Exercise 6.1: Overfitting and model selection (16 points)

This exercise demonstrates the problem of overfitting and how model comparison can provide a way to choose the optimal model given the measured data. This week, we will look at a polynomial model of order P:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_P x^P + \varepsilon = \boldsymbol{x}\boldsymbol{\theta} + \varepsilon, \tag{1}$$

where $\boldsymbol{x} = \left(x, x^2, \cdots, x^P\right)$, $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_P)^T$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ iid.

(a) (*4 points*) Assuming N independent observations $\boldsymbol{y} = (y_1, ..., y_N)$ are obtained, show that the log-likelihood for the model defined above is given by:

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{(\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta})}{2\sigma^2} \tag{2}$$

(b) (*1 point*) Write down the expression for the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for the P^{th}-order polynomial model defined in eq (1). *Hint:* Both terms are a combination of an accuracy term (which is given by the log likelihood that you derived in part (a) of this exercise) and a complexity term.

(c) (*5 points*) Now, we consider a Gaussian prior distribution with mean $\boldsymbol{\mu}_0 = (0, \cdots, 0)^T$ and covariance $\Sigma_0 = I$ over the model parameters $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}; \boldsymbol{\mu_0}, \Sigma_0) = \frac{1}{\sqrt{|2\pi I|}}\exp\left(-\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{2}\right) \tag{3}$$

For the linear Gaussian model specified by eqs (1) and (3), it is possible to obtain an analytical expression for the model evidence $p(\boldsymbol{y})$. Show that the model evidence is given by the following expression

$$\begin{aligned}\log p\left(\boldsymbol{y}\right) = &-\frac{\left(\boldsymbol{y} - X\boldsymbol{\mu}_{\theta|y}\right)^T\left(\boldsymbol{y} - X\boldsymbol{\mu}_{\theta|y}\right)}{2\sigma^2} - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi \\ &- \frac{1}{2}\boldsymbol{\mu}_{\theta|y}^T\boldsymbol{\mu}_{\theta|y} + \frac{1}{2}\log\left|\Sigma_{\theta|y}\right|\end{aligned} \tag{4}$$

where $\boldsymbol{\mu}_{\theta|y}$ and $\Sigma_{\theta|y}$ are the mean and covariance of the posterior distribution over model parameters $\boldsymbol{\theta}$. *Hint*: Plug the expression for the likelihood and the prior density into the definition of the model evidence and solve the integral by *completing-the-squares* in the exponent. *Additional hint*: To obtain the expressions for $\boldsymbol{\mu}_{\theta|y}$ and $\Sigma_{\theta|y}$, plug in $\Sigma_y = \sigma^2 I$, $\boldsymbol{\mu}_0 = (0, \cdots, 0)^T$ and $\Sigma_0 = I$ into the expressions on the slides from Lecture 12 and replace the respective terms here.

The remaining part of this exercise requires a computer. From now on, we will consider the quadratic model:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \varepsilon. \tag{5}$$

with the following model parameter values: $\theta_0 = 0.3$, $\theta_1 = -0.1$, $\theta_2 = 0.5$ and $\sigma^2 = 0.001$. Let x vary from -0.5 to 0.2 in steps of 0.1.

(d) (*4 points*) Create a function that generates data from the model in eq (5) and evaluates the log-likelihood, AIC, BIC and the log model evidence. Plot those four measures for $P = 1, 2$ and 7. *Hint:* Log-likelihood, AIC and BIC should be evaluated at the maximum likelihood estimate (MLE) which for the polynomial model is given by $\boldsymbol{\theta}_{MLE} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

(e) (*2 points*) Repeat the steps in (c) for $N = 100$ times and plot the log-likelihood, AIC, BIC, and the log model evidence obtained with $P = 1, 2$ and 7. What do you notice about the different model selection criteria?

Exercise 6.2: Practical application of Bayesian model selection and averaging (11 points)

This exercise is a simple demonstration of how to implement fixed-effects and random-effects Bayesian model selection (BMS), and how to perform Bayesian model averaging (BMA). Here, we use a set of exemplary Dynamic Causal Models (DCMs) of fMRI data that can be downloaded from the lecture webpage on Moodle. The dataset comprises fMRI data from 20 synthetic subjects that have been fitted used two DCMs that differed in the exact network architecture. The goal is to compare the two models across the 20 subjects using BMS, and to compute BMA parameter estimates across models in a subsequent step.

