# Chapter 8:

# Bayesian Inference and Models in AP

**CHAPTER 8 Contents:**

[8.1. Bayesian Inference and Models in Artificial Psychology 190](#_Toc125225858)

[8.2. A critique on the use of P-value 191](#_Toc125225859)

[8.3. Bayesian independent 2 samples t-test 194](#_Toc125225860)

[8.4. Naïve Bayes Classifier 195](#_Toc125225861)

[8.5. Cross validation 197](#_Toc125225862)

[8.6. Bayesian Binary Logistic Regression 201](#_Toc125225863)

[8.7. Bayesian Network Analysis 204](#_Toc125225864)

[8.8. Bayesian Model Averaging 208](#_Toc125225865)



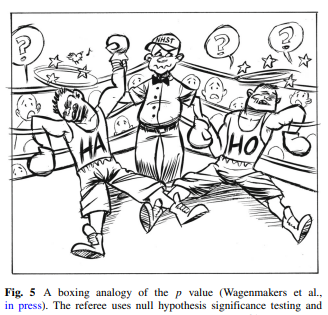
**Bayesian Inference**

Abstract: We begin with Bayes theorem which combines apriori information with information from the data to create a posterior conditional probability such as the probability of a patient having a severe memory impairment given their gender. We then compare this approach with classical Fisherian inference giving Bayesian analogues based on summarising the posterior distributions of estimates using percentiles to give medians and credible regions. Other Bayesian diagnostics which are used in inference based upon posterior distributions of estimates are also defined. These include Maximum posterior estimates (MAP), the probability of direction (Pd), highest density regions (HDI) and Bayes factors (BF). The use of these diagnostics in R is illustrated on an example of introversion and extroversion in students. A family of machine learning models, naïve Bayes, are then presented and illustrated in R. These models are an extension of the classical logistic regression model to incorporate apriori information into classification. The use of cross-validation in unbiased model assessment is presented and illustrated via k fold cross validation which splits the data into design and test sets. Finally theory underlying the use of Bayesian linear regression models is illustrated in R using Bayesian model averaging (BMA) which summarises regression estimates based upon sets of models which each use different subsets of the features.

Keywords: Bayesian inference; Bayesian logistic regression; Bayesian linear regression; Naïve Bayes; Bayesian model averaging; Posterior distributions; Cross-validation

## Bayesian Inference and Models in Artificial Psychology

Bayesian inference includes Bayesian parameter estimation and Bayesian hypothesis testing. In recent years, this approach has been proposed as an attractive alternative to estimation and hypothesis testing in classical statistics.

In classic Fisherian or Frequentist statistics, confidence interval and p-value are used for estimation and hypothesis testing. (Wagenmakers, et al., 2018).

The background of psychological researches is full of p-value reports (Farahani et al., 2020). It can be said that the use of p-values causes a crisis of confidence in the results of psychological researches that are mainly conducted using classical statistics, in other words, p-value hacks the results of psychological research. The frequent use of p-value in Null Hypothesis Statistical Testing (NHST) has been seriously criticized by a large number of researchers repeatedly from different points of view (Wagenmakers et al., 2018).

1. a boxing analogy of the p value from Wagenmakers et al,2018).

Likelihood

Class prior probability

Predictor prior probability

Posterior probability

**(8.1)**

## Bayesian Statistics in a Nutshell

A quick look at Bayesian statistics may help ease the concepts of this section and other sections of this chapter. Suppose an artificial psychologist examines 100 boys with ADHD and finds that 43 of them are the first child in the family. Therefore, the probability that a boy is the first child in the family is equal to: %

Here we can talk about two other terms in probability and they are dependent and independent probabilities. If the occurrence or non-occurrence of a phenomenon has no effect on the occurrence or non-occurrence of another phenomenon, then the two phenomena are independent and , for example, the probability that a child with ADHD is the first child in the family and the probability of catching a goldfish How big is a river in Hawaii? These two phenomena are completely independent of each other because catching a large goldfish in the Hanalei River in Hawaii has no effect on the probability of a child having ADHD as a first child, and vice versa. If the probability of catching a large goldfish in that river in Hawaii is 10%, then the two probabilities must be multiplied together to calculate the coincidence.

**(8.2)**

If two phenomena are supposed to be dependent, then a different approach is used to calculate the probability. Pay attention to this example: suppose that the artificial psychologist examines 100 patients with MS, he also examines the memory and gender of these 100 people. The results are shown in the table below.

