Introduction to NeuraBASE: Neuronal Network Modelling From A Bayesian Perspective

Robert Hercus
Choong-Ming Chin
Kim-Fong Ho

neuraMATIX

Introduction

The Bayesian probability theory is one of many approaches that provide a framework into the understanding of neural network modelling. Conventional Bayesian learning for neural networks can be construed as an inference technique for estimating the parameters for the model given the training datasets. The work of Bayesian neural network learning is currently focused on the application of a multilayer perceptron to approximation and curve fitting problems, where the data are mostly quantitative (see MacKay [3], Mueller and Insua [4] and Ripley [5]). For these problems, neural networks are viewed as highly nonlinear approximation functions, where the parameters or network weights are typically estimated by least squares optimisation techniques.

This paper analyses the modelling aspect of neuronal networks in an unsupervised learning environment from a Bayesian perspective. In this study, the data under consideration are categorical variables, whereby, the neuronal network architecture incorporates a decision analysis of propagating child nodes based on their parents' association, in which prior knowledge plays a role. Based on a neuronal network approach described in Hercus [6], we will discuss how the iterative Markov chain Monte Carlo method can be exploited to estimate the unknown parameters of the neuronal network model, where its simulated equilibrium distribution approaches the posterior distribution for the parameters.

In principal, Bayesian methods are sufficiently dynamic to be applied to standard neural network models, provided the model can be interpreted in terms of statistical terminology. In this case, the Bayesian approach is used to elucidate the relationship between parent and child nodes based on the probabilistic nature of the prior distribution and, to provide a good predictive measure of the decision to activate a child node based on the observed knowledge of its parental nodes.

Self-Learning Neuronal Network Architecture

This section provides an introduction to the NeuraBASE self-learning neuronal network architecture. The main focus of this technique is to allow the network to detect meaningful representations of patterns from the dataset autonomously. This is achieved by setting up a built-in unsupervised learning rule that represents the statistical structure implicit in the network. The result is a powerful unsupervised learning method with potential widespread applications.

In essence, the NeuraBASE concept is based on linking two parent nodes to form an association, the child node, which will encapsulate the information of both its parents, and form links with other nodes, as depicted in **Figure 1**.

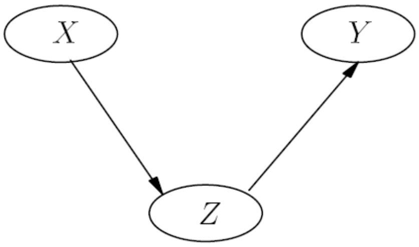

Figure 1. *An example of a forward linking process that links information from X to Y to form a child node Z*

In **Figure 1**, we can see that the central point of this new technique is the “firing" direction of information from the parent nodes to the child node. As opposed to many conventional network architectures where, both parents fire their information concurrently to the child node, the node *Z* is formed by first firing of information from node *X* to the child node, followed by node *Y*. This is known as a *forward linking direction*. A *backward linking direction* is based on firing of information from node *Y* to node *X* to form a new child node that is different from *Z* due to a different combination of information.

To activate a child node that is deemed to be significant, the NeuraBASE unsupervised learning rule consists of two phases of assessments. Using **Figure 1** as an illustration, in Phase 1 of the assessment, we perform a hypothesis test of independence where the null model is written as,

$$\mathrm{H}_0 : X \text{ and } Y \text{ are mutually independent.}$$

against the alternative model,

$$\mathrm{H}_1 : X \text{ and } Y \text{ are not mutually independent.}$$

Under H_0, the estimated probability of node *Z* is,

$$\widehat{\mathrm{P}}(\mathrm{H}_0) = \frac{\mathrm{n_x}}{\mathrm{N}}\frac{\mathrm{n_y}}{\mathrm{N}}$$

and under the alternative,

$$\widehat{\mathrm{P}}(\mathrm{Z}|\mathrm{H}_1) = \frac{\mathrm{n_z}}{\mathrm{N}}$$

where $\mathrm{n_x}, \mathrm{n_y}, \mathrm{n_z}$ are the frequencies of prevalent patterns in nodes *X*, *Y* and *Z* respectively, while *N* denotes the total size of the dataset.

Our hypothesis test uses a Bayes factor, $\mathrm{B}_{01}(\mathrm{Z})$ defined as,

$$\mathrm{B}_{01}(\mathrm{Z}) = \frac{\widehat{\mathrm{P}}\ (\mathrm{Z}|\mathrm{H}_0)}{\widehat{\mathrm{P}}\ (\mathrm{Z}|\mathrm{H}_1)}$$

Eq (1)

where it is an intuitive measure of the extent to which the information in *Z* will change the odds of the null model relative to the alternative model.

