_p.

In order to find the actual perceived stimulus intensity, I_p, Fechner found his law by integrating Equation 1.30

$$I_p = \int_{I_0}^{I_s} \frac{1}{k_w}\frac{dI_s}{I_s} = \frac{1}{k_w}\ln\left(\frac{I_s}{I_0}\right) \tag{1.31}$$

where I_0 is the absolute threshold, the minimal intensity value that can be perceived. Comparing Fechner's law in Equation 1.31 to the decibel scale in Equation 1.27 we see that they are identical except for the choice of base in the logarithm.

Fechner published his studies in 1860 and it was only seriously challenged 100 years later when Stevens showed that it does not apply generally to all sensory modalities. It fails, for example, to describe magnitude perception for color saturation, temperature on the skin and electric shock. Stevens introduced an alternative law of magnitude perception now known as Stevens' law, which describes perceived stimulus intensity as a power function of physical stimulus intensity.

$$I_p = k_s I_s^a \tag{1.32}$$

where the parameters k_s and the exponent, a varies across stimulus modalities.

Stevens' law seem to apply reasonably well to a very wide range of stimulus modalities including those, like sound intensity, for which Fechner's law applies. The reason for this is that power laws with an exponent $a < 1$ can have a shape similar to that of a logarithmic function.

Fechner and Stevens' work informs us on the choice of unit to use for the stimulus intensity when fitting a psychometric function. If previous studies have shown that perceived intensity follows one of Fechner and Stevens's laws then the physical unit should be chosen accordingly. It is therefore common to use dB as the unit of sound intensity because perceived sound intensity has been shown to be approximated well by a logarithmic function. For other units may apply for other modalities and experimental paradigms.

1.4.1 Magnitude estimation exercise

Although mathematically different, Fechner and Stevens' laws of perceptual intensity provide fairly good fits to perceived brightness as a function of luminance. This is because the exponent of Stevens' law is approximately $a = 0.33 < 1$ so that

$$I_p = 10 I_s^{0.33}$$

To see that this relationship might be mistaken for a logarithmic relationship, first calculate the perceived stimulus intensity, I_p, for physical intensities $I_s = 1, 2, \ldots, 10$ using Steven's law. This simulates an observer that rates the perceived intensity according to Stevens' law. Fit Fechner's law to the simulated data. Note that Fechner's law is linear with respect to I_s.

- List the parameter values for Fechner's law
- Plot the simulated data and curve showing Fechner's law

- Evaluate whether Fechner's law provides a reasonable fit by visual inspection of the simulated data and the model

For electric shock, Stevens found the exponent to be approximately $a = 3.3 > 1$. As before, calculate the perceived stimulus intensity for physical intensities $I_s = 1, 2, \ldots, 10$ using Stevens' law and fit Fechner's law to the simulated data.

- List the parameter values for Fechner's law
- Plot the simulated data and curve showing Fechner's law
- Evaluate whether Fechner's law provides a reasonable fit by visual inspection of the simulated data and the model

CHAPTER 2

Linear encoding

2.1 Introduction

Our senses are able to encode high-dimensional stimuli to lower-dimensional feature spaces. Typically, these features are of particular relevance to our actions and, hence, ultimately, to our survival.

As an example, humans can encode images to extract relevant features such as facial expression. If we represent a monochrome image as a vector where each pixel is a dimension then the value of each dimension represents the brightness of the corresponding pixel. If we assume that an image is *linearly* encoded onto an internal representation, x, of some feature then

$$x = \boldsymbol{i}^T \boldsymbol{w} + \delta + \epsilon \tag{2.1}$$

where, $\boldsymbol{i}$, is the image column vector, $\boldsymbol{w}$ is a weight column vector, δ is the intercept of the model, and $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ is the noise added in the encoding process.

The linear encoding model is a very simplified model, which, in general is too simple. Still, it does capture some important aspects of the encoding process. It does, however, rely on some assumptions. Importantly, the linear encoding model relies on the assumption that the weight of each pixel is the same for all images. A pixel must, in other words, represent the same aspect across all images. This is a reasonable assumption if the images are aligned. For frontal images of faces, this can be achieved by scaling, shifting and rotating the images so that the two eyes are in the same pixels in all images.

In the previous chapters we assumed that the stimuli were carefully designed and delivered to have a stimulus intensity known with great precision. If we want to study how the observer perceives facial features, such as a smile or gender characteristics, we cannot control the stimulus in the same way. First, we may not know the relevant image features and even if we did, we would probably be unable to control them precisely. How should we generate a image of a face with typical feminine or masculine features?

Unable to control the stimulus precisely we can instead sample from a repository or generate our own samples, i.e. we can obtain a random selection of N photographs of faces. We can then ask observers to rate the images with respect to a certain facial feature. We will assume that their rating reflect the internal representation value, x, and therefore refer to both the rating and the internal representation value as x. This experiment will give us a N-dimensional column vector, $\boldsymbol{x}$, of ratings, where each entry reflects the internal representation value for a particular image, and a matrix, $\boldsymbol{I}$, of images, where each row represents an image and each column represents a pixel that, according to Equation 2.1, are related as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}_1^T \\ \vdots \\ \boldsymbol{i}_N^T \end{bmatrix} \boldsymbol{w} + \delta + \boldsymbol{\epsilon} = \boldsymbol{I}\boldsymbol{w} + \delta + \boldsymbol{\epsilon} \tag{2.2}$$

If the images each consist of M pixels then the image vectors $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_N$ and the weight vector, $\boldsymbol{w}$, are M-dimensional and $\boldsymbol{I}$ is an N-by-M matrix. Note that a certain image can be rated multiple times. In that case the image will be repeated across multiple rows of the image matrix, $\boldsymbol{I}$. The *intercept*, denoted as δ, and the weight vector, $\boldsymbol{w}$ are free parameters that can be stacked into a parameter vector by rewriting Equation 2.2 to

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}_1^T & 1 \\ \vdots & \vdots \\ \boldsymbol{i}_N^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ \delta \end{bmatrix} + \epsilon = \boldsymbol{I}_\delta \begin{bmatrix} \boldsymbol{w} \\ \delta \end{bmatrix} + \epsilon \tag{2.3}$$

If we have more equations than unknowns, $(N > M)$, then Equation 2.3 can be solved using the normal equations to find

$$\begin{bmatrix} \boldsymbol{w} \\ \delta \end{bmatrix} = \left(\boldsymbol{I}_\delta^T \boldsymbol{I}_\delta \right)^{-1} \boldsymbol{I}_\delta^T \boldsymbol{x} \tag{2.4}$$

where $\hat{\boldsymbol{w}} = \left(\boldsymbol{I}^T \boldsymbol{I} \right)^{-1} \boldsymbol{I}^T$ is called the *pseudo-inverse* matrix of $\boldsymbol{I}$. This solution is described in more detail in Chapter 8 in the Introduction to Machine Learning book by Herlau, Schmidt and Mørup [HMS21].

Before we attempt to solve Equation 2.2 using Equation 2.4 we must consider the relationship between the number of, $M + 1$, of free parameters and the number of equations, N. The resolution of digital images is typically measured in megapixels, meaning that $M > 10^6$. This corresponds reasonably well with the number of photo-receptors in the human eye, normally tens of millions. Compare this number to the number of images that an observer, or a group of observers, can rate in a reasonable amount of time. If we assume that it takes 2 s to rate one image then an observer can rate 1800 images in half an hour. The task is tedious and we cannot expect an observer to perform well for more than an hour without breaks. Assume that we can obtain approximately 1000 reliable ratings in one session with one observer, then we will need thousands of experimental sessions before $N = M$. Fortunately, we can use far fewer ratings even though this leaves Equation 2.2 extremely *ill-posed*, meaning that we have more parameters than data points (images). In order to do so, we must *regularise* Equation 2.2. We can do this by reducing the dimensionality of the images.

The simplest steps to reduce the dimensionality of the images rated by observers is to use cropping, down-sampling and gray-scaling. If the task of the observers is to rate facial features then we may crop the images to contain only the face and not the hair or any irrelevant background. For cropped images of faces, it is possible to reliably detect facial features in images with a resolution of only thousands of pixels, which is a significant reduction in the dimensionality, M, of the images compared to the millions of dimensions in uncropped megapixel images. Many facial features are capture by gray-scale photos, which have only one value per pixel as opposed to typical color images. Therefore, in most cases, using gray-scale images instead of color images will reduce the dimensionality of the images with a factor of 3 without reducing the relevant information in the images significantly.