(a) (*2 point*) Compute the individual log Bayes factors and plot the results. *Hint*: The negative free energies (stored in the structure $DCM.F$) serve as a lower bound approximation to the log model evidence and can thus be used to compute the log Bayes factors.

(b) (*3 point*) Perform fixed-effects Bayesian model selection by computing the Group Bayes factor and the posterior model probabilities. What do you notice when you compare the model comparison results with the model preferences observed at the single-subject level from (a)? Why?

(c) (*3 point*) Perform random-effects Bayesian model selection by computing the protected exceedance probabilities. *Hint*: Use the Statistical

Parametric Mapping (SPM) software, which you can download for free from `http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`. *Additional hint*: Either use the GUI (Click *Dynamic Causal Modeling* and *Action: compare*) or call the relevant SPM routines (*spm_BMS.m*) in your script. Depending on the approach that you choose, please provide the matlabbatch or your script as the solution to this exercise.

(d) (*3 point*) Perform random-effects Bayesian model averaging for the two DCMs, using SPM. Use the BMA results to compute, for each connection, the posterior probability that the connection strength differs from zero. *Hint*: Either use the GUI or call the relevant SPM routines (*spm_dcm_bma.m* for BMA, and *spm_BMS_gibbs.m* for obtaining the random-effects posterior model probabilities necessary for BMA) in your script. *Additional hint*: Group BMA posterior mean and standard deviation necessary for computing the posterior probabilities are stored in *bma.mEp* and *bma.sEp*, respectively. Depending on the approach that you choose, please provide the matlabbatch or your script as the solution to this exercise.

Exercise 6.3: Derivation of random-effects Bayesian model selection scheme (8 points)

In this exercise, you will derive the update equations for the random-effects Bayesian model selection (BMS) scheme introduced in Stephan et al. (2009). This exercise does not require a computer.

(a) (*2 point*) Write down the joint probability $p(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{m})$ of the model parameters $\boldsymbol{r} = [r_1, \cdots, r_k]$ and $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_k]$, and the data $\boldsymbol{y} = [y_1, \cdots, y_N]$ for the random-effects BMS scheme. Make use of

$$p(\boldsymbol{r}|\boldsymbol{\alpha}) = Dir(\boldsymbol{r}, \boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \prod_k r_k^{\alpha_k - 1} \tag{6}$$

$$p(\boldsymbol{m}_n|\boldsymbol{r}) = \prod_k r_k^{m_{nk}} \tag{7}$$

where $Z(\boldsymbol{\alpha}) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$, k is the model index and n denotes the subject index. *Hint*: Apply the product rule and exploit the conditional independence assumptions of the model. *Additional hint*: The joint probability should be expressed as a function of the individual model evidences $p(y_n|m_{nk})$.

(b) (*4 point*) Use the expression for the joint probability obtained in (a) to derive the variational energies for $\boldsymbol{r}$ and $\boldsymbol{m}$

$$I(\boldsymbol{r}) = \langle \log p(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{m}) \rangle_{q(\boldsymbol{m})} \tag{8}$$

$$I(\boldsymbol{m}) = \langle \log p(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{m}) \rangle_{q(\boldsymbol{r})} \tag{9}$$

under the following mean-field assumption $q(\boldsymbol{r}, \boldsymbol{m}) = q(\boldsymbol{r})q(\boldsymbol{m})$. *Hint*: The expected value is defined as

$$\langle \log p(y, \Theta_1, \Theta_2) \rangle_{q(\Theta_1)} = \int \log p(y, \Theta_1, \Theta_2) q(\Theta_1) d\Theta_1 \tag{10}$$

Additional hint: For the derivation of the update equations, the definition of the digamma function will be helpful.

(c) (*2 point*) Provide the final iterative variational Bayesian update scheme. *Hint*: Evaluate the approximate posterior distributions $q(\boldsymbol{r}) \propto \exp(I(\boldsymbol{r}))$ and $q(\boldsymbol{m}) \propto \exp(I(\boldsymbol{m}))$ and read of the update equations by comparing the results to known distributions.

References

Stephan, K., Penny, W., Daunizeau, J., Moran, R., Friston, K., 2009. Bayesian model selection for group studies. *NeuroImage* 46, 1004-1017.

Announcements:

- You are encouraged to solve the exercises in groups of 2-3.
- Please submit your solutions via moodle `https://moodle-app2.let.ethz.ch/mod/assign/view.php?id=883108`.
- Post your questions on `https://moodle-app2.let.ethz.ch/mod/assign/view.php?id=1042014`.