Hence,

$$B_{01}(Z)\begin{cases} > 1 \ \textit{data support null hypothesis} \\ < 1 \ \textit{data support alternative hypothesis} \\ = 1 \ \textit{no conclusive evidence to support either hypothesis} \end{cases}$$

From the hypothesis test using the Bayes factor and in the event that the null hypothesis is accepted, we can say that, intuitively *X* and *Y* have such a weak association that the child node *Z* can be inferred to be *non-significant* relative to its parents and hence, should not be formed.

On the other hand, if the null hypothesis of independence is rejected or when $B_{01}(Z) = 1$, we can infer that relative to its parents node, *Z* is *significant or weakly significant*. Based on this, we can therefore make a concrete decision on, whether or not, to activate the node. The Phase 2 assessment of our unsupervised learning rule has a simple probabilistic test. Activation of the child nodes is based on a conditional probability criterion by comparing them with their parents. However, this test is only performed for a three level nodal structure where all the parental nodes are active (see **Figure 2**).

In the figure below we consider the active parent nodes X_1 and X_2, X_3 and X_4, of *X* and *Y* respectively, which are also active.

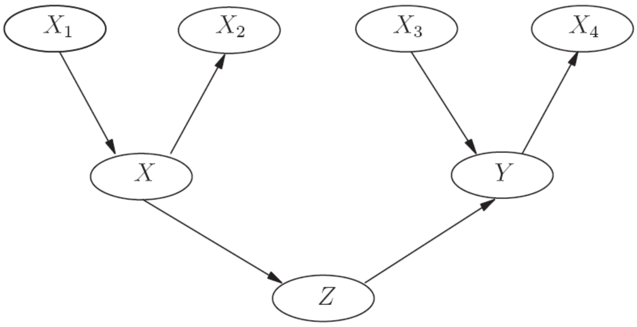

Figure 2. *A schematic diagram of a three-generation forward linking neuronal network process*

And let the following be the frequencies of information stored in each node as compared in the training dataset.

$$n(X_1) = N_{x1},\ n(X_2) = N_{x2},\ n(X_3) = N_{x3},\ n(X_4) = N_{x4},$$

and let the following be the frequencies of information stored in each node:

$$n(X) = n_x, \qquad n(Y) = n_y, \qquad n(Z) = n_z.$$

As child nodes encapsulate certain information from both of its parents, therefore,

$$n_x \leq \min\{N_{x1}, N_{x2}\}, n_y \leq \min\{N_{x3}, N_{x4}\},\ n_z \leq \min\{n_x, n_y\}.$$

Our activation test for node *Z* then takes the form,

$$\hat{P}\,(Z|X,Y) > \min\{\,\hat{P}\,(X_1, X_2), \hat{P}\,(Y|X_3, X_4)\}$$

Eq (2)

or equivalently,

$$\frac{n_z}{n_x + n_y - n_z} > \min\{\frac{n_x}{N_{x1} + N_{x2} - n_x}, \frac{n_y}{N_{x3} + N_{x4} - n_y}\}$$

Eq (3)

where, if **Eq (2)** holds, we will activate node *Z*; otherwise we will not activate the node. The inequality **Eq (2)** can also be explained as a measuring criterion of the conditional strength of node *Z*, given *X* and *Y*, which must be greater than either of its parents' conditional strengths. By performing this test we ensure that patterns that are strongly featured in the training dataset will continue to dominate at each level while phasing out weak active nodes even though they are significant at Phase 1 of the hypothesis testing stage.

Instead of linking only active nodes and forming new nodes at each level, the unsupervised learning rule of Phase 1 and 2 can also be utilised in a *multi-layer* fashion, where, active nodes from different levels can be linked to form new patterns. A simple strategy is to first assemble all active nodes that could not be linked to form new active nodes in their respective levels to become a set of input nodes in a new layer.

From the unsupervised learning rule discussed, we have the following algorithm:

Algorithm: Activating Meaningful Nodes

Given a sequence of length N of a training dataset, set the initial layer of the network, $l = 1$, and construct a set of input nodes $\mathrm{X}_1^{(l)}, \mathrm{X}_2^{(l)}, \ldots, \mathrm{X}_\mathrm{N}^{(l)}$, where each node occupies a single-element of the training data. Set the level of the network as $\mathrm{m} = 1$.