1. The patients’ memory problems based on sex in MS patients

|  |  |  |
| --- | --- | --- |
| Male | Female |  |
| 12 | 8 | severe |
|  |  | Memory problems |
| 48 | 32 | mild |

Based on this table, P = 0.2 12/60 (A │ B) = in which (A = severe, B = male) can be divided into the intersection of problems, having severe memory of A and the gender of the patient being B being male.

**(8.3)**

Similarly:

**(8.4)**

Bayes' rule, which is used for NB, can be considered as the result of dividing these two expressions. Therefore, Bayes' rule can be considered as follows:

**(8.5)**

In this formula, P(Evidence │ outcome) is obtained from training data.

**(8.6)**

In this formula, P (outcome │ Evidence) is checked, which is used to predict test data.

**Bayes Rule:**

**(8.7)**

**(8.8)**

In the above formula, the left part is called the Posterior Probability or simply posterior, and the first term is called likelihood of evidence, which is actually the conditional probability for a particular class, and provided that the predicted variables are independent, all of them can be multiplied. The likelihood value is determined from the training data for Y=c (a specific group) and the second part, the prior, shows the overall probability of y=c where c is a class of Y.

**(8.9)**

With this introduction, you can get a general understanding of the necessary terms in NB. Terms such as conditional probability, Bayes' rule, independent and dependent probability, and posterior, prior, and likelihood.

The most important design parameter in NB is the smoothing method. The idea of smoothing goes back to the efforts called Cromwell's rule, and based on that, if the estimate of a probability is equal to zero, it should not be used in probabilistic reasoning, because as discussed, to combine the probabilities, we multiply them together and, so, if one of them is zero, regardless of the probabilities of the other variables, it will be zero. The most common form of smoothing is called Laplace smoothing, in which the number of desirable cases (K) out of n trial attempts is considered as the desirable ratio ((k+1))/((n+1)) and not as k/ n.

Classical statisticians consider smoothing as a form of regularization and Bayesian statisticians consider smoothing as a prior.

## A critique on the use of p-value

As stated, the classic or Fisherian statistic relies on the central core of the p-value in the test hypothesis. P-values are easy to obtain using routine software such as SPSS. Interpreting the p-value, however, is challenging. Concluding that p<0.05 guarantees the rejection of the null hypothesis (H0) and thus supports the acceptance of the alternative hypothesis (H1) is a misinterpretation. Let’s get a little more specific. The p-value indicates the probability of obtaining a result at least as large as the observed result, provided that the null hypothesis is correct. Therefore, p-value cannot recognize the fact that the data that are unusual under H0 can also be unusual under H1. However, the p-value is still of interest to psychologists. Some of the reasons for this interest is that most psychologists, like other people, are addicted to their own beliefs, so they tend to teach others what they have learned and do not take steps to change their statistical knowledge. In addition, it seems incorrectly, that the interpretation of p-value and p<0.05 is enough to reject H0 and confirm H1. Psychologists may also worry about reducing their chances of publishing their research articles, if they use new methods. Perhaps these reasons are the general reasons why there is resistance to new statistical methods alternatives to the p-value (Sharpe, 2013).

This book aims to break down this resistance. To overcome the weaknesses of the p-value, researchers have made efforts to replace it. One of these attempts is to replace the confidence interval (CI) with the p-value. The confidence interval has also been criticized, such as the fact that it considers the real value to be estimated as a fixed value.

The 95% confidence interval (CI) for an effect shows that if the confidence interval is calculated repeatedly from the data, there is a 95% probability that the desired effect or parameter is in a given range. This interpretation is somewhat counter-intuitive. In Bayesian statistics, similar to the confidence interval, there is a credible interval, which shows that according to the observed data, there is a 95% probability that the desired effect falls within this domain. By examining the background of the research done on the benefits of Bayesian inference, it can be said that this method can be more useful with high dimensional data than the classical statistical method, and the information is more reliable (Etz, 2016). Bayesian approaches are more accurate in conditions where there is noisy data and there are small samples (Kruschke, J. K., Aguinis, H., & Joo, H. 2012). Bayesian inference provides two possibilities for combining prior knowledge in the final analysis. (Andrews & Baguley, 2013, Kruschke, et al, 2012 ) .Bayesian analysis also gives straightforward, intuitive results (Kruschke,2018; Wagenmakers, et al, 2018). Bayesian inference includes a measure of evidence that the data additionally provides in favor of H0 versus H1, the Bayes factor, which unlike the p-value, does not have a serious bias, against H0 (Edwards, 1965; Sellke et al, 2001)

In summary, it can be said that the main focus of classical statistics compared to Bayesian statistics is that classical statistics is strongly focused on the statistical test of the null hypothesis (NHST) and the misinterpretation of the p-value, and this extreme focus causes serious criticisms and accompanying lack of trust in the results of psychological research.

