Week 5 - Model fitting

The goal of this exercise is to:

- Get familiar with linear regression, GLMs and their applications to neuronal tuning.
- Compute the optimal parameters of the models through maximum likelihood estimation.
- Understand how to evaluate the goodness of fit.

1 Linear regression

One of the fundamental problems in Neuroscience is the modeling of the functional relationship between sensory stimuli and neural activity. Consider the example of seeing an object, which generates specific neural activity in the visual brain areas. Is it possible to construct a model that accurately captures this relationship?

One possible approach consists in assuming a linear relationship between the stimulus and the neural activity. Therefore, we can use linear regression to model the relationship between the neural activity (dependent variable) and the sensory stimulus (independent variable). For instance, we can measure the neural activity of an animal's visual cortex while presenting different visual stimuli. By fitting the neural activity with linear regression, we can obtain an equation that describes how the neural activity **y** changes in response to the characteristics of different stimuli **X**. How can we learn the parameters of the linear regression? How much of the neural activity can we reconstruct with linear regression? Are there better model that can be used to reconstruct the neural activity?

We start addressing these questions by building our first linear regression model. Given the neural activity y of a neuron, we can retrieve its firing rate using the input stimulus $\mathbf{x}$. Since we cannot achieve perfect reconstruction (e.g. measurement noise), we can include everything that our model cannot capture under noise which we initially assume to be Gaussian distributed. Therefore, we can build a Linear-Gaussian regression model:

$$\mathbf{y} = \mathbf{X}\theta + \eta, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ with N the number of observations, $\mathbf{X} \in \mathbb{R}^{N,F}$ represents the stimulus matrix with F the number of features, $\theta \in \mathbb{R}^F$ are the linear model parameters and $\eta \in \mathbb{R}^N$ is a multivariate normal Gaussian distribution ($\eta_i \sim \mathcal{N}(0, \sigma)$). The quality of our fitting can be assessed with the mean square error (MSE), i.e. the average squared difference between the predicted values and the actual ones. Therefore, we can find the optimal parameters for our model by minimizing the MSE. In the following exercises, you will need to retrieve the equation for learning the parameters θ.

Exercise 1.1

Find the optimal θ^* by minimizing[a] the mean square error (MSE) between the ground truth neural activity (y) and the one reconstructed by the model ($\hat{y}$):

$$\theta^* = \arg\min_\theta MSE(\mathbf{y}, \hat{\mathbf{y}}) \tag{2}$$

[a] Hint: Remember that the stationary points of a differentiable function can be obtained by setting the derivative equal to zero.

Solutions:

$$\theta^* = \arg\min_\theta MSE(\mathbf{y}, \hat{\mathbf{y}}) = \arg\min_\theta ||\mathbf{y} - \mathbf{X}\theta||_2^2 \tag{3}$$

We set the gradient with respect to θ to 0 to find the minimum of the MSE: [1]

$$\begin{aligned}
\nabla_\theta ||\mathbf{y} - \mathbf{X}\theta||_2^2 = \nabla_\theta \left((\mathbf{y} - \mathbf{X}\theta)^T(\mathbf{y} - \mathbf{X}\theta)\right) &= 0 \\
\nabla_\theta \left(\mathbf{y}^T\mathbf{y} - \theta^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\theta + \theta^T\mathbf{X}^T\mathbf{X}\theta\right) &= 0 \\
\nabla_\theta \left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\theta + \theta^T\mathbf{X}^T\mathbf{X}\theta\right) &= 0 \\
-2\mathbf{y}^T\mathbf{X} + 2\theta^T\mathbf{X}^T\mathbf{X} &= 0
\end{aligned}$$

where $\theta^T\mathbf{X}^T\mathbf{y}$ is equal to $\mathbf{y}^T\mathbf{X}\theta$ as they are scalars and the following gradient $\nabla_\theta \left(\theta^T\mathbf{X}^T\mathbf{X}\theta\right)$ can be computed using the following identity $\nabla_\mathbf{x} \left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) = 2\mathbf{x}^T\mathbf{A}$. In our case, we have $\mathbf{x} = \theta$ and $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ which results in $\nabla_\theta \left(\theta^T\mathbf{X}^T\mathbf{X}\theta\right) = 2\theta^T\mathbf{X}^T\mathbf{X}$.

We isolate now the model parameters θ:

$$\begin{aligned}
2\theta^T\mathbf{X}^T\mathbf{X} &= 2\mathbf{y}^T\mathbf{X} \\
\mathbf{X}^T\mathbf{X}\theta &= \mathbf{X}^T\mathbf{y} \\
(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\theta &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

$$\theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

Exercise 1.2

Find the optimal θ^* by computing the maximum likelihood estimation (MLE):

$$\theta^* = \arg\max_\theta \log \mathcal{L}(\theta|\mathbf{X}, \mathbf{y}) \tag{5}$$

Solutions:

Differently from before, now we want to maximize the log likelihood. Since the errors are gaussian distributed, we have that the log likelihood is:

$$\log \mathcal{L}(\theta|\mathbf{X}, \mathbf{y}) = \log \left(\frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\theta)\right)\right) \tag{6}$$

Again, we compute the derivative and set it to be equal to 0.

$$\begin{aligned}
\nabla_\theta \left(\log \mathcal{L}(\theta|\mathbf{X}, \mathbf{y})\right) &= 0 \\
\nabla_\theta \left(-\frac{k}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2}((\mathbf{y} - \mathbf{X}\theta)^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\theta))\right) &= 0 \\
\nabla_\theta \left(-\frac{k}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(\mathbf{y}^T\Sigma^{-1}\mathbf{y} - 2\mathbf{y}^T\Sigma^{-1}\mathbf{X}\theta + \theta^T\mathbf{X}^T\Sigma^{-1}\mathbf{X}\theta)\right) &= 0 \\
\frac{1}{2}(2\mathbf{X}^T\Sigma^{-1}\mathbf{y} - 2\mathbf{X}^T\Sigma^{-1}\mathbf{X}\theta) &= 0
\end{aligned}$$