Step 1
Initial Phase 1 test for Level 2 nodes. Set $\mathrm{m} \leftarrow \mathrm{m} + 1$. Set a child node in level m, $\mathrm{X}_{\mathrm{ij}}^{(l)}$ as an *active node* given its parents $\mathrm{X}_\mathrm{i}^{(l)}$ and $\mathrm{X}_\mathrm{j}^{(l)}$ if,

$$\frac{\mathrm{n_{ij}}}{\mathrm{n_i n_j}} \geq \frac{\mathrm{N^{(ij)}}}{\mathrm{N^{(i)} N^{(j)}}}$$

Eq (4)

where $\mathrm{n_i}, \mathrm{n_j}$ and $\mathrm{n_{ij}}$ are frequencies of nodes $\mathrm{X}_\mathrm{i}^{(l)}$, $\mathrm{X}_\mathrm{j}^{(l)}$ and $\mathrm{X}_{\mathrm{ij}}^{(l)}$ respectively, and $\mathrm{N^{(i)}}$, $\mathrm{N^{(j)}}$and $\mathrm{N^{(ij)}}$ are the total pattern lengths for node types $\mathrm{X}_\mathrm{i}^{(l)}$, $\mathrm{X}_\mathrm{j}^{(l)}$ and $\mathrm{X}_{\mathrm{ij}}^{(l)}$ that can be constructed from the training dataset.

Step 2
If $= \mathrm{N}$, then STOP. Or else set $\mathrm{m} \leftarrow \mathrm{m} + 1$, and go to Step 3.

Step 3
Phase 1 test for Level 3 onwards. Set a child node in level m, $\mathrm{X}_{\mathrm{ijkl}}^{(l)}$ as a *significant node* given its active parents, $\mathrm{X}_{\mathrm{ij}}^{(l)}$ and $\mathrm{X}_{\mathrm{kl}}^{(l)}$ if,

$$\frac{\mathrm{n_{ijkl}}}{\mathrm{n_{ij} n_{kl}}} \geq \frac{\mathrm{N^{(ijkl)}}}{\mathrm{N^{(ij)} N^{(kl)}}}$$

Eq (5)

where $\mathrm{n_{ij,}n_{kl}}$ and $\mathrm{n_{ijkl}}$ are frequencies of nodes $\mathrm{X}_\mathrm{i}^{(l)}$, $\mathrm{X}_\mathrm{j}^{(l)}$ and $\mathrm{X}_{\mathrm{ij}}^{(l)}$ respectively, and $\mathrm{N^{(ij)}}$, $\mathrm{N^{(kl)}}$and $\mathrm{N^{(ijkl)}}$ are the total pattern lengths for node types $\mathrm{X}_{\mathrm{ij}}^{(l)}$, $\mathrm{X}_{\mathrm{kl}}^{(l)}$ and $\mathrm{X}_{\mathrm{ijkl}}^{(l)}$ that can be constructed from the training dataset.

Step 4
Phase 2: At level m, given active nodes $\mathrm{X}_{\mathrm{ij}}^{(l)}, \mathrm{X}_{\mathrm{kl}}^{(l)}$ from level $\mathrm{m} - 1$, and their parents, $\mathrm{X}_\mathrm{i}^{(l)}$, $\mathrm{X}_\mathrm{j}^{(l)}$, $\mathrm{X}_\mathrm{k}^{(l)}$, $\mathrm{X}_\mathrm{l}^{(l)}$, *activate* the significant node $\mathrm{X}_{\mathrm{ijkl}}^{(l)}$ if,

$$\mathrm{P}(\mathrm{X}_{\mathrm{ijkl}}^{(l)} |\mathrm{X}_{\mathrm{ij}}^{(l)}, \mathrm{X}_{\mathrm{kl}}^{(l)}) > \min \{\mathrm{X}_\mathrm{i}^{(l)}\mathrm{X}_\mathrm{j}^{(l)}),\ \mathrm{P}(\mathrm{X}_{\mathrm{kl}}^{(l)}|\mathrm{X}_\mathrm{k}^{(l)}\mathrm{X}_\mathrm{l}^{(l)})\}$$

or equivalently,

$$\frac{n_{ijkl}}{n_{ij} + n_{kl} - n_{ijkl}} > \min\{\frac{n_{ij}}{n_i + n_j - n_{ij}}, \frac{n_{kl}}{n_k + n_l - n_{kl}}\}$$

where,

n_{ijkl} = frequency of child node $X_{ijkl}^{(l)}$ of $X_{ij}^{(l)}$ and $X_{kl}^{(l)}$

n_{ij} = frequency of node $X_{ij}^{(l)}$

n_{kl} = frequency of node $X_{kl}^{(l)}$

n_i, n_j = frequency of node $X_i^{(l)}$, $X_j^{(l)}$, parents of node $X_{ij}^{(l)}$

n_k, n_l = frequency of node $X_k^{(l)}$, $X_l^{(l)}$, parents of node $X_{kl}^{(l)}$

Step 5
If there are no new active child nodes that can be formed, go to Step 6. Else, go to Step 2.

Step 6
If all active nodes that could not be linked have meaningful patterns, then STOP.

Or else, assemble all the active nodes that could not be linked from Level 1 to Level *m*, and let them become a set of *input nodes* in layer $l + 1$.

Set $l \leftarrow l + 1$, $m = 1$ and return to Step 1.

