Assignment 8: Feed Forward Neural Networks

Consider the following FFNN as an implementation of a nonlinear regression of response y_i on covariates $x_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$:

$$\begin{aligned} y_i &= \sum_{j=1}^{M} \beta_j \psi(x_i'\gamma_j) + \epsilon_i, \quad i = 1, \ldots, N, \\ & \epsilon_i \sim N(0, \sigma^2), \\ & \psi(\eta) = \exp(\eta)/(1 + \exp(\eta)). \end{aligned} \tag{1}$$

Here $\gamma_j = (\gamma_{j0}, \ldots, \gamma_{jp})$ denotes the logistic regression coefficients in the potential for the j-th hidden node. To undertake Bayesian inference with this model, we shall use the following prior.

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2), \quad \gamma_j \sim N(\mu_\gamma, S_\gamma^2), \; j = 1, \ldots, M. \tag{2}$$

Question 1

Proof the following lemma. The lemma provides the marginalised likelihood, where for the sake of an easier notation, we shall omit dependence on the hyperparameters.

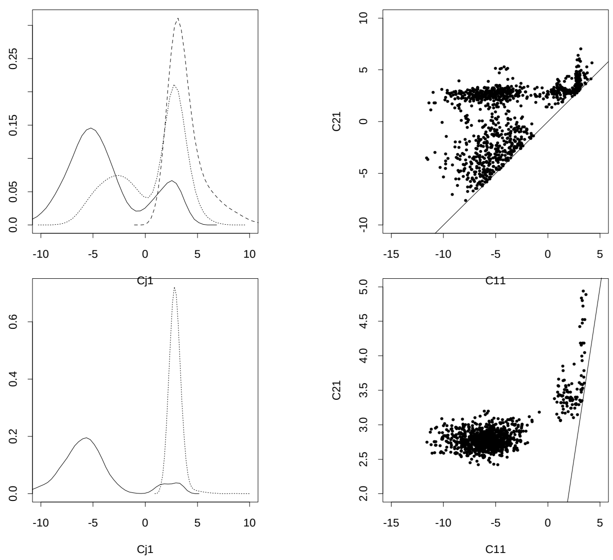
Lemma 1 *Let* $z_{ij} = z_{ij}(\gamma) = \psi(x_i'\gamma_j)$, $Z = (z_{ij})_{i=1,\ldots,N}^{j=1,\ldots,M}$, $\mathbf{1} = (1)_{i=1,\ldots,M}$, $A = Z'Z/\sigma^2$, $\rho = Z'y/\sigma^2$, $C = 1/\sigma_\beta^2 I$, $\delta = \mu_\beta/\sigma_\beta^2 \mathbf{1}$. *Let* $m_b(\gamma) = (A+C)^{-1}(\rho+\delta)$ *and* $S_b(\gamma) = (A+C)^{-1}$. *Then,*

$$\begin{aligned} p(y|\gamma) &= \frac{p[\beta = m_b(\gamma)]}{p[\beta = m_b(\gamma)|y, \gamma]} \prod_{i=1}^{N} p[y_i|\beta = m_b(\gamma), \gamma] \\ &= p[\beta = m_b(\gamma)]|S_b(\gamma)|^{1/2} \prod_{i=1}^{N} p[y_i|\beta = m_b(\gamma), \gamma]. \end{aligned}$$

Hint: Use the "candidate's formula": $p(x) = p(y|x)/p(x, y)$.

Question 2.

Consider the following figure:



The figures show posterior distributions $p(\gamma_{j1}|y)$ and $p(\gamma_{11}, \gamma_{21}|y)$ for neural network models with $M = 3$ hidden neurons (top row) and $M = 2$ hidden neurons (bottom row). Both are estimated using the same data set. To avoid multimodality due to permuation of the indices $j = 1, \ldots, M$ we added a constraint $\gamma_{11} \leq \gamma_{21} \leq \gamma_{31}$.

Explain the multimodality in the posterior distributions shown in the top two panels. Why did most of the posterior multimodality disappear under $M = 2$? Can you give any possible explaination for the remaining secondary mode in $p(\gamma_{11}, \gamma_{21}|M = 2, y)$ (right lower panel)?

Question 3.

The answer to Question 2 might lead you to include M in the parameter vector to allow for a random M. We can achieve this by adding to model (1)-(2) a prior on M:

$$p(M) \propto \alpha^M (1 - \alpha), \; M \leq M^*, \tag{3}$$

where M^* is some upper bound, and α is a fixed hyperparameter.

Assume that we have a Markov chain Monte Carlo simulation scheme to simulate from $p(\beta, \gamma|M, y)$ (i.e., posterior simulation under model (1)-(2)). How would you extend this scheme to implement inference in the "variable architecture" model (1)-(3)?

Note: This is an open-ended question with many possible answers. I don't expect you to perfectly solve the question, just try to come up with some reasonable suggestions how you would go about it.