Solutions to the 4th exercise sheet

Attribution of points

You can obtain up to 20 points, plus a variable number of bonus points for the exercises labeled as ‘advanced’. This time points were attributed as follows:

<table>
<thead>
<tr>
<th>exercise</th>
<th>points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>3</td>
</tr>
<tr>
<td>1b</td>
<td>3</td>
</tr>
<tr>
<td>1c</td>
<td>2</td>
</tr>
<tr>
<td>1d</td>
<td>1</td>
</tr>
<tr>
<td>2a</td>
<td>2</td>
</tr>
<tr>
<td>2b</td>
<td>3</td>
</tr>
<tr>
<td>2c</td>
<td>3</td>
</tr>
<tr>
<td>2d</td>
<td>3</td>
</tr>
<tr>
<td>3a</td>
<td>+2</td>
</tr>
<tr>
<td>3b</td>
<td>+1</td>
</tr>
<tr>
<td>3c</td>
<td>+1</td>
</tr>
</tbody>
</table>

Exercise 1a: Covariance and correlation

When analyzing data, we often want to estimate the variance from a particular sample. For a data set \( v = (v_1, \ldots, v_N) \), this variance is defined as:

\[
\text{Var}(v) = \frac{1}{N} \sum_{i=1}^{N} (v_i - \mu)^2,
\]

where \( \mu \) is the mean of the probability distribution underlying the measured variable \( v \). Note that we need to know the real mean of this distribution to apply this formula.

If we don’t know \( \mu \), we have to estimate it from our sample, which introduces a certain bias. To correct for this bias, we must use the following formula when computing the variance:

\[
\text{Var}(v) = \frac{1}{N-1} \sum_{i=1}^{N} (v_i - \bar{v})^2,
\]

where \( \bar{v} \) is the sample mean of the values in our sample. The sample mean \( \bar{v} \) is probably different from \( \mu \). Note that for very large values of \( N \), this bias can be neglected.

In this exercise, we do not know the real mean \( \mu \) of the distribution. Hence, we should compute our variance using:

\[
\text{Var}(r_1) = \frac{1}{N-1} \sum_{i=1}^{N} (r_{1,i} - \bar{r}_1)^2 \quad (1)
\]

This did not become clear from the exercise sheet (which even asks for \( 1/N \)), so we also accepted solutions assuming the real mean to be known and equal to \( \bar{r}_1 \).

On the other hand, the norm of an \( N \)-dimensional vector \( x \) is defined as:

\[
\|x\| = \sqrt{\sum_{i=1}^{N} x_i^2},
\]

We have therefore:

\[
\frac{1}{N - 1} \|x\|^2 = \frac{1}{N - 1} \left( \sum_{i=1}^{N} x_i^2 \right)^2
\]

\[
= \frac{1}{N - 1} \sum_{i=1}^{N} x_i^2
\]

\[
= \frac{1}{N - 1} \sum_{i=1}^{N} (r_{1,i} - \bar{r}_1)^2
\]

\[
= \text{Var}(r_1),
\]

which is the variance specified in equation (1). An analogous derivation yields:

\[
\frac{1}{N - 1} \|y\|^2 = \text{Var}(r_2).
\]
As you probably remember, the cosine between two vectors \( \mathbf{x} \) and \( \mathbf{y} \) plays an important role in the definition of their inner product (or dot product):

\[
\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos \alpha,
\]

where \( \alpha \) is the angle between the two vectors. From exercise 1a we know that

\[
\|\mathbf{x}\| = \sqrt{(N-1) \cdot \text{Var}(r_1)},
\]

\[
\|\mathbf{y}\| = \sqrt{(N-1) \cdot \text{Var}(r_2)}.
\]

Furthermore, we know that the definition of the covariance between \( r_1 \) and \( r_2 \) is:

\[
\text{Cov}(r_1, r_2) = \frac{1}{N-1} \sum_{i=1}^{N} (r_{1,i} - \bar{r}_1) (r_{2,i} - \bar{r}_2)
\]

We can then rewrite the cosine of \( \alpha \) as:

\[
\cos \alpha = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}
= \frac{1}{\sqrt{(N-1) \cdot \text{Var}(r_1)} \sqrt{(N-1) \cdot \text{Var}(r_2)}} \mathbf{x} \cdot \mathbf{y}
= \frac{1}{N-1} \sum_{i=1}^{N} (r_{1,i} - \bar{r}_1) (r_{2,i} - \bar{r}_2)
= \text{Cov}(r_1, r_2) \sqrt{\text{Var}(r_1) \cdot \text{Var}(r_2)},
\]

which is the definition of correlation coefficient.