The algorithm begins by constructing a sequence of *N* single-element nodes from the training dataset. In every input level of the network build-up, child nodes at Level 2 that satisfy a simple Bayes factor test are considered as active nodes. They are then allowed to grow nodes in the next subsequent level. However for levels $m \geq 3$ onwards, propagation of new nodes are subject to two different phases of tests. Besides the Bayes factor test for identifying statistically significant nodes, nodes that are found to be statistically significant are further subjected to a conditional probabilistic test in order to activate them.

The network will continue to grow and propagate nodes until either one of the following scenarios occurs:

a) The number of levels *m* is equal to the length of the training dataset sequences;

b) All the unlinked active nodes generated by the algorithm have meaningful patterns of information.

In the event neither condition (a) nor (b) occurs, the algorithm is flexible enough to consider linking active nodes from different levels in order to produce meaningful patterns. This is easily achieved by assembling different levels of active nodes, and placing them as new layers of input nodes. Then continue growing nodes by following the steps given in the algorithm until the termination criteria has been met.

In addition, we will define the status of a node to be *strongly active* if,

- Bayes factor test < 1; and

- Phase 2 test requirement is fulfilled

and for a node to be *weakly active* if,

- Bayes factor test = 1; and
- Phase 2 test requirement is fulfilled

and for a node to be *non-active* if,

- Bayes factor test > 1; or
- fails Phase 2 test requirement.

Take note that for nodes at Level 2, we only need to utilise the result of the Bayes factor test to determine the status of a particular node, that is, a node is *strongly active* if,

- Bayes factor test < 1

and for a node to be *weakly active* if,

- Bayes factor = 1

and a node to be *non-active* if,

- Bayes factor > 1

The next section of this study is concerned with the application of Bayesian techniques in the neuronal network. With the introduction of suitable prior distributions, statistical inference of the network architecture can be conducted using simulation or numerical techniques.

Bayesian Modelling to Neuronal Networks

The application of Bayesian theory to neural networks provides an alternative approach to statistical inference, which is different to the classical interpretation using the frequentist approach. Theoretically, any statistical problem (survival analysis, regression models etc.) can be modelled using the Bayesian approach. The underlying philosophy of the Bayesian inference is that, the only sensible measure of uncertainty is probability. It assumes that data comes from one of a parameterised family of distributions. Parameters are to be treated as random variables and their distribution is known as *prior beliefs*.

Given such a scenario, a Bayesian statistical model consists of three parts:

i. A parametric statistical model for the data $\mathbf{x}$ given the parameter, $\theta \in \boldsymbol{\Theta}$, that is a family of distribution $f(\mathbf{x}|\theta)$, where $\mathbf{x}$ and θ may be multidimensional. The function $f(\mathbf{x}|\theta)$ is also known as the likelihood of data given the unknown parameter θ.

ii. Prior beliefs about θ are represented by a prior distribution $\pi(\theta)$, a probability density (or mass) function such that $\int_{\boldsymbol{\Theta}} \pi(\theta)d\theta = 1$ (or $\Sigma_{\theta}\pi(\theta) = 1$).

iii. A posterior distribution of θ, $\pi(\theta|x)$ as having the property of describing the uncertainty about θ after observing the data.

Part (i) is a common nomenclature in statistics. In Bayesian context, the likelihood is denoted by $f(x|\theta)$, since in Bayesian statistics we assume the data is dependent on the random variable θ. In the frequentist approach, the parameter θ is not thought of as random. The prior distribution $\pi(\theta)$, on the other hand, aims to summarise existing information about the parameter before the current data is observed.

In the light of the observed data **x**, the posterior distribution of θ represents our *modified* belief and the distribution of θ given **x** is defined as follows, and the Bayesian inference proceeds from this posterior distribution.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta}$$

Eq (6)

The first process is to fit a probability model to a set of observed data, and then, to summarise the result by a probability distribution on the parameters of the model, and if required, on unobserved quantities such as predictions. Because the denominator of the right-hand-side of **Eq (6)** is independent of θ, hence $\pi(x|\theta)$ is usually represented by,

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$

In the subjective world of Bayesian statistics, the prior distribution is intended to capture the information available about the parameter before the data is observed. In order to choose an appropriate prior for a class of likelihood function, we note that if $\pi(\theta|\alpha)$ is a prior for the likelihood $f(x|\theta)$, and if $\pi(\theta|x, \alpha)$ belongs to the same parametric family as $\pi(\theta|\alpha)$ then, we say $\pi(\theta|\alpha)$ is a conjugate prior for θ. The parameter α of the prior is often referred to as a hyper parameter.