The next step in reducing the dimensionality of images is to use *Principal Component Analysis* (PCA). PCA provides a new orthogonal basis, the principal components, (PCs) for the image vector. The first PC points in the direction of the most variance in image space. The second PC points in the direction of the second most variance in image space and so forth. The total number of PCs equals the number of dimensions, M, in the original image space *if* we have more images than than the image dimensionality, i.e. if $M > N$. If, as in our case, the images have more dimensions than the number of images, i.e. $N > M$ then the number of PCs will equal the number of images, N.

Since the PCs are orthogonal, the sum of the variance across PCs equals the total amount of variance of the images. By projecting the images into PC space we will often find that most of the variance is contained in the first few PCs. These PCs define a *subspace* that has a dimensionality, which is typically much lower than the original image space. By projecting the images onto this subspace we obtain a representation, of the images, known as the *score*, which is of a much lower dimensionality than the original images. Replacing the images, $\boldsymbol{I}$, with their lower-dimensional scores, $\tilde{\boldsymbol{I}}$ in Equations 2.1–2.3 thus provides us with equations that are not ill-posed. This approach is explained in detail in Chapter 3, especially subsection 3.4.2, in the Introduction to Machine Learning book by Herlau, Schmidt and Mørup [HMS21].

We may be able to reduce the dimensionality of the images even further. Variance in image space will be due to variance in many image features. For images of faces, the variance can be related to

the angle of the light source, gender characteristics, eyewear and facial expression. Selecting the PCs containing most of the variance will include the variance across all those features. If possible, we would like to select the PCs that include variance related to the image feature that we intend to capture in the weight vector, $\boldsymbol{w}$. We can use methods like *forward selection* to achieve this. Forward selection is an iterative procedure in which the model is first fitted using each of the PCs included in the subspace, one at a time, so that the number of model fits equals the number of PCs included in the subspace. Each model is then evaluated. The PC that gives the best model will be included in the final model. Then, the model is fitted again using each of the remaining PCs, one at a time, *and* the PC that was already selected in the first iteration. This process is repeated until the model is no longer improved by adding additional PCs. In this way a number of PCs that are relevant for improving the model is selected.

Note that in forward selection the models should not be evaluated by their goodness-of-fit. In that case, forward selection would select all PCs in the subspace more free parameters generally provide better fits. Instead the models should be evaluated using *cross-validation* where the models are fitted to a *training set* and evaluated on a *test set*. Once the PCs that will be included in the final model have been selected using forward selection, the model can be fitted to all the data to obtain an estimate of the model parameters. Cross-validation and the forward selection approach to feature selection is described in more detail in Chapter 10 in Introduction to Machine Learning book by Herlau, Schmidt and Mørup [HMS21].

The model described in Equation 2.2 describes a line in image space, or, in a subspace of image space. In order to validate the model we can visualise a point on the line corresponding to a rating, $\boldsymbol{x_0}$, as an image, $\boldsymbol{i_0}$. In order to do so we need to solve Equation 2.1 for $\boldsymbol{i}$ given the weight vector, $\boldsymbol{w}$, and the intercept, δ, which are parameter values obtained by fitting the model. Note that there is, an $(M-1)$-dimensional, space of images that will have the same rating according to Equation 2.1. This subspace consists of images that all have the same rating but varies in image components not relevant to the rating, i.e. they are orthogonal to the weight vector, $\boldsymbol{w}$. Note that adding such components to $\boldsymbol{i}$ in Equation 2.1 do not change the rating, x. We can find a unique image, $\boldsymbol{i_0}$, containing no such irrelevant components for a given rating, $\boldsymbol{x_0}$, by constraining the solution to be parallel to the weight vector, $\boldsymbol{w}$, so that $\boldsymbol{i_0 = \alpha w}$ where α is a scalar.

$$\boldsymbol{x_0} = \boldsymbol{i}_0^T \boldsymbol{w} + \delta = \alpha \boldsymbol{w}^T \boldsymbol{w} + \delta = \boldsymbol{\alpha \|w\|^2 + \delta} \tag{2.5}$$

Solving Equation 2.5 for alpha gives us

$$\boldsymbol{\alpha} = \frac{\boldsymbol{x_0} - \boldsymbol{\delta}}{\boldsymbol{\|w\|}^2} \tag{2.6}$$

and hence

$$\boldsymbol{i_0 = \alpha w} = (\boldsymbol{x_0} - \boldsymbol{\delta}) \frac{\boldsymbol{w}}{\boldsymbol{\|w\|^2}} \tag{2.7}$$

The solution to 2.7 can be interpreted in a fairly intuitive manner: It is the scaled weight vector multiplied by the difference between the rating, x_0 and the intercept, δ. Compare this to Equation 2.1 to note that the intercept, δ, is the average rating for an image that is orthorgonal to the weight vector meaning that the image contains zero information relevant to the rating task. The intercept can thus be seen as a baseline, or bias, since it is the rating given to a neutral face.

Fagertun, Andersen and Paulsen [FAP12] used a similar approach for creating synthetic images with varying gender characteristics. In their approach the rating, x, of gender characteristics, ranging from feminine to masculine, was calculated using binary responses and reaction times, so that the rating was deemed more extreme if for faster responses. They validated the synthetic images informally using visual inspection and found that typical masculine characteristics such as facial hair was present in images rated more masculine. This result was confirmed across three publicly available image sets that were constrained to contain cropped frontal facial images and in one image set that was obtained by collecting images with greater variability from LinkedIn™. Fagertun, Andersen, Hansen and Paulsen

[Fag+13] used the same approach on high resolution three-dimensional images dividing the images into shape and texture components. They observed that synthetically generated more masculine images showed characteristic masculine head shape features such as a stronger jaw line and also characteristic masculine texture features such as beard stubble. This confirms that the approach can be used to create synthetic images varying in only in a single perceptual feature.

2.2 Linear encocding exercise

2.2.1 Overview

In this exercise, you will run two simple experiments. You will also create and validate a linear encoding model.

In Experiment 1, photos of faces are presented to the test participants. The participants must rate each face on a scale to indicate a facial feature. The facial feature can be the emotional expression of the face (e.g. happy, sad or angry) or it could be the perceived gender of the face. It could also be some other feature that select.

The participants' ratings are used to create a linear model in which the participants' rating is the dependent variable and the images, or rather, their principal component scores, are the predictors. You will use the linear model to create synthetic images of faces that correspond to given ratings. Some of the methods you will use are described in more detail in the Introduction to Machine Learning book by Herlau, Schmidt and Mørup [HMS21].

In Experiment 2 synthetic images of faces are presented to the test participants. The participants must rate each face on a scale similar to the one used in Experiment 1. The data from Experiment 2 are analysed using the psychometric function and ROC curves. The purpose of Experiment 2 is to validate the linear model experimentally.

This exercise is inspired by the papers by Fagertun et al. [FAP12; Fag+13]. You can find the papers on DTU Findit.

2.2.2 Instructions

Follow the instructions below to complete this exercise. Note that there are questions in bold in the instructions. Your report should consist of the answers to these questions. Please hand in your answers to the exercise as a pdf file. If you like, you can also submit your code and data (the images you used and the ratings).

1. Obtain photos of faces for your experiment. First, choose a database from which you will obtain the photos of face that you will use as stimuli. Also, decide on the facial feature that you want to use. It can be an emotional expression, perceived gender or some other feature that you would like to use.

   - 'Aligned & Cropped' images from UTKFace (https://susanqq.github.io/UTKFace/)
   - Karolinska Directed Emotional Faces (KDEF) (https://kdef.se/)
   - FEI Face Database (https://fei.edu.br/ cet/facedatabase.html)

   Then select the images that you want to use from the database you have selected. Choose a subset of at least 200 images from the database. For this part, you will have to manually review the files and/or use the tags set by the creators of the database (often contained in the filenames). Select the images based on the following two criteria

   - The images should vary across the facial feature that you want to use. For instance, if the participants are rating the images on a scale ranging from happy to angry then you should

make sure to include an approximately equal number of faces that appear happy, neutral and angry.