**Exercise 1c**

Since the correlation coefficient \( r \) can be expressed as the cosine of the angle between the two vectors described above, its maximum value is 1, and its minimum value is -1.

- \( r = 1 \) implies that both vectors are parallel, that is: in every single trial, the deviations of \( r_1 \) and \( r_2 \) from their respective means have the same sign and are proportional to each other. The two neurons are perfectly correlated.

- \( r = -1 \) implies that both vectors are antiparallel, that is: in every single trial, the deviations of \( r_1 \) and \( r_2 \) from their respective means have opposite signs and are proportional to each other. The two neurons are perfectly anti-correlated.

- \( r = 0 \) implies that there is no consistent link between \( r_1 \) and \( r_2 \). The two neurons are entirely uncorrelated.

**Exercise 1d**

Every probability distribution of a variable \( v \) is somewhat “centered” on its mean \( \bar{v} \). But this is not what we refer to in this exercise.

In fact, there are two kinds of probability distributions: those with a mean equal to zero, and those with a nonzero mean. The former kind of distribution is much easier to deal with mathematically, because it is centered on the origin.

For example, the mean of the elements of \( \mathbf{x} \) is zero, because

\[
\frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} \sum_{i=1}^{N} (r_{1,i} - \bar{r}_1)
= \frac{1}{N} \sum_{i=1}^{N} r_{1,i} - \frac{1}{N} \sum_{i=1}^{N} \bar{r}_1
= \bar{r}_1 - \frac{1}{N} N \bar{r}_1
= \bar{r}_1 - \bar{r}_1 = 0.
\]

The same is true for \( \mathbf{y} \). Only because \( \mathbf{x} \) and \( \mathbf{y} \) are centered in this sense, the cosine between them equals their correlation coefficient.

**Exercise 2a: Bayes’ theorem**

Quite intuitively, the probability for the firing rate \( r \) is the sum of the probabilities of observing it either due to stimulus \( \leftarrow \) or due to stimulus \( \rightarrow \). These two contributions have to be weighted by the prior probability of the respective stimulus to be present

\[
p(r) = p(r \mid \leftarrow)p(\leftarrow) + p(r \mid \rightarrow)p(\rightarrow)
= \frac{1}{2} \left(p(r \mid \leftarrow) + p(r \mid \rightarrow)\right),
\]

where in the last line we factored out the common prior \( p(\leftarrow) = p(\rightarrow) = 1/2 \).

The shape of \( p(r) \) will depend of course of \( p(r \mid \leftarrow) \) and \( p(r \mid \rightarrow) \). To make things more concrete we will assume that the probabilities \( p(r \mid \rightarrow) \) and \( p(r \mid \leftarrow) \) are like those shown in figure 2.
Having specified \( p(r \mid \rightarrow) \) and \( p(r \mid \leftarrow) \), the distribution \( p(r) \) is straightforward to plot:

![Figure 1](image1.png)

**Figure 1.** Distributions of the firing rate \( r \) of a neuron, conditioned to the stimuli \( s = \leftarrow \) (red) and \( s = \rightarrow \) (blue). Parameters: \( \mu_{\rightarrow} = 36 \text{ spikes/s}, \mu_{\leftarrow} = 64 \text{ spikes/s}, \sigma = 12 \text{ spikes/s} \).