Examples of natural conjugacies are:

Likelihood	Prior
Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Normal	Normal
Exponential	Gamma

In the context of neuronal networks, by assuming the training dataset, **x** follows a probability distribution $G(\theta)$, with a corresponding conjugate prior $\theta \sim H(\alpha)$, we can use Bayesian statistics to help us answer the following queries:

- The predictive probability of a child node under the test of hypothesis of mutual independence of its parents.
- The predictive conditional probability of the child node *Z* given by the observed values of its parents *X* and *Y,* that is we wish to find $P(Z|X, Y)$.
- The predictive probability that the child node *Z* is strongly active, weakly active or non-active given that its parents are active.

The following **Proposition 1** addresses the predictive probability of a child node in a *Phase 1* testing environment.

Proposition 1

Suppose under the Phase 1 hypothesis test, given the three random variables X, Y and Z of a nodal network structure in ***Figure 1*** *above, with each variable following a probability distribution* $G(\theta)$, *with* $\pi(\theta)$, *being a probability density function for the random variable* $\theta \in \Theta$. *The Bayes factor of testing the null hypothesis of mutual independence of nodes X and Y,* $X \perp\!\!\!\perp Y$ *against the alternative hypothesis of mutual dependence of nodes X and Y,* $X \not\perp\!\!\!\perp Y$ *is,*

$$B_{01}(Z) = \begin{cases} \displaystyle\int_{\Theta} \frac{f(x|\theta)f(y|\theta)}{f(x,y|\theta)}\,\pi(\theta|x,y)\,d\theta & \text{if } \theta \text{ is continuous} \\ \displaystyle\sum_{\text{all }\theta} \frac{f(x|\theta)f(y|\theta)}{f(x,y|\theta)}\,\pi(\theta|x,y) & \text{if } \theta \text{ is discrete} \end{cases}$$

where $f(.|\theta)$ *is the likelihood of data given* θ, *and* $\pi(\theta)$ *is the posterior probability of* θ *given the data.*

Proof

The proof for this result is quite straightforward and we only show the proof for the case when $\boldsymbol{\theta}$ follows a continuous probability distribution. Note that the proof is analogous with the continuous one for the case when $\boldsymbol{\theta}$ is discrete. From **Figure 1**, under the null hypothesis of mutual independence between *X* and *Y*, the predicted probability of node *Z* is.

$$\begin{aligned} P(Z|H_0) &= P(X,Y|X \perp Y) \\ &= \int_{\Theta} f(x,y|\theta)\pi(\theta)d\theta \\ &= \int_{\Theta} f(x|\theta)f(y|\theta)\pi(\theta)d\theta \end{aligned}$$

Under the alternative hypothesis where $X \not\perp\!\!\!\perp Y$, the predicted probability of node *Z* is,

$$\begin{aligned} P(Z|H_1) &= P\big(X,Y\big|X \not\perp\!\!\!\perp Y\big) \\ &= \int_{\Theta} f(x,y|\theta)\pi(\theta)d\theta \end{aligned}$$

Hence, the Bayes factor, $B_{01}(Z)$ can be written as,

$$\begin{aligned} B_{01}(Z) \quad &= \frac{\int_{\Theta} f(x|\theta)f(y|\theta)\pi(\theta)d\theta}{\int_{\Theta} f(x,y|\theta)\pi(\theta)d\theta} \\ &= \int_{\Theta} \frac{f(x|\theta)f(y|\theta)}{f(x,y|\theta)} \left\{ \frac{f(x,y|\theta)\pi(\theta)}{\int_{\Theta} f(x,y|\theta)\pi(\theta)d\theta}\, d\theta \right\} \\ &= \int_{\Theta} \frac{f(x|\theta)f(y|\theta)}{f(x,y|\theta)}\,\pi(\theta|x,y)d\theta \end{aligned}$$

The next result shows the predictive conditional probability of a child node given the observed information of its parents.

Proposition 2

*Suppose the three random variables X, Y and Z of a nodal network structure given in **Figure 1** have the properties $X \perp Y \mid Z$ and $X \not\perp\!\!\!\perp Y$, and each variable follows a probability distribution $\mathrm{G}(\theta)$ with $\pi(\theta)$, being a probability density function for the random variable $\theta \in$.*

Hence, the predictive conditional probability of Z given both of its parents X and Y is given as,

$$\mathrm{P}(\mathrm{Z}|\mathrm{X},\mathrm{Y})$$

$$= \begin{cases} \displaystyle\int_{\Theta} \frac{\mathrm{f}(\mathrm{z}|\mathrm{x},\theta)\mathrm{f}(\mathrm{z}|\mathrm{y},\theta)}{\mathrm{f}(\mathrm{x},\mathrm{y}|\theta)} \frac{\mathrm{f}(\mathrm{x}|\theta)\mathrm{f}(\mathrm{y}|\theta)}{\mathrm{f}(\mathrm{x},\mathrm{y}|\theta)} \pi(\theta|\mathrm{x},\mathrm{y})\, \mathrm{d}\theta \text{ if } \theta \text{ is continuous} \\ \displaystyle\sum_{\text{all } \theta} \frac{\mathrm{f}(\mathrm{z}|\mathrm{x},\theta)\mathrm{f}(\mathrm{z}|\mathrm{y},\theta)}{\mathrm{f}(\mathrm{z}|\theta)} \frac{\mathrm{f}(\mathrm{x}|\theta)\mathrm{f}(\mathrm{y}|\theta)}{\mathrm{f}(\mathrm{x},\mathrm{y}|\theta)} \pi(\theta|\mathrm{x},\mathrm{y}) \text{ if } \theta \text{ is discrete} \end{cases}$$