- The images should *not* vary much in other features, such as the orientation of the face, lighting conditions and the age of the person in the photo.

Consider if you want to reduce the dimensionality of the images using cropping, down-sampling and gray-scaling.

What data base did you use and why? What facial feature did you select for Experiment 1. Describe in detail how you selected the images based on the feature of interest and on features not of interest.

2. Conduct Experiment 1

   - Write a script that presents images and accept ratings from an observer. Importantly, the program must save the ratings and information relating each rating to image that was rated.
   - Collect the data. Start the script and have the test person rate the stimuli. Your group members can be your test participants. You can also ask students from other groups, friends or family to participate.

   Describe the experiment in detail. How many test participants did you include? How many images did they rate? Did they all rate the same images?

3. Pool the (normalised) rating data

   - **Show a histogram of the ratings for each participant**
   - **Do all your participants use the full range of the rating scale?** In case they do not then you can normalise the ratings for each participant using min-max normalisation. **Did you min-max normalise and why (not)?**

4. PCA and dimension reduction

   - Subtract the average image from all the images. Do not standardise (divide by the standard deviation).
   - Run PCA on the images (with the average image subtracted). The scores (i.e., the representation of the images in PCA space) will be the predictors in the linear encoding model.
   - Visually present the first few PCA components as images. Visualise a component by reconstructing the images that has the max/min score for that component. Compare these to the mean image. You can also create synthetic image using scores that lie somewhere between the max/min score. In this way you can visualise the effect that one principal component has on the images. **Show figures of your visualisation. What image features do the PCs represent?**
   - **Show a bar plot of the variance explained for all the PCs. Use it to select which PCs you want to include in your model. You should reduce the dimensionality significantly. How did you decide?**

5. Build a linear regression model that predicts the (normalised) ratings from the selected PC scores.

   - Use forward selection to select the relevant PCs for the model. You can use a routine from a toolbox or you can setup a homemade stepwise selection with cross validation. Some examples of routines from toolbox are
     - sklearn.feature_selection.SequentialFeatureSelector (Python)
     - sequentialfs (Matlab)

Explain how your feature selection method works (your own or a routine from a toolbox). Visualise the PCs that were selected in the same way as you did above. Do the selected PCs represent facial features relevant in the rating task?

- Finally you should fit the linear model using only the PCs that you selected using forward selection. **What are the parameter values for this fit?**

6. Generate synthetic images from the model

   - Use your model fit to generate synthetic face images with a given rating using the method described in Section 2.1. You should generate one synthetic image for each possible rating (e.g. 7 images if your rating scale was 1-7). Also generate synthetic images that extends the rating scale in your data set (at least one image in each end of the range). **Explain in detail how you generated the synthetic faces from the model. Show the synthetic images. Discuss if the images appear as expected and, if not, what the reason could be.**

7. Conduct Experiment 2.

   - Modify the script you used to present images in Experiment 1, so that it can present all the synthetic images that you have created at least 20 times in random order. Use the modified script to run the experiment. You can use the same test participants or some other participants. **Describe the experiment in detail like you did for Experiment 1.**
   - **Make a table with the response counts or response proportions for each test participant.**

8. Analyse the data from Experiment 2.

   - Analyse your results using ROC curves. **Choose a baseline stimulus and fit and plot ROC curves for all the other stimuli with respect to the baseline stimulus. Also plot the data in the same plot. Do this for each test participant. Do the ROC curves appear like you would expect? Explain your reasoning.**
   - Analyse your results using psychometric functions. **Fit and plot psychometric functions for each response criterion. Also plot the data in the same plot. Do the psychometric functions appear like you would expect? Explain your reasoning.**

CHAPTER 3

Bayesian models of perception

Perception often seems effortless. When we recognise the face of a friend it seems to happen easily, quickly and automatically. This can lure us into believing that perception is a simple process and that our perceptual experience is merely a reflection of the world as it is. This belief is far from the truth. Yet, an anecdote from the early years of AI research tells that this false belief lured pioneering researchers. In 1966 Seymour Papert set up a summer project for students at MIT. The goal of the project was to detect and identify objects from images. It took several decades for the field to reach useful solutions and the problem has still not been completely solved, so Papert seem to have underestimated the challenge.

From the earliest studies of human perception, it has become clear that perception is an extremely complex process. From the most recent studies we learn that we still do not fully understand it. What we do know is that perception depends not only on sensory information but also on the state of the observer, an understanding of the world around us and on perceptual biases.

Perception can be thought of as the brain's solution to an inverse problem: estimate the state of the world given sensory information. This type of problem is quite similar to the scientific endeavour of finding the state of the world given measurements. In both perception and in science such inverse problems are typically ill-posed or under-determined. This means that there is not enough data to find a unique solution. Yet, perception and science do solve this type of problems by adding information. This information comes in the form of models of how the world works and prior knowledge of the state of the world.

We can think of inverse problems in terms of Bayes' rule. Let, d, denote data that could be sensory data sensed by the eye, or it could be measurements in a scientific experiment. Let, w, denote the state of the world. Then Bayes' rule can be written as

$$P(w \mid d) = \frac{P(d \mid w)P(w)}{P(d)}$$

The problem of perception is to determine the posterior probability of the state of world given sensory information, or data, d. The posterior probability is proportional to the likelihood, $P(d \mid w)$, and the prior probability, $P(w)$.

Typically, we are not interested in the state of the entire world but only a small part of the world. For example, we might be interested in whether there is a tiger hiding in the bushes. In this example, the sensory data could be the sound of a low rumble from the bushes. The likelihood would then the probability of a low rumble given that there is a tiger hiding in the bushes. Tigers sometimes rumble and sometimes do not. In order to determine the likelihood, we need to estimate the probability that they do, or, even better, that this particular albeit hypothetical tiger rumbles. The likelihood thus contains a probabilistic model of how the world works, or more precisely, how it generates sensory data.

The prior probability is the probability of the state of the world prior to any sensory input as opposed to the posterior probability, which can be determined only after sensory input. It could be the probability of a tiger hiding in the bushes in general and would be based on previous experience with tigers and bushes.

The denominator in Bayes' rule, $P(d)$ is sometimes referred to as the evidence. For some purposes, this term can be ignored. To see why, let us rephrase the problem of determining whether there is a tiger hiding in the bushes as

$$P(tiger \mid rumble) > P(no\ tiger \mid rumble)$$

This inequality implements the *maximum a posteriori* (MAP) decision rule: we decide that there is a tiger if this is the more probable outcome. If we insert Bayes' rule on both sides, we get

$$\frac{P(rumble \mid tiger)P(tiger)}{P(rumble)} > \frac{P(rumble \mid no\ tiger)P(no\ tiger)}{P(rumble)}$$

Note that the evidence, $P(rumble)$, is the only term that does not depend on the state of the world. It is therefore the same on both sides of the MAP decision rule inequality. We can thus multiply the by $P(rumble)$ on both sides and remove it from the inequality. In effect, we may ignore the evidence term when making decisions by comparing the posterior probabilities of various outcomes.

Another way to understand the role of the evidence is to rewrite it using the law of total probability so that Bayes' rule can be written in the form

$$P(w \mid d) = \frac{P(d \mid w)P(w)}{\sum_w P(d \mid w)P(w)}$$

In this form, it is clearer that the evidence in the denominator is merely a normalisation term ensuring that $\sum_w P(w \mid d) = 1$, which must hold for any probability distributions in general.