Before going further, it is convenient to simplify the notation a bit to avoid writing \( p(r \mid \rightarrow)/p(r \mid \leftarrow) \) all the time. Let us define

\[
z = \frac{p(r \mid \rightarrow)}{p(r \mid \leftarrow)}. \tag{4}
\]

In terms of \( z \) equation (3) reads

\[
p(\leftarrow \mid r) = \frac{1}{1 + z} \tag{5}
\]

To compute the other posterior \( p(\rightarrow \mid r) \), we isolate it from the relation \( p(\rightarrow \mid r) + p(\leftarrow \mid r) = 1 \). We obtain:

\[
p(\rightarrow \mid r) = 1 - p(\leftarrow \mid r) = 1 - \frac{1}{1 + z} = \frac{1 + z}{1 + z} = \frac{1}{z^{-1} + 1}.
\]

Plugging back the definition of \( z \) into this last relation we finally get

\[
p(\rightarrow \mid r) = \frac{1}{1 + \frac{p(r \mid \leftarrow)}{p(r \mid \rightarrow)}}, \tag{6}
\]

which is equation (3) with all symbols \( (\leftarrow) \) and \( (\rightarrow) \) swapped.

To plot the two posteriors, recall that they only depend on \( z \):

\[
p(\leftarrow \mid r) = \frac{1}{1 + z}, \quad p(\rightarrow \mid r) = \frac{1}{1 + z^{-1}}. \tag{7}
\]

Computing \( z \) is easy once we know that the probabilities \( p(r \mid \leftarrow) \) and \( p(r \mid \rightarrow) \) are Gaussian:

\[
p(r \mid \leftarrow) \propto \exp\left(-\frac{(r - \mu_{\leftarrow})^2}{2\sigma^2}\right)
\]

\[
p(r \mid \rightarrow) \propto \exp\left(-\frac{(r - \mu_{\rightarrow})^2}{2\sigma^2}\right).
\]

Let us compute \( z \) explicitly

\[
z = \frac{p(r \mid \rightarrow)}{p(r \mid \leftarrow)} = \frac{\exp\left(-\frac{(r - \mu_{\rightarrow})^2}{2\sigma^2}\right)}{\exp\left(-\frac{(r - \mu_{\leftarrow})^2}{2\sigma^2}\right)}
\]

\[
= \exp\left(-\frac{(r - \mu_{\rightarrow})^2}{2\sigma^2}\right) \exp\left(\frac{(r - \mu_{\leftarrow})^2}{2\sigma^2}\right)
\]

\[
= \exp\left(\frac{(r - \mu_{\rightarrow})^2 + (r - \mu_{\leftarrow})^2}{2\sigma^2}\right)
\]

\[
= \exp\left(\frac{2r - (\mu_{\rightarrow} + \mu_{\leftarrow})(\mu_{\rightarrow} - \mu_{\leftarrow})}{2\sigma^2}\right)
\]

We can arrange the exponent in a more appealing way:

\[
z = \exp\left(\frac{\mu_{\rightarrow} - \mu_{\leftarrow}}{\sigma^2}\left[r - \frac{\mu_{\rightarrow} + \mu_{\leftarrow}}{2}\right]\right). \tag{8}
\]

**Exercise 2b**

The posterior probability for \( s = \leftarrow \) is the probability that, given we observed the firing rate \( r \), this firing rates was caused by stimulus \( (\leftarrow) \). We find the posterior with Bayes’ theorem:

\[
p(s \mid r) = \frac{p(r \mid s)p(s)}{p(r)}
\]

where \( s \) is a particular stimulus. In our case, Bayes’ theorem takes the form

\[
p(\leftarrow \mid r) = \frac{p(r \mid \leftarrow)p(\leftarrow)}{p(r)}
\]

\[
= \frac{p(r \mid \leftarrow)p(\leftarrow)}{p(r \mid \leftarrow)p(\leftarrow) + p(r \mid \rightarrow)p(\rightarrow)}
\]

\[
= \frac{1}{1 + \frac{p(r \mid \rightarrow)}{p(r \mid \leftarrow)}}. \tag{2}
\]

If we assume that the prior probabilities of the two stimuli are the same, \( p(\leftarrow) = p(\rightarrow) = 1/2 \), equation (2) simplifies to

\[
p(\leftarrow \mid r) = \frac{1}{1 + \frac{p(r \mid \rightarrow)}{p(r \mid \leftarrow)}}. \tag{3}
\]
The first factor in the exponent measures how easy it is to discriminate the two signals: if the two means are far apart compared to the standard deviation, this factor will be large—in absolute value. The second factor is the difference between the rate and the midpoint between the two peaks. The product of the two factors is positive if \( r \) is closer to \( \mu_\rightarrow \) than to \( \mu_\leftarrow \), and negative otherwise.