where $\mathrm{f}(.|\theta)$ *is the likelihood of data given* θ, *and* $\pi(\theta)$ *is the posterior probability of* θ *given the data.*

Proof

We only show the proof for the case when $\boldsymbol{\theta}$ follows a continuous probability distribution for the proof is analogous when $\boldsymbol{\theta}$ is discrete. From the definition of Bayes' rule

$$\mathrm{P}(\mathrm{Z}|\mathrm{X},\mathrm{Y}) = \frac{\mathrm{P}(\mathrm{Z},\mathrm{X},\mathrm{Y})}{\mathrm{P}(\mathrm{X},\mathrm{Y})}$$

Eq (7)

and since $X, Y, Z \sim G(\boldsymbol{\theta})$, hence the prior predictive distribution of the triplets (Z, X, Y) is,

$$\mathrm{P}(\mathrm{Z},\mathrm{X},\mathrm{Y}) = \int_{\Theta} \mathrm{f}(\mathrm{z},\mathrm{x},\mathrm{y}|\theta)\pi(\theta)\mathrm{d}\theta$$

Eq (8)

and similarly,

$$\mathrm{P}(\mathrm{X},\mathrm{Y}) = \int_{\Theta} \mathrm{f}(\mathrm{x},\mathrm{y}|\theta)\pi(\theta)\mathrm{d}\theta$$

Eq (9)

By substituting **Eq (8)** and **Eq (9)** into **Eq (7)** we have,

$$\mathrm{P}(\mathrm{Z}|\mathrm{X},\mathrm{Y}) = \frac{\int_{\Theta} \mathrm{f}(\mathrm{z},\mathrm{x},\mathrm{y}|\theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta} \mathrm{f}(\mathrm{x},\mathrm{y}|\theta)\pi(\theta)\mathrm{d}\theta}$$

Since $Z \not\perp\!\!\!\perp X$ and $Z \not\perp\!\!\!\perp Y$, the predictive conditional probability of Z given X and Y is,

$$\mathrm{P}(\mathrm{Z}|\mathrm{X},\mathrm{Y}) = \frac{\int_{\Theta} \mathrm{f}(\mathrm{z}|\mathrm{x},\mathrm{y},\theta)\mathrm{f}(\mathrm{x},\mathrm{y}|\theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta} \mathrm{f}(\mathrm{x},\mathrm{y}|\theta)\pi(\theta)\mathrm{d}\theta}$$

$$= \int_{\Theta} \mathrm{f(z|x,y,\theta)} \left\{ \frac{\mathrm{f(x,y|\theta)\pi(\theta)}}{\int_{\Theta} \mathrm{f(x,y|\theta)\pi(\theta)d\theta}} \right\} \mathrm{d\theta}$$

$$= \int_{\Theta} \mathrm{f(z|x,y,\theta)}\ \pi\mathrm{(\theta|x,y)d\theta}$$

Eq (10)

From Bayes' rule again, and because $X \perp Y|Z$, we can write,

$$\mathrm{f(z|x,y,\theta)} = \begin{cases} \dfrac{\mathrm{f(x,y|z,\theta)\ f(z|\theta)}}{\int_{\Theta} \mathrm{f(x,y|z,\theta)\ f(z|\theta)dz}} \text{if Z is continuous} \\ \dfrac{\mathrm{f(x,y|z,\theta)\ f(z|\theta)}}{\sum_{\text{all z}} \mathrm{f(x,y|z,\theta)f(z|\theta)}} \text{ Z is discrete} \end{cases}$$

$$= \frac{\mathrm{f(x,y|z,\theta)f(z|\theta)}}{\mathrm{f(x,y|\theta)}}$$

$$= \frac{\mathrm{f(x|z,\theta)f(y|z,\theta)f(z|\theta)}}{\mathrm{f(x,y|\theta)}}$$

$$= \frac{\dfrac{\mathrm{f(z|x,\theta)f(x|\theta)}}{\mathrm{f(z|\theta)}} \dfrac{\mathrm{f(z|y,\theta)f(y|\theta)}}{\mathrm{f(z|\theta)}} \mathrm{f(z|\theta)}}{\mathrm{f(x,y|\theta)}}$$

$$= \frac{\mathrm{f(z|x,\theta)f(z|y,\theta)}}{\mathrm{f(z|\theta)}} \frac{\mathrm{f(x|\theta)f(y|\theta)}}{\mathrm{f(x,y|\theta)}}$$