3.1 Bayesian Signal Detection Theory

We can analyse the signal detection problem in terms of Bayesian observer theory. Here, the data, d, available is the stimulus encoded onto the internal representation, x. In that case, the MAP decision rule states that the observer should respond 'yes' if it is more probable that the signal, s, is present given an internal representation value, x.

$$P(s \mid x) > P(s_0 \mid x) \tag{3.1}$$

By inserting Bayes' rule and multiplying by the denominator, $P(x)$, on both sides we get

$$P(x \mid s)\,P(s) > P(x \mid s_0)\,P(s_0) \tag{3.2}$$

We then insert the Gaussian model for the likelihood and note that $P(s_0) = 1 - P(s)$

$$\phi\left(\frac{x-\mu}{\sigma}\right) P(s) > \phi(x)\,(1 - P(s)) \tag{3.3}$$

We first analyse the equal variance signal detection model. Recall that for this model, $\mu = d'$ and $\sigma = 1$. Inserting the expression for the Gaussian probability density function, taking the logarithm on both sides of the equation and rearranging gives us

$$\log\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-d')^2\right)\right) - \log\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\right) > \log\left(\frac{1-P(s)}{P(s)}\right) \tag{3.4}$$

which can be reduced to

$$-\frac{1}{2}(x-d')^2 + \frac{1}{2}x^2 > \log\left(\frac{(1-P(s))}{P(s)}\right) \tag{3.5}$$

Solving for x we find that the MAP rule corresponds to the decision rule in signal detection theory, $x > c$

$$x > \frac{1}{d'} \log\left(\frac{(1-P(s))}{P(s)}\right) + \frac{d'}{2} = c \tag{3.6}$$

Recasting the criterion, c, in terms of the prior probability, $P(s)$, provides us with a more interpretable measure of response bias. Whereas a value of c should be seen in relation to the optimal value of $c = \frac{d'}{2}$ we can immediately interpret the prior probability, $P(s)$, as the observer's *a priori* knowledge of the probability that a signal will be present. Note that for the unbiased observer, $P(s) = 1 - P(s)$, so that Equation 3.6, reduces to $c = \frac{d'}{2}$, which is the criterion value for the unbiased observer that we previously found.

We then proceed to analyse the unequal variance model. Again, we insert the expression for the Gaussian probability density function, take the logarithm on both sides of Equation 3.3.

$$\log\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) - \log\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\right) > \log\left(\frac{1-P(s)}{P(s)}\right) \tag{3.7}$$

This expression can be reduced to

$$\frac{1}{2}\left(1-\frac{1}{\sigma^2}\right)x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \log(\sigma) > \log\left(\frac{(1-P(s))}{P(s)}\right) \tag{3.8}$$

The expression in Equation 3.8 may seem a bit daunting but all we will focus on here is that it is quadratic in x. Therefore the solution can not be expressed in terms of a single criterion as in Equation 3.6. We can develop a better understanding of why that is by investigating the case where $\sigma > 1$. In this case the coefficient, $\frac{1}{2}\left(1-\frac{1}{\sigma^2}\right)$, of the squared term, x^2, is positive and the solution will be in the form of *two* criteria $x < c_{lo} \wedge x > c_{hi}$. The greater criterion, c_{hi}, is in perfect analogy to the criterion, c, that we have already encountered in signal detection theory: the observer will perceive the stimulus, s, when x falls above the criterion, c. But the smaller criterion, c_{lo}, may seem counter intuitive relative to signal detection theory. Why would the observer perceive the signal, s, for *low* values of x?

We can see why this is so in Figure 1.2, which illustrates the unequal variance model. Going back to Equation 3.3, we recall that an unbiased observer will perceive the stimulus, s, if $P(x \mid s) > P(x \mid s_0)$. Inspecting the Gaussian likelihoods drawn in Figure 1.2, we see that, unsurprisingly, $P(x \mid s) > P(x \mid s_0)$ for large values of x because the mean, μ, of $P(x \mid s)$ is greater than the zero mean of $P(x \mid s_0)$. However, we can also just barely see that for the smallest values of x, $P(x \mid s)$ is also greater than $P(x \mid s_0)$ because $P(x \mid s)$ is so wide that its probability density spills over to small values of x.

We have now learned that the unequal variance model is, in fact, not an optimal observer model even though the equal variance model is. This leaves us with the question of which model to use. In theory, we could, of course, test both models to see how they predict actual data. In practice, the unequal variance signal detection model and the Bayesian observer model, will most often lead to very similar predictions because the probability mass that lies below c_{lo} is very small. It will therefore require large amounts of data to distinguish between the two models. Worse is that the observer should behave exactly according to one of the two models while we collect this large amount of data. If the observer lapses, during such long experiments, it will still be difficult to determine which model describe behavior better.

3.2 Bayesian multisensory integration of continuous responses

Intuitively, we perceive our sensory modalities to be independent: We are, for example, never in doubt of whether we heard or saw something. In agreement with this, we experience that the senses complement each other so that, the sight of the colour of our food complement the gustatory experience. However, in many cases, the senses do not only provide independent, complementary information but also contain redundant information. Take, for example, a cat hunting for mice. The cat might both see and hear a mouse, so the location of the mouse will be encoded in the auditory modality as x_a and in the visual

modality as x_v. In this case, the information contained in the sensory modalities is not independent. Both sensory modalities contain information that is important to the task of localising the mouse, but in order to catch the mouse, the cat needs a single integrated estimate of the location, S, of the mouse.

3.2.1 The strong fusion observer model for continuous responses

We are now searching for a model for integrating auditory and visual internal representation values, $\boldsymbol{x_a}$ and $\boldsymbol{x_v}$, using Bayes' rule. First, let us consider the case where the observer (the cat) has access to only one sensory modality. If the observer can only hear the mouse, then it has access only to $\boldsymbol{x_a}$. It seems reasonable to assume that the most probable location, i.e. the *maximum a posteriori* (MAP) estimate, $\hat{\boldsymbol{S}}_{\boldsymbol{MAP}}$, of the true location of the sound of the mouse, $\boldsymbol{S_a}$, should correspond to the internal representation value, $\boldsymbol{x_a}$. Even so, we will show this more formally as it will help us in establishing a useful mathematical framework. Using Bayes' rule, the observer model can estimate the posterior probability of an estimate $\hat{S}$, of the location, S.

$$P(\hat{S} \mid x_a) = \frac{P(x_a \mid \hat{S})P(\hat{S})}{P(x_a)} \tag{3.9}$$

For simplicity, we will assume that the observer has no prior information on the location, S, so that the prior, $P(\hat{S})$, is constant across all locations. In this case, the prior probability does not influence the posterior. We can show this by expanding the denominator

$$P(\hat{S} \mid x_a) = \frac{P(x_a \mid \hat{S})P(\hat{S})}{P(x_a)} = \frac{P(x_a \mid \hat{S})P(\hat{S})}{\int_{\hat{S}} P(x_a \mid \hat{S})P(\hat{S})} = \frac{P(x_a \mid \hat{S})}{\int_{\hat{S}} P(x_a \mid \hat{S})} \tag{3.10}$$

In this case, the MAP estimate is equal to the maximum likelihood estimate, which is why the strong fusion model is often referred to as the maximum likelihood estimation (MLE) model.

As in signal detection theory, the observer model assumes that the internal representation value, $\boldsymbol{x_a}$, vary randomly due to Gaussian sensory noise added in the encoding process, so that it centered around the true location of the sound, $\boldsymbol{S_a}$. The, observer can thus calculate the likelihood as

$$\boldsymbol{P(x_a \mid \hat{S}) = f(x_a \mid \hat{S}, \sigma_a^2)} \tag{3.11}$$

Inserting Equation 3.11 into Bayes' rule in the form of Equation 3.10 gives us

$$P(\hat{S} \mid x_a) = \frac{f(x_a \mid \hat{S}, \sigma_a^2)}{\int_{\hat{S}} f(x_a \mid \hat{S}, \sigma_a^2)} = f(\hat{S} \mid x_a, \sigma_a^2) \tag{3.12}$$

Note that the normalisation of the Gaussian likelihood, $f(x_a \mid \hat{S}, \sigma_a^2)$, with respect to the estimate, $\hat{S}$, of the location, S, results in a posterior probability density, $P(\hat{S} \mid x_a)$, which is centered around the internal representation value, x_a. Now, since the maximum of a Gaussian distribution is its mean, the MAP solution, $\hat{\boldsymbol{S}}_{\boldsymbol{MAP}}$, to Equation 3.12 is the mean of the posterior so that

$$\boldsymbol{\hat{S}_{MAP} = x_a} \tag{3.13}$$

In order to fit the simple unisensory observer model to observable data, we will assume that the observer responds according to its MAP solution, $\hat{\boldsymbol{S}}_{\boldsymbol{MAP}} = \boldsymbol{x_a}$. In the case of a cat chasing a mouse, the response could be a pounce or an orienting response to a sound. According to the observer model, the distribution of the internal representation value, $\boldsymbol{x_a}$, will centered around the true location of the sound, $\boldsymbol{S_a}$.

$$\boldsymbol{P(\hat{S}_{MAP} \mid S_a, \sigma_a^2) = f(x_a \mid S_a, \sigma_a^2)} \tag{3.14}$$

In this case the only free parameter of the model is σ_a. However, the observer model may not hold. Therefore, the true location of the sound, S_a, is typically replaced by a free parameter for the mean, μ_a, of the distribution of the observer's MAP solution.

$$P(\hat{S}_{MAP} \mid S_a, \sigma_a^2) = f(x_a \mid \mu_a, \sigma_a^2) \tag{3.15}$$

Trivially, the case in which the observer can only see the stimulus, but not hear it, is described by replacing μ_a with μ_v and σ_a with σ_v in Equations 3.9-3.14.