We can finally plot the posteriors. Given the probabilities \( p(r|\leftarrow) \) and \( p(r|\rightarrow) \) specified in figure 1, the posteriors look like in Figure 3.

![Figure 3. Posterior probabilities for the stimuli \( \leftarrow \) and \( \rightarrow \), for the case where priors are identical.](image)

Factoring out the common factor leads to

\[
\begin{align*}
&= \exp \left( \frac{1}{2\sigma^2} \right) \exp \left( -(r_1 - \mu_2)^2 - (r_2 - \mu_1)^2 \right) \\
&\quad + (r_1 - \mu_1)^2 + (r_2 - \mu_2)^2 \\
&= \exp \left( \frac{1}{2\sigma^2} \right) \\
&\quad \cdot \exp \left[ -\left( r_1 - 2r_1\mu_2 + \mu_2^2 \right) - \left( r_2 - 2r_1\mu_1 + \mu_1^2 \right) \\
&\quad + \left( r_1^2 - 2r_1\mu_1 + \mu_1^2 \right) + \left( r_2^2 - 2r_2\mu_2 + \mu_2^2 \right) \right] \\
&= \exp \left( \frac{1}{2\sigma^2} \right) \exp \left( 2r_1\mu_2 + 2r_2\mu_1 - 2r_1\mu_1 - 2r_2\mu_2 \right) \\
&= \exp \left( \frac{r_1\mu_2 + r_2\mu_1 - r_1\mu_1 - r_2\mu_2}{\sigma^2} \right) \\
&= \exp \left( -\left( r_1 - r_2 \right)/\mu_2 \right),
\end{align*}
\]

which, substituted into (3), gives:

\[
p(+ | r_1, r_2) = \frac{1}{1 + \exp \left[ -\left( r_1 - r_2 \right)/\mu_2 \right]}.
\]

This relation, for given values of \( \mu_1, \mu_2 \) and \( \sigma \), turns out to be a sigmoidal function of \( r_1 - r_2 \).

The corresponding posterior probability for stimulus \((-)\) can be computed by making use of the fact that in our paradigm, \((+)\) and \((-)\) are the only possible stimuli, and one of them is presented in every trial. Hence:

\[
p(- | r_1, r_2) = 1 - p(+ | r_1, r_2),
\]

\[
= \frac{1}{1 + \frac{p(r_1 | +)p(r_2 | -)}{p(r_1 | -)p(r_2 | +)}},
\]

which is the same as exchanging \((+)\) and \((-)\) in equation (3).

**Exercise 2c**

If the two priors \( p(\leftarrow) \) and \( p(\rightarrow) \) are not identical, we need to keep track of the factor \( p(\rightarrow)/p(\leftarrow) \) appearing in equation (2). It is not difficult to see that, in that case, the derivation to find the posterior is essentially the same once we replace the definition (4) with:

\[
z = \frac{p(r \rightarrow)}{p(r \leftarrow)} \frac{p(\rightarrow)}{p(\leftarrow)}.
\]

There are several ways to study the effect of this additional factor. A useful method is to concentrate on the firing rate \( r_{1/2} \) at which \( p(\leftarrow | r) = p(\rightarrow | r) = 1/2 \). At this point is equally probable that the firing rate is caused by either stimulus, which corresponds to maximum uncertainty. Looking at equation (7) it
is not difficult to convince oneself that \(p(-| r) = p(→| r) = 1/2\) can only occur if \(z = 1\). When the two priors \(p(-)\) and \(p(→)\) are equal, \(z\) is given by equation (8), which is 1 when \(r_{1/2}\) is exactly at the midpoint between \(μ_-\) and \(μ_+\).

When the priors are different, the condition \(z = 1\) reads:

\[
p(→) \frac{p(-)}{p(→)} \exp \left( \frac{μ_+ - μ_-}{σ^2} \left[ r_{1/2} - \frac{μ_+ + μ_-}{2} \right] \right) = 1. \tag{9}
\]

What happens if \(p(-) > p(→)\)? In that case the ratio \(p(→)/p(-) < 1\), which means that the exponential has to be larger than 1 to fulfill the condition \(z = 1\). This can only happen if \(r\) is closer to \(μ_+\) than to \(μ_-\). The effect of unequal priors is illustrated in Figure 4.