Eq (11)

By substituting **Eq (11)** into **Eq (10)**, we therefore have,

$$\mathrm{P(Z|X,Y)} = \int_{\Theta} \frac{\mathrm{f(z|x,\theta)f(z|y,\theta)}}{\mathrm{f(z|\theta)}} \frac{\mathrm{f(x|\theta)f(y|\theta)}}{\mathrm{f(x,y|\theta)}} \pi\mathrm{(\theta|x,y)d\theta}$$

Finally the probability of predicting the status of a child node whether it is *strongly active*, *weakly active* or *non-active* given its parents status is given in **Proposition 3**, where, this type of evaluation of probability is derived from *posterior predictive distribution* (see Gelman *et al.* [1] and, Leonard and Hsu [2]). Before providing the theoretical results, we have to define the status of *n* nodes in the network, $\mathrm{X1, X2, \dots, Xn}$ as an independent and identically distributed sample from the distribution $\mathrm{f(x|\theta)}$, where $\pi(\theta)$, $\theta \in \Theta$ be the prior distribution and $\pi(\theta|x)$, be the posterior distribution.

Let,

$$X_i = \begin{cases} -1 \text{ if node } i \text{ is non} - \text{active} \\ 0 \text{ is node } i \text{ is weakly active} \\ 1 \text{ is node } i \text{ is strongly active} \end{cases}$$

and we want to predict the distribution of $\mathrm{X_{n+1}|X_1, \dots, X_n}$ where $\mathrm{X_1, \dots, X_n}$ have already been observed. Hence, the posterior predictive distribution of $\mathrm{X_{n+1}|X_1, \dots, X_n}$ is,

$$f(x_{n+1}|x_1,\dots,x_n) = \begin{cases} \int_{\Theta} f(x_{n+1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|x_1 \dots, x_n)d\boldsymbol{\theta} & \text{if } \boldsymbol{\theta} \text{ is continuous} \\ \sum_{\text{all } \theta} f(x_{n+1}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}|x_1 \dots, x_n) & \text{if } \boldsymbol{\theta} \text{ is discrete} \end{cases}$$

From the variables set-up, we can infer that,

$$X_1 \dots, X_n, X_{n+1} \text{ Multinomial } (\theta)$$

where, we assume $X_1 \perp X_j, i \neq j$ and $\boldsymbol{\theta} = (\theta_{-1}, \theta_0, \theta_1)$, $\theta_{-1}, \theta_0, \theta_1 \geq 0, \theta_{-1} + \theta_0 + \theta_1 = 1$

Hence,

$$P(X_i = k|\theta_{-1}, \theta_0, \theta_1) = \theta_k, \quad k = -1, 0, 1$$

As for the choice of the prior, the standard conjugate prior θ for a multinomial likelihood is a Dirichlet probability density function where,

$$\theta \sim \text{Dirichlet}(\alpha)$$

such that $\alpha = (\alpha_{-1}, \alpha_0, \alpha_1), \alpha_{-1}, \alpha_0, \alpha_1 \geq 0$, and the α is known as the hyper parameter of the prior θ. Thus,

$$\pi(\theta_{-1}, \theta_0, \theta_1) = \frac{\Gamma(\alpha_{-1} + \alpha_0 + \alpha_1)}{\Gamma(\alpha_{-1})\Gamma(\alpha_0)\,\Gamma(\alpha_1)} \theta_{-1}^{\alpha_{-1}{}^{-1}}\ \theta_0^{\alpha_0{}^{-1}}\ \theta_1^{\alpha_1{}^{-1}}$$

where for $j = -1, 0, 1$

$$E(\theta_j) = \frac{\alpha_j}{\sum_{i=-1,0,1} \alpha_i}, \text{Var}(\theta_j) = \frac{(\sum_{i=-1,0,1} \alpha_i - \alpha_j)\alpha_j}{(\sum_{i=-1,0,1,} \alpha_i)^2 (\sum_{i=-1,0,1,} \alpha_i + 1)}$$

We now establish the posterior predictive distribution of $X_{n+1}|X_1, \dots X_n$ in the following Proposition 3.