We have now established a framework, which will help us analyse the case of multisensory stimuli for which the observer has access to a bivariate internal representation, (x_a, x_v), but still needs a single integrated estimate, $\hat{S}$, of the location of the mouse. This is the basis of strong fusion: It only makes sense to fuse the two sensory modalities into a single estimate if they both convey information about the same location. Still assuming a uninformative prior, as in Equation 3.10, we can write Bayes' rule as

$$P(\hat{S} \mid x_a, x_v) = \frac{P(x_a, x_v \mid \hat{S})}{P(x_a, x_v)} = \frac{P(x_a, x_v \mid \hat{S})}{\int_{\hat{S}} P(x_a, x_v \mid \hat{S})} \tag{3.16}$$

Since the sensory noise is added in the encoding process, the observer model assumes that the noise added to the auditory internal representation, x_a, is independent from the noise added to the visual internal representation. This means that the internal representation values are *conditionally* independent on S, so that we can write the likelihood for the multisensory case as

$$P(x_a, x_v \mid \hat{S}) = P(x_a \mid \hat{S})P(x_v \mid \hat{S}) = f(x_a \mid \hat{S}, \sigma_a^2)f(x_v \mid \hat{S}, \sigma_v^2) \tag{3.17}$$

Note that conditional independence, as described in Equation 3.17, does not imply unconditional independence, $P(x_a, x_v) = P(x_a)P(x_v)$. This would be a poor assumption because the internal representation values, x_a and x_v will, in general, be dependent as they are both influenced by location of co-occurring changes in sound and light. Conditional independence means that the internal representation values, x_a and x_v, are independent across multiple occurrences of the exact same stimulus because their values would, in that case, vary only due to the noise added in the encoding process.

Equation 3.17 describes the likelihood for audiovisual stimuli as a product of two Gaussian probability density distributions. In general, the product of two Gaussian distributions with random variables, x_a and x_v is proportional to a Gaussian distribution of a random variable, x_{av}, that is equal to a weighted average of x_a and x_v

$$x_{av} = w_a x_a + (1 - w_a)x_v \tag{3.18}$$

where the weight, w_a, is given by

$$w_a = \frac{\sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{3.19}$$

so that

$$f(x_a \mid \mu_a, \sigma_a^2)f(x_v \mid \mu_v, \sigma_v^2) = Af(x_{av} \mid \mu_{av}, \sigma_{av}^2) \tag{3.20}$$

where A is a constant that will be of little importance.

Since x_{av} is a weighted mean of x_a and x_v, its mean, μ_{av}, is a weighted average of the mean, μ_a, of x_a, and the mean μ_v, of x_v

$$\mu_{av} = w_a \mu_a + (1 - w_a)\mu_v \tag{3.21}$$

and the variance, σ_{av}^2, is given by

$$\sigma_{av}^2 = w_a^2\sigma_a^2 + (1 - w_a)^2\sigma_v^2 = \frac{\sigma_a^2\sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{3.22}$$

For the strong fusion observer model in Equation 3.20, the means of the distributions of x_a and x_v, are both assumed to be equal to one and the same estimate, $\hat{S}$. A weighted average of the two means is therefore also equal to $\hat{S}$. Inserting this into the observer model gives us

$$P(\hat{S} \mid x_a, x_v) = \frac{Af(x_{av} \mid \hat{S}, \sigma_{av}^2)}{\int_{\hat{S}} Af(x_{av} \mid \hat{S}, \sigma_{av}^2)} = \frac{f(x_{av} \mid \hat{S}, \sigma_{av}^2)}{\int_{\hat{S}} f(x_{av} \mid \hat{S}, \sigma_{av}^2)} = f(\hat{S} \mid x_{av}, \sigma_{av}^2) \tag{3.23}$$

As in Equation 3.13 we find that the MAP estimate, $\hat{S}_{MAP}$, of the location S is

$$\hat{S}_{MAP} = x_{av} = w_a x_a + (1 - w_a) x_v \tag{3.24}$$

3.2.2 Properties of the strong fusion observer model

Here we will pause and reflect on the properties of the strong fusion model. The first thing that we will note is that its implementation is actually quite simple: The auditory and visual stimuli are encoded as scalars, x_a and x_v. When the observer has access to both, the combined estimate is simply a weighted sum. The fairly complex probability theoretical considerations above are not necessary for the implementation. They only help for us analyse the model.

In order to better understand the weight, w_a, we rephrase it in terms of precision, $r = \sigma^{-2}$, by dividing the numerator and denominator on the right hand side of Equation 3.19 by $\sigma_a^2 \sigma_v^2$. This gives us

$$w_a = \frac{r_a}{r_a + r_v} \qquad w_v = (1 - w_a) = \frac{r_v}{r_a + r_v} \tag{3.25}$$

Equation 3.25 shows that the strong fusion observer weighs the sensory modalities by their precision. Somewhat confusingly, this is called the *information reliability* principle, because the term reliability is sometimes used to mean precision. Regardless of the name, the principle of weighing information according to its reliability, or precision, makes intuitive sense: If you were to make a decision based on, say, two eye-witness accounts, you would weigh the account of the more reliable, or precise, witness higher than the account of the less reliable witness.

Calculating the precision, r_{av} of the internal representation value x_{av} reveals another interesting property of the strong fusion model. We can do this by rearranging equation 3.22

$$r_{av} = \frac{1}{\sigma_{av}^2} = \frac{\sigma_a^2 + \sigma_v^2}{\sigma_a^2 \sigma_v^2} = r_a + r_v \tag{3.26}$$

Equation 3.26 reveals that the precision of the observer given audiovisual stimuli is the sum of the precision given the auditory stimulus and precision given the visual stimulus. A strong fusion observer will thus always be more precise when integrating the multisensory internal representations, x_a and x_v.

We can even show that the strong fusion observer weight, w_a, is the Bayes' optimal weight in terms of maximising the precision, r_{av}, or, equivalently, minimising the variance, σ_{av}^2, by expressing the variance as a function of the weight as in Equation 3.22 and finding the minimum by differentiation

$$\frac{\partial \sigma_{av}^2}{\partial w} = \frac{\partial w^2 \sigma_a^2 + (1 - w)^2 \sigma_v^2}{\partial w} = \frac{\partial (\sigma_a^2 + \sigma_v^2) w^2 - 2w\sigma_v^2 + \sigma_v^2}{\partial w} = 2(\sigma_a^2 + \sigma_v^2) w - 2\sigma_v^2 = 0 \tag{3.27}$$

The solution to Equation 3.27, is $w = w_a = \frac{\sigma_v^2}{\sigma_a^2 + \sigma_v^2}$, which is the minimum $\arg\min_w(\sigma_{av}^2) = w_a$, which shows that the choice of weight in the strong fusion observer model is optimal for maximising the precision, r_{av}.