![Figure 4. Posterior probabilities for the stimuli ← and →, for the case where \(p(-) = 0.85\) and \(p(→) = 0.15\) (thick curves). The posterior probabilities for equal priors are shown in thin lines as a reference.](image)

**Advanced** A more rigorous way to prove this is to apply the logarithm at both sides of (9), use the property \(\ln \exp(x) = x\), and isolate \(r_{1/2}\). The final result is

\[
r_{1/2} = \frac{μ_+ + μ_-}{2} + \frac{σ^2}{μ_+ - μ_-} \ln \frac{p(-)}{p(→)}. \tag{10}
\]

**Exercise 2d**

The decision rule by which there’s leftward motion whenever \(r > \frac{1}{2}(μ_+ + μ_-)\) is the right rule to apply when both priors are equal. This no longer true when \(p(-)\) and \(p(→)\) are different. If, for instance, \(p(-)\) is much larger than \(p(→)\), stimulus ← is more likely to be the cause of the firing activity observed (loosely speaking). This means that we should favor the presence of ← in our decision rule by lowering the threshold.

The proper way to set the decision threshold is to assign leftward motion whenever \(p(-| r) > p(→| r)\), and rightward motion whenever \(p(→| r) > p(-| r)\). The threshold is therefore the firing rate at which \(p(-| r) = p(→| r)\). In the previous exercise we denoted by \(r_{1/2}\) this precise value, and we saw that it depends on the priors—see, for instance Figure 4 or equation (10). We can strengthen even more this point by plotting the dependence of the threshold on the ratio \(p(-)/p(→)\):

![Figure 5. Linear discriminant analysis. Exercise 3a](image)

Given the expression \(p(r | s)\) given in the exercise, it is easy to compute \(l(r)\):

\[
l(r) = \log \frac{\exp \left( \frac{1}{2}(r - μ_+)^T C^{-1} (r - μ_+) \right)}{\exp \left( \frac{1}{2}(r - μ_-)^T C^{-1} (r - μ_-) \right)} = \frac{1}{2}(r - μ_+)^T C^{-1} (r - μ_+) + \frac{1}{2}(r - μ_-)^T C^{-1} (r - μ_-)
\]

\[
= \frac{1}{2} \left( T C^{-1} μ_+ + T C^{-1} r - μ_+^T C^{-1} μ_- - r^T C^{-1} μ_- - μ_+^T C^{-1} r + μ_-^T C^{-1} μ_- \right).
\]

We can simplify this expression using the obvious property that the transpose of a scalar is a scalar, which leads to the following relations

\[
a^T b = b^T a, \quad a^T M b = b^T M^T a.
\]

We also use the fact that \(C\) (and hence \(C^{-1}\)) is symmetric: \((C^{-1})^T = (C^T)^{-1} = C^{-1}\), so that we can finally write

\[
l(r) = (μ_+ - μ_-)^T C^{-1} r - \frac{1}{2} μ_+^T C^{-1} μ_+ + \frac{1}{2} μ_-^T C^{-1} μ_- \tag{11}
\]

Note that \(l(r)\) is linear on \(r\). We can actually write

\[
l(r) = w^T r + w_0, \tag{12}
\]

where

\[
w = C^{-1}(μ_+ - μ_-), \tag{13}
\]

\[
w_0 = -\frac{1}{2} μ_+^T C^{-1} μ_+ + \frac{1}{2} μ_-^T C^{-1} μ_- \tag{14}
\]

It is also worth checking that the result is consistent with the previous exercise, where the firing rate \(r\)
was one-dimensional. Recall that in exercise 2b we defined \( z \) as the ratio of \( p(r \rightarrow) \) and \( p(r \leftarrow) \). According to that definition, \( l(r) = -\log z \). It is not difficult to see that \( -\log z \), with \( z \) given in equation (8), corresponds to the one-dimensional version of equation (11). Check it!

**Exercise 3b**

We have done all the hard work. Given equation (12), it is clear that the decision boundary will be given by all the \( r \) that fulfill the equation

\[
 w^T r + w_0 = 0,
\]

with \( w \) and \( w_0 \) given by equations (13)–(14). If the number of neurons is \( n \), equation (15) defines a hyperplane of \( n - 1 \) dimensions embedded in the \( n \)-dimensional space.