Proposition 3
Let $X_1, \dots, X_n, X_{n+1}$ be an independent and identically distributed sample where each variable X_i

$$X_i \sim \text{Multinomial}(\theta_{-1}, \theta_0, \theta_1), \quad \theta_{-1} + \theta_0 + \theta_1 = 1$$

with the conjugate prior $\theta = (\theta_{-1}, \theta_0, \theta_1)$ *following,*

$$\theta_{-1}, \theta_0, \theta_1 \sim \text{Dirichlet}(\alpha_{-1}, \alpha_0, \alpha_1)$$

Hence the posterior distribution,

$$\theta | X_1, \dots, X_n \sim \text{Dirichlet}(n_{-1} + \alpha_{-1}, n_0 + \alpha_0, n_1 + \alpha_1)$$

and the predictive posterior distribution of $X_{n+1} | X_1, \dots X_n$ *is*

$$f\left(x_{n+1} = k \middle| x_{1,\dots,} x_n\right) = \frac{n_k + \alpha_k}{n + \alpha_{-1} + \alpha_0 + \alpha_1}$$

Where $k = -1, 0, 1, n_k = \sum_{i=1}^{n} I(x_i - k)$, such that

$$I(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$$

And $n = n_{-1} + n_0 + n_1$

Proof
To find the posterior distribution, for a variable $X_i = x_i, x_i = -1, 0$ or 1 we can write,

$$f(x_i | \theta) = \prod_{j=-1,0,1} \theta_j^{I(x_i - j)}$$

where,

$$I(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$$

Hence, from the definition of posterior distribution,

$$\pi(\theta | x_1, \dots, x_n) = \frac{f(x_1 \dots, x_n | \theta)\pi(\theta)}{\int_\Theta f(x_1 \dots, x_n \, | \theta)\pi(\theta)}$$

$$\propto f(x_1 \dots, x_n | \theta)\pi(\theta)$$

$$= f(x_1 | \theta) f(x_2 | \theta) \dots \, f(x_n | \theta)\pi(\theta)$$

$$= \prod_{i=1}^{n} f(x_i | \theta)\pi(\theta)$$

$$\propto \prod_{i=1}^{n} \left\{ \prod_{j=-1,0,1} \theta_j^{I(x_i-j)} \right\} \theta_{-1}^{\alpha-1^{-1}} \theta_0^{\alpha_0{}^{-1}} \theta_1^{\alpha_1{}^{-1}}$$

$$= \prod_{j=-1,0,1} \theta_j^{n_j+\alpha_j-1}$$

where, $n_j = \sum_{i=1}^{n} I(x_i - j)$. Therefore the posterior distribution follows a Dirichlet distribution where,

$$\theta | X_1 \ldots, X_n \sim \text{Dirichlet}(n_{-1} + \alpha_{-1}, n_0 + \alpha_0, n_1 + \alpha_1)$$

To evaluate the posterior predictive distribution of $X_{n+1}|X_1, \ldots, X_n$, we note that for $k = -1, 0$ or 1,

$$f\left(x_{n+1} = k \middle| x_{1,\ldots,} x_n\right) = \int_{\Theta} f(x_{n+1} = k|\boldsymbol{\theta})\pi(\theta|x_1, \ldots, x_n)\ d\boldsymbol{\theta}$$

$$= \int_{\Theta} \theta_k \pi(\boldsymbol{\theta}|x_1, \ldots, x_n) d\boldsymbol{\theta}$$

$$= E(\Theta_k|X_1, \ldots, X_n)$$

$$= \frac{n_k + \alpha_k}{\sum_{i=-1,0,1}(n_i + \alpha_i)}$$

$$= \frac{n_k + \alpha_k}{n + \alpha_{-1} + \alpha_0 + \alpha_1}$$

With the construction of appropriate Bayesian models with suitable priors, the next stage is to try to compute the posterior distributions, and then subsequently make statistical inference based on the parameter estimates given the data set.

Conclusion

In this report, we presented a Bayesian analysis of the NeuraBASE neuronal network model from a Bayesian perspective. From the Bayesian point of view, the neuronal network's strength lies in its ability to associate different information from many sources, and hence, allows greater "objectivity" in final conclusions.

The Bayesian statistical relationship between the parent and child nodes of the neuronal network structure can also be obtained by selecting appropriate likelihood and prior distributions. This is confirmed by the theoretical results presented, which addresses the various predictive probabilities of a child node in relation with its parents.

These demonstrate that the neuronal network with its implicit statistical structure present in the nodes can exploit the subjective world of Bayesian statistics to provide an alternate viewpoint in the way we model the neuronal network.

References

[1] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, Bayesian Data Analysis, Texts in Statistical Science, Chapman and Hall, London, New York, (1995).

[2] Springer Series in Statistics, Springer, New York, (2001).

[3] T. Leonard and J.S.J. Hsu, Bayesian Methods, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, UK, (1999).

[4] D.J.C. MacKay, *Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks*, Bayesian methods for backpropagation networks, Springer, (1995).

[5] P. Müller and D.R. Insua, *Issues in Bayesian analysis of neural network models*, Neural Computation, 10, pp. 571{592, (1995).

[6] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge, United Kingdom, (1996).