Finally, we will look at the case where the true location of the sound, S_a, is different from the true location of the visual stimulus, S_v. Small differences could occur in the natural world if, say, a cat sees the tail of the mouse while the sound originates from the snout of the mouse. The strong fusion estimate, $\hat{S}$, of the location would, on average, lie somewhere between the snout and the tail and in that case a pounce may well prove successful. Much larger differences occur mostly in artificial settings where the location of the sound, S_a, and the location of the visual stimulus, S_v, are designed to be different. A striking example of this is the ventriloquist illusion. When we observe a ventriloquist, we often perceive

the sound as appearing from a dummy even though it does, of course, originate from the ventriloquist. A more common version of this illusion is experienced when watching a movie of a conversation with a monaural sound source. Although the sound of the voices originate from one location, observers tend to perceive that the voice of each actor originates from the location of the image of that actor. In order to test how the strong fusion accounts for the ventriloquist illusion, we may measure an observer's estimate of the location of the sound, the image and the audiovisual combination of the two. It has been found that the strong fusion model is able to account for this type of data. Since vision has a higher spatial resolution than hearing in humans, estimates of the location of the visual stimulus are typically much more precise than estimates of the location of the sound. Therefore, according to Equation 3.25 observers will weigh the location of the visual stimulus much higher when estimating the location of the audiovisual stimulus. This, according to the strong fusion model, is the basis of the ventriloquist illusion.

3.3 Bayesian multisensory integration of discrete responses

The brain integrates information across the sensory modalities for many types of stimuli. In Section 3.2 we described models of audiovisual integration for stimuli, such as the location of a mouse, that are naturally perceived on a continuum. In this section we will describe models of audiovisual integration that are naturally perceived in discrete categories.

Audiovisual speech perception serves as a good example of a stimulus that is integrated across the senses and which is perceived in categories. Speech perception is enhanced when we can see the face of the talker compared to when we can only hear the voice. This is called the *enchancement effect*. Although there could be several reasons for this effect, it indicates that perceptual precision is increased when we perceive speech from two modalities rather than just one. This is similar to the effect of audiovisual integration for spatial localisation discussed in Section 3.2. As in the ventriloquist illusion for spatial localisation, illusory percepts occur also in speech perception when the two senses convey incongruent information. The most commonly known example of this is the McGurk fusion illusion where a voice saying /ba/ is dubbed on a video of a face articulating /ga/, in which case observers typically hear /da/. In this case, auditory and visual information seem to be fused into a percept different from that contained in either modality much as in the ventriloquist illusion.

Unlike spatial location, speech is perceived in discrete categories that are called phonemes. Phonemes are often thought of as speech sounds such as vocals and consonants. Defining phonemes in terms of the acoustic properties is, however, very difficult. One reason for this is that a phoneme, such as the 'p' in the word 'sport', sounds slightly differently every time you say it. Worse is that the sound depends on many characteristics of the speaker such as the speaker's gender, age and level of inebriation. Even worse, the the 'p' in the word 'sport' is actually articulated more like the 'b' in the word 'bored' than the 'p' in the word 'port'. This is an example of co-articulation meaning that the sound of phonemes depends on the phonetic context. The acoustic variability within phonetic categories is thus large compared to the variability across phonetic categories, which is why it is difficult to define phonemes acoustically.

Phonemes are better characterised by the way they are articulated. A full description of all the articulatory features is beyond the scope of this text and we will focus on just one, the *place of articulation*, since it is particularly important for audiovisual speech perception. The consonant /b/ is a bilabial phoneme as it is articulated by a brief closure of the lips. A /b/ is thus articulated at the very front of the mouth. The consonants /d/ is articulated further back, at the alveolar ridge, right behind the teeth. It is therefore an alveolar consonant. The consonant /g/ is called a velar consonants and is articulated at the very back of the mouth. Obviously, the exact place of articulation may vary depending on the same factors that influence the acoustic characteristics of phonemes as we do not close the lips exactly the same way every time we articulate a /b/. Yet, this variability in the place of articulation within phonetic categories is small compared to the variability across phonetic categories. We do not, for example, ever close our lips when articulating a /d/. This is why phonemes are better characterised by articulatory features.

The place of articulation is important in audiovisual integration of speech. The *manner-place* hypothesis has been has found much support in the study of audiovisual integration of speech. Its claim is that it is information about the place of articulation that is integrated across modalities because it can be perceived visually, as you can see if someone is articulating a /b/ or a /g/, and auditorily, as you can also hear the difference. Other phonetic features can not be seen and are therefore not integrated audiovisually.

3.3.1 The early strong fusion model of audiovisual integration of speech

Perception of phonemes in discrete categories may rely on a continuous internal representation of the place of articulation much as the discrete responses, 'yes' and 'no', relies on a continuous internal representation of signal intensity in signal detection theory. As we have already described models of audiovisual integration for continuous responses in Section 3.2 it is natural to combine these models with signal detection theory, to build models of audiovisual integration of speech. We will refer to these models as *early* models of audiovisual integration of speech because they posit that integration occurs based on the continuous internal representation before decoding the internal representation into discrete response probabilities.

As in Section 3.2, the early strong fusion model posits that the observers estimate, $\hat{S}_{MAP}$, of the place of articulation, S is distributed as a Gaussian distribution so that

$$P(\hat{S}_{MAP} \mid S_a) = f(x_a \mid \mu_a, \sigma_a^2) \tag{3.28}$$

$$P(\hat{S}_{MAP} \mid S_v) = f(x_v \mid \mu_v, \sigma_v^2) \tag{3.29}$$

$$P(\hat{S}_{MAP} \mid S_{av}) = f(x_{av} \mid \mu_{av}, \sigma_{av}^2) \tag{3.30}$$

where expressions for μ_{av}, w_a and σ_{av} are given by Equations 3.21, 3.19 and 3.22 respectively.

Just as in signal detection theory, criteria values, c, are introduced to decode the continuous values for $\hat{S}_{MAP}$ onto discrete response categories. We will, arbitrarily, choose that a bilabial place of articulation (towards the front of the mouth) corresponds to smaller values of x, so that the probability, $P(r_b)$, of a b-response, r_b, is given by

$$P(r_b) = P(\hat{S}_{MAP} < c_{bd}) = \int_{-\infty}^{c_{bd}} P(x \mid S) = \Phi(c_{bd}, \mu, \sigma) \tag{3.31}$$

where c_{bd} denotes the response criteria that separates b- and d-responses. Note that the stimulus, S, should be replaced by S_a, S_v and S_av for auditory, visual and audiovisual stimuli respectively.

Likewise, the probability, $P(r_d)$, for a d-response, r_b, is given by

$$P(r_d) = P(c_{bd} < \hat{S}_{MAP} < c_{dg}) = \int_{c_{bd}}^{c_{dg}} P(x \mid S) = \Phi(c_{dg}, \mu, \sigma) - \Phi(c_{bd}, \mu, \sigma) \tag{3.32}$$

where c_{dg} denotes the response criteria that separates d- and g-responses.

Finally, the probability, $P(r_g)$, for a g-response, r_g, is given by

$$P(r_g) = P(\hat{S}_{MAP} > c_{dg}) = \int_{c_{dg}}^{\infty} P(x \mid S) = 1 - \Phi(c_{dg}, \mu, \sigma) \tag{3.33}$$

In general, the means μ_a and μ_v, for each stimulus are free parameters of the early strong fusion model. Also the standard deviations, σ_a and σ_v, are free parameters, which typically are assumed not to depend on the specific stimulus. Finally, the response criteria, c_{bd} and c_{dg} are also free parameters. Depending on the experimental paradigm, the strong fusion model may be over-parameterised making it a highly flexible model that is likely to over-fit. This can, of course, be tested using cross-validation as over-fitting models tend to have a high validation error despite a low training error. In the next section we will describe how we sometimes can use observer theory to constrain the model so that it will be less likely to over-fit.

3.3.2 Applying the early strong fusion model of audiovisual integration

Andersen [And15] applied the early strong fusion model to categorical responses to auditory, visual and audiovisual speech stimuli. In a previous study, auditory speech had been generated using a speech synthesizer and videos of a face articulating speech had been generated using an animation. Five auditory and five visual stimuli had been generated so that they were perceptually evenly spaced on a continuum ranging from a clear /ba/ to a clear /da/. Additionally, 25 Audiovisual stimuli had been generated as the combination of the five auditory and five visual stimuli. Note that this continuum spans only the bilabial and alveolar (front to mid) range of the full range of place of articulation. Accordingly, the participants in the experiments could only respond with b- or d-responses. This simplifies the early strong fusion model, as the only dependent variable is the number of d-responses.