[1] The gradient with respect to $\theta \in \mathbb{R}^F$ of a function that depends also on other variables is the vector of the partial derivatives with respect to the components of the variable vector θ : $[\nabla_\theta f(\mathbf{x}, \theta)]_i = \frac{\partial}{\partial\theta_i} f(\mathbf{x}, \theta)$, for $i = 1, ..., F$

Let's isolate again the model parameters θ:

$$\mathbf{X}^T\Sigma^{-1}\mathbf{X}\theta = \mathbf{X}^T\Sigma^{-1}\mathbf{y}$$

Therefore, we obtain that θ_{MSE} is equal to θ_{MLE} (as the errors are normally distributed with constant variance):

$$\theta^* = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{7}$$

2 Generalized linear models (GLM)

Previously, we assumed that the probability of a spike occurring at any given time is a linear function of some input variables. However, it is usually assumed that neural spike counts follow a Poisson distribution. In this case, we can build the following model called the linear nonlinear Poisson (LNP) model:

$$\mathbf{y} \sim \text{Poisson}(f(\mathbf{X}\theta)) \tag{8}$$

where $\mathbf{X} \in \mathbb{R}^{N,F}$ represents the stimulus matrix, $\theta\mathbf{X}$ is the linear projections and $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is an element-wise rectifying nonlinearity, such as the exponential function ($f(k) = e^k$) or the softplus function ($f(k) = \log(1 + e^k)$), where $\mathcal{M}$ is the space of the conditional expectation of y.

Unfortunately, we cannot compute the optimal parameter as we did before for any non linear function because the derivative of the log likelihood has not always a closed form solution. However, we can take advantage of optimization methods such as gradient descent to find the optimal parameters. To this end, we can specify the loss function to optimize our model as the minimization of the negative log likelihood of the spikes (Poisson loss).

Exercise 2.1

Derive the Poisson loss as the negative log likelihood of the spikes given the expected spike counts ($\lambda = f(\mathbf{X}\theta)$) predicted by the model:

Solutions:
First, we compute the negative log likelihood of the Poisson distributed data:

$$-\log \mathcal{L}(\lambda|\mathbf{X}, \mathbf{y}) = -\log \prod_{n=1}^{N} \frac{\lambda_n^{y_n} e^{-\lambda_n}}{y_n!} = -\sum_{n=1}^{N} y_n \log(\lambda_n) - \lambda_n - \log(y_n!) \tag{9}$$

where N is the number of datapoints. The poisson loss can be defined as the average over the data points:

$$loss = \frac{1}{N}\sum_{n=1}^{N} \lambda_n - y_n \log(\lambda_n) \tag{10}$$

where we removed the $\log(y_n!)$ as it doesn't depend on λ.

3 Model evaluation

To assess the performance of a model in fitting data, various metrics can be used. There are two popular metrics that can be used the coefficient of determination (R^2) and the fraction of explained variance (EV). The goal of these metrics is to quantify how well a model captures the variance in the observed data or how well the predicted values fit the observed values. To calculate these metrics, we start with a set of observed data points ($\mathbf{y}$) and a corresponding set of predicted data points ($\hat{\mathbf{y}}$).

Exercise 3.1

Derive the fraction of explained variance (EV) by computing the variance of the residuals and the total variance of the observed data. The fraction of explained variance is defined as:

$$EV = 1 - \frac{Var(y - \hat{y})}{Var(y)} \tag{11}$$

Solutions:
We can expand the equation of the explained variance as:

$$EV = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y_n} - \mathbb{E}(y - \hat{y}))^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{12}$$

Exercise 3.2

Derive now the coefficient of determination (R^2) by computing the total sum of squares (TSS), the residual sum of squares (RSS) and the explained sum of squares (ESS)[a] . Then, the coefficient of determination is defined as:

$$R^2 = \frac{ESS}{TSS} \tag{13}$$

When can the fraction of explained variance (EV) be considered the same as the coefficient of determination (R^2)?

[a] Hint: Consider that you need to use the average neural activity $\bar{y}$ for the TSS and the predicted neural activity $\hat{y}$ for the RSS.

Solutions:
First, we compute the the total sum of squares (TSS):

$$TSS = \sum_{n=1}^{N}(y_n - \bar{y})^2 \tag{14}$$

where $\bar{y}$ is the average of the observed data. Then, we compute the residual sum of squares (RSS):

$$RSS = \sum_{n=1}^{N}(y_n - \hat{y_n})^2 \tag{15}$$

The explained sum of squares (ESS) is then computed as:

$$ESS = TSS - RSS = \sum_{n=1}^{N}(y_n - \bar{y})^2 - \sum_{n=1}^{N}(y_n - \hat{y_n})^2 \tag{16}$$

Finally, the coefficient of determination is:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{n=1}^{N}(y_n - \bar{y})^2 - \sum_{n=1}^{N}(y_n - \hat{y_n})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} = 1 - \frac{\sum_{n=1}^{N}(y - \hat{y_n})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{17}$$

By comparing eq. 12 and eq. 17, you can observe that the explained variance (EV) becomes the same as the coefficient of determination (R^2) when the mean of the residuals ($\mathbb{E}(y - \hat{y})$) is equal to 0.