With these simplifications, the probability, $P_a(r_d)$, of a d-response given an auditory stimulus is given by

$$P_a(r_d) = P(\tilde{S}_{MAP} \geq c \mid S_a) = \int_c^\infty P(\tilde{S}_{MAP} \mid S_a) = 1 - F(c_a \mid \mu_a, \sigma_a) = \Phi\left(\frac{\mu_a - c_a}{\sigma_a}\right) \tag{3.34}$$

where $F(c_a \mid \mu_a, \sigma_a)$ denotes the cumulative Gaussian probability function with mean, μ_a, and standard deviation, σ_a.

Note that the last step in Equation 3.34 is based on the mathematical identity

$$1 - F(x \mid \mu, \sigma) = F(-x \mid \mu, \sigma) = \Phi\left(\frac{-x - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - x}{\sigma}\right) \tag{3.35}$$

Note also that since the auditory stimuli were perceptually evenly spaced along the /b/-/d/ continuum, we can recognise the last step in Equation 3.34 as the psychometric function from Equation 1.19 if we replace μ by the stimulus intensity, I_s. This is in line with the observer model illustrated in Figure 1.4 where each point on the psychometric function corresponds to the mean of the Gaussian distribution of internal representation values for that perceived stimulus intensity. Since the intervals of the synthetic continuum was designed to match the perceived intensity, we can replace $\mu = I_s$ by $1, \ldots, 5$ for the five auditory stimuli. Similarly the response probability, $P_v(r_d)$, of a d-response given a visual stimulus is given by replacing changing the subscript a by v.

Now that we have described the psychometric functions for the auditory and visual stimuli, we are ready to derive the response probabilities for audiovisual stimuli. The approach we will use is first to use the strong fusion model described in Equations 3.21-3.22 to derive expressions for the mean and standard deviation for audiovisual stimuli and then use Equation 3.34 to derive an expression for the response probabilities. We must, however, first align the internal representations. Recall that we have replaced the mean by the perceived signal intensity so that $\mu = I_s = 1, \ldots, 5$ for the five auditory and the five visual stimuli so that Equation 3.34 represents a psychometric function. We did this for both auditory and visual stimuli. However, Unless $c_a = c_v$, the two psychometric functions are misaligned, which will also misalign the auditory and visual internal representations. The means of the distributions for auditory and visual stimuli are, however, defined with respect to the response criterion, c. Introducing $\tilde{\mu} = \mu - c = I_s - c$ will fix this problem. This is illustrated in Figure 3.1.

To summarise, applied to the experimental paradigm used by Andersen [And15], the early strong fusion model has four free parameters: c_a, c_v, σ_a and σ_v. From these we can directly calculate the response probabilities for auditory and visual stimuli according to Equation 3.34 by setting $\mu = I_s = 1, \ldots, 5$. In order to calculate the response probabilities for audiovisual stimuli we must first calculate $\tilde{\mu}_a = \mu_a - c_a$ and $\tilde{\mu}_v = \mu_v - c_v$. We can then calculate $\tilde{\mu}_{av} = w_a\tilde{\mu}_a + (1 - w_a)\tilde{\mu}_v$ according to Equations 3.21. Finally, we can calculate $P_{av}(r_d) = \Phi\left(\frac{\tilde{\mu}_{av}}{\sigma_{av}}\right)$ according to Equation 3.34.

With this parameterization, Andersen [And15] could show that the early strong fusion model provided a good fit to the experimental data. Importantly, the use of the observer model to limit the number of free parameters ensured that validation error in a leave-one-stimulus cross validation test was low.

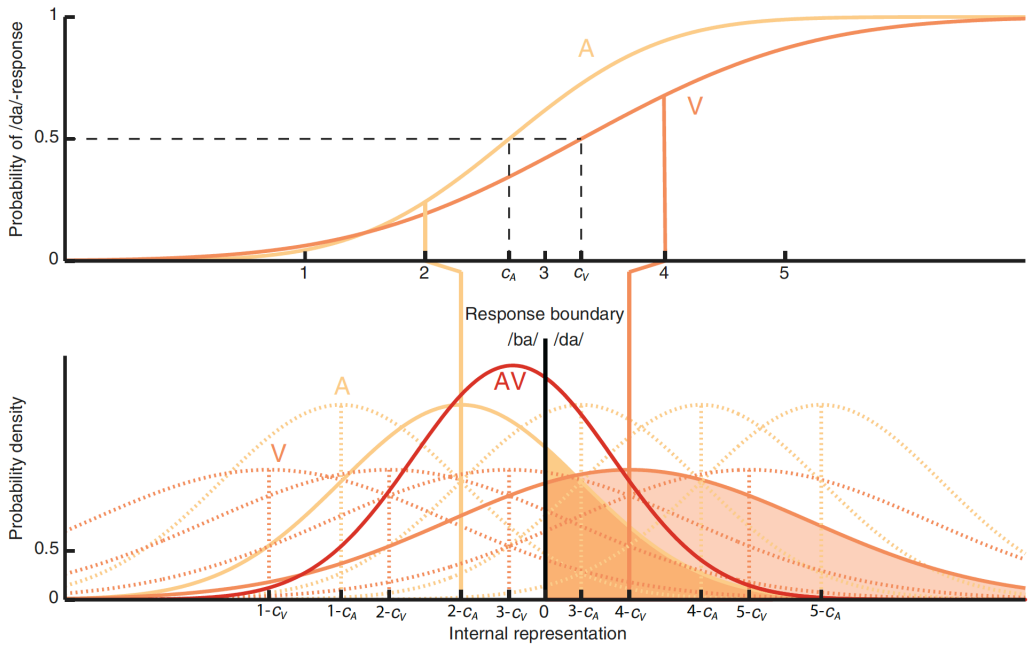


Figure 3.1: The psychometric function and observer model for audiovisual speech stimuli. The top panel depicts psychometric functions for auditory and visual stimuli. Note that they have different thresholds, c_a and c_v. The lower panel depicts the corresponding internal representation. Substituting the means, μ_a and μ_v, by $\tilde{\mu}_a = I_s - c_a$ and $\tilde{\mu}_v = I_s - c_v$ where $I_s = 1, \ldots, 5$ denotes the perceived stimulus intensity aligns the internal representations to a common threshold arbitrarily set to zero. Figure from Andersen [And15]..

3.3.3 The strong fusion probability matching model

The strong fusion probability matching model shares one key aspect with the other strong fusion models: since the observer model assumes one single underlying cause for the auditory and visual internal representation values, x_a and x_v, it assumes that they are independent conditioned on the stimulus, so that, as in Equation 3.17,

$$P(x_a, x_v \mid \hat{S}) = P(x_a \mid \hat{S}) P(x_v \mid \hat{S}) \tag{3.36}$$

The strong fusion probability matching model does not, however, share the assumption, that the internal representation values, x, are normally distributed scalars in Equation 3.11. The decoding process of mapping the internal representation to discrete responses is thus not described by the model. Instead it writes the *discrete* posterior directly as $P(\hat{S}_i \mid x)$ where $\hat{S}_i$ indicates that the stimulus is estimated to belong to the i^{th} response category. Without any assumptions on the distribution or dimensionality of x this expression for the posterior cannot be simplified.

Even without simpler expressions for the unisensory posterior probability of the stimulus estimate, $\hat{S}$, based on either the auditory or the visual internal representations we can still derive an expression for the multisensory posterior probability of the stimulus estimate, $\hat{S}$, based on both the auditory or the visual internal representations beginning with the conditional independence assumption

$$P(\hat{S}_i \mid x_a, x_v) = \frac{P(x_a, x_v \mid \hat{S}_i)}{P(x_a, x_v)} = \frac{P(x_a, x_v \mid \hat{S}_i)}{\sum_j^{N_r} P(x_a, x_v \mid \hat{S}_j)} = \frac{P(x_a \mid \hat{S}_i) P(x_v \mid \hat{S}_i)}{\sum_j^{N_r} P(x_a \mid \hat{S}_j) P(x_v \mid \hat{S}_j)} \tag{3.37}$$

Then we 'flip' Bayes rule so that we get expressions for the unisensory likelihoods, $P(x_a \mid \hat{S}_i)$ and $P(x_v \mid \hat{S}_i)$,

$$P(x_a \mid \hat{S}_i) = \frac{P(\hat{S}_i \mid x_a)P(x_a)}{P(\hat{S}_i)} \qquad P(x_v \mid \hat{S}_i) = \frac{P(\hat{S}_i \mid x_v)P(x_v)}{P(\hat{S}_i)} \tag{3.38}$$

Inserting this into Equation 3.37

$$P(\hat{S}_i \mid x_a, x_v) = \frac{\frac{P(\hat{S}_i|x_a)P(x_a)}{P(\hat{S}_i)} \frac{P(\hat{S}_i|x_v)P(x_v)}{P(\hat{S}_i)}}{\sum_j^{N_r} \frac{P(\hat{S}_j|x_a)P(x_a)}{P(\hat{S}_j)} \frac{P(\hat{S}_j|x_v)P(x_v)}{P(\hat{S}_j)}} = \frac{P(\hat{S}_i \mid x_a)P(\hat{S}_i \mid x_v)}{\sum_j^{N_r} P(\hat{S}_j \mid x_a)P(\hat{S}_j \mid x_v)} \tag{3.39}$$

where we have used that the prior is uniform so that $P(\hat{S}_i) = P(\hat{S}_j)$ for all i, j.

We now have the posterior probability of $\hat{S}_i$ and only need a decision rule. A MAP decision rule would imply that the observer would consistently respond in response category for which the posterior is the greatest. This does not align with the variability in responses across repeated presentations of the same stimulus that is typically seen in experimental data. Instead, the model assumes a *probability matching rule*, in which the observer randomly chooses a response with the probability given by the posterior so that

$$P(r_i) = P(\hat{S}_i \mid x) \tag{3.40}$$

This decision rule is not optimal in the sense that it will minimise the error rate. Assume, for example, that we were flipping an unfair coin, where probability of heads up is 0.8 and the observer knows the probability of the outcome. A MAP observer would always choose heads and therefore be correct with a probability of 0.8. A probability matching observer would choose heads with a probability of 0.8 and tails with a probability of 0.2. This observer would be correct with a probability of $0.8 \cdot 0.8 + 0.2 \cdot 0.2 = 0.64 + 0.04 = 0.68$. Nevertheless this suboptimal behavior is often seen in humans. It may be an example of a cognitive bias called the *gambler's fallacy* where the observer tries to predict a random sequence, which is, in nature, impossible. It may, however, be optimal if we take the value of the discovery of new knowledge into account. For example, assume that we sample from a distribution with probabilities of three outcomes given by 0.4, 0.3 and 0.3. The MAP observer would always choose the outcome with a probability of 0.4. This is only optimal if we assume that the probabilities are constant. If the probabilities of the outcomes would change to 0.4, 0.5 and 0.1 the MAP observer would never learn about this change and his chance of a correct answer would remain constant at 0.4. The probability matching observer, however, would perhaps notice that the probabilities have changed, and change his choices accordingly leading to a better outcome.

It may appear a bit unclear how to parameterise the probability matching strong fusion model. Equation 3.39 provides us with the response probabilities (given the probability matching rule) for audiovisual stimuli but we have no parameterised expression for the response probabilities for auditory and visual stimuli. These response probabilities are only given by the posterior probabilities, $P(\hat{S}_j \mid x_a)$ and $P(\hat{S}_j \mid x_v)$. We therefore need to parameterise these. We can do this by the softmax function so that

$$P(\hat{S}_i \mid x_a) = \frac{\exp z_i}{\sum_{j=1}^{N_r} \exp z_j} \tag{3.41}$$

where N_r is the number of response probabilities and $z_{N_r} = 0$ to ensure that the model is not over-parameterised as we need only $N_r - 1$ free parameters to model N_r response probabilities under the constrain that $\sum_{j=1}^{N_r} P(\hat{S}_j \mid x_a) = 1$.

The strong fusion probability matching model was first phrased in terms of *fuzzy logic* truth values, which are mathematically very similar to probabilities in that they are real-valued scalars that lie in the interval $[0; 1]$. Therefore the strong fusion probability matching model is commonly know as the Fuzzy Logical Model of Perception (FLMP). Here, we have chosen another name that describes the model in the Bayesian framework described here.

3.3.4 Applying the strong fusion probability matching model audiovisual integration

In Section 3.3.2 we described how Andersen [And15] applied the early strong fusion model to a data set consisting of 5 auditory stimuli, 5 visual stimuli and the 25 corresponding audiovisual stimuli. Notably the auditory and visual stimuli was generated synthetically, by a speech synthesizer and an animated head, so that they were perceptually evenly spaced. Based on this experimental design, the early strong fusion model was applied to the data using only four free parameters: two for the psychometric function for auditory stimuli and two for the psychometric function for visual stimuli. The integration of signal detection theory, the psychometric function and the strong fusion model thus led to a parsimonious model design.

Andersen [And15] also applied the strong fusion probability matching model to the same data although they referred to the model as the Fuzzy Logical Model of Perception (FLMP) since the model had previously been applied to the same data under that name. Since this model, in its basic form, lacks an underlying observer model relating it to the psychometric function it needs one free parameter for the posterior distribution, $P(\hat{S}_i \mid x)$ for each of the five auditory and five visual stimuli totalling 10 free parameters. Andersen [And15] showed that, in this form, used in previous studies, the probability matching model is likely to be over-parameterised as its validation error in a leave-one-stimulus out cross-validation procedure was higher than the validation error for the corresponding early strong fusion model.

So far, the probability matching model and the early strong fusion model differs in two aspects. First, the early strong fusion model is based underlying observer model, so that it is more parsimonious than the probability matching model. Secondly, and perhaps more importantly, the two models differ in model of integration. Whereas the early strong fusion model posits that integration is based on a continuous internal representation, prior to categorisation, the probability matching model posits that integration is based on a discrete internal representation as specified in Equation 3.39. In order to isolate the effect of the integration mechanism, Andersen [And15] designed a new model in which the auditory and visual response probabilities were based a continuous internal representation, so that they could be modeled using the psychometric function as in Equation 3.34 while integration was based on the discrete representation as described in Equation 3.39. This model was named the late strong fusion, or MLE, model, since it posits that integration occurs after classification. The validation error of this model was comparable to that of the early strong fusion model. This indicates that the high validation error of the probability matching model may be due to its high number of free parameters and that, controlling for this, it is not possible to distinguish between models of early and late stages of integration based on the available data. Although this finding is negative, it still added new knowledge to our understanding of audiovisual integration of speech. Previous studies had concluded that the probability matching model accurately described the integration mechanism since it provided good fits (low training errors) to the data set. By designing more parsimonious models and using cross-validation to control for model flexibility, Andersen showed that this conclusion is not valid.

Bibliography

[And15] Tobias S. Andersen. "The early maximum likelihood estimation model of audiovisual integration in speech perception." In: *The Journal of the Acoustical Society of America* 137.5 (2015), pages 2884–2891. DOI: 10.1121/1.4916691. eprint: https://doi.org/10.1121/1.4916691. URL: https://doi.org/10.1121/1.4916691.

[Fag+13] Jens Fagertun et al. "3D gender recognition using cognitive modeling." English. In: *2013 International Workshop on Biometrics and Forensics (IWBF)*. 2013 International Workshop on Biometrics and Forensics (IWBF) ; Conference date: 04-04-2013 Through 05-04-2013. United States: IEEE, 2013. ISBN: 978-1-4673-4987-1. DOI: 10.1109/IWBF.2013.6547324. URL: http://www.img.lx.it.pt/iwbf2013/.

[FAP12] Jens Fagertun, Tobias Andersen, and Rasmus Reinhold Paulsen. "Gender Recognition Using Cognitive Modeling." English. In: *Computer Vision – ECCV 2012*. Lecture Notes in Computer Science. 12th European Conference on Computer Vision (ECCV 2012) ; Conference date: 07-10-2012 Through 13-10-2012. Springer, 2012, pages 300–308. ISBN: 978-3-642-33867-0. DOI: 10.1007/978-3-642-33868-7_30. URL: http://eccv2012.unifi.it/.

[HMS21] Tue Herlau, Morten Mørup, and Mikkel N. Schmidt. *Introduction to Machine Learning and Data Mining*. DTU lecture notes, 2021. URL: https://gitlab.compute.dtu.dk/tuhe/books.

$\sqrt{17}$
$\Omega \int \delta e^{i\pi} = -1$
{2.7182818284}
ϑφερτυθιοπσδφγηξκλ
χ^2
$\sum$