The theoretical framework of Bayesian statistics is based on Bayes theory. In this book, our goal is not to focus on the theoretical basis which the reader can follow and study elsewhere. What is considered in the theoretical framework of Bayesian statistics is different from what is considered in the theoretical framework of classical statistics. In the theoretical framework of classical statistics, the focus is on hypothesis testing and the p-value, which assumes that the effects are fixed, and unknown and that the data are random. That is, it is assumed that the unknown parameter is a unique value that the researcher tries to estimate with statistical methods using the data obtained from the sample. In the framework of Bayesian statistics, the true effect is not estimated, but instead the probability of different effects is calculated according to the data obtained from the sample, which itself leads to the posterior distribution, a distribution of possible values for the parameters.

In Bayesian statistics, indicators such as the median of the posterior distribution and the range of values of that distribution that includes 95% of the most probable values, the 95% credible interval, are calculated to model uncertainty in an estimated parameter. In classical statistics, point estimation and confidence interval are calculated.

In Bayesian statistics, using Bayesian sampling algorithms, which will be explained later, a possible (posterior) distribution is obtained from an effect that is compatible with the observed data.

It follows that, in Bayesian analysis, based on the resulting data and sometimes the prior belief or distribution about the results using Bayesian sampling algorithms a possible distribution is produced called the posterior. For example, suppose we assume that the correlation between affective metallization and mental health is equal to 0.54 in a sample of 100 people. Based on this distribution, Bayesian analysis tells us that the most probable effect (correlation) is 0.54, but the data is consistent with the correlations of 0.74 to 0.85, each of which has specific probabilities. To determine the significance of an effect in Bayesian analysis, a p-value is not needed, despite it being commonly used in classical statistics, with it, instead, being sufficient to describe the posterior distribution of that effect. One of these very important indicators is the credible interval. The credible interval is a key concept in Bayesian inference and analysis, whose purpose is to provide a summary of the uncertainty related to the estimated parameter. A credible interval in Bayesian statistics is a range of the posterior distribution that includes the possible magnitudes of the investigated effect with certain probabilities. Instead of a 95% confidence interval that is common in classical statistics in Bayesian statistics, McElreath (2020) suggested using a threshold of 89% when specifying ranges of a credible region.

Kruschke in 2014 states that the credible interval with 89% coverage is more stable than one with 95% and that to calculate the credible interval with a 95% interval, it is necessary that the number of samples in Bayesian sampling is at least 10000, which is routinely the default number. Posterior samples that are used in most Bayesian statistics software packages, on the other hand, usually only use 4000 samples.

In Bayesian analysis, just like classical statistical analysis, there are descriptive and analytical indicators that should be taken into account when reporting the results.

The median, in general, is a more robust index compared to the mean.

MAP is the maximum posterior probability estimate (MAP). The MAP index in a posterior distribution indicates the value that has the highest probability. The peak of the posterior distributioncan be considered the mode of the posterior distribution. The median is more robust compared to this index, but if the distribution has extreme skewness. MAP is more appropriate than the median (Markowski, et al, 2019).

95% or 98% credible intervals (CI) also indicate uncertainty. Generally, a CI is calculated based on the highest density interval (HDI). The highest density interval (HDI) produces an interval containing values which have the highest probability density and always contains the most likely value of the parameter value corresponding to the mode.

**Significance or existence of a network:**

In Bayesian analysis, there is no p-value index, but there is a more interpretable index that is more straightforward. It is interpreted as describing the existence of an effect. This probability index is called the probability of direction (Pd) and indicates the most likely direction (positive or negative) of an effect. In the interpretation of this index, it is possible to have a cutoff point like a p-value. A Pd > 97% is indicative of a likely effect, a Pd > 99% suggests an effect probably exists and a pd > 99.9% indicates the effect exists with certainty. Region of Practical Equivalence (Rope) is a region that signifies values of a parameter estimate, such as amount of change in an outcome, corresponding to denoting practically no effect. This indicator shows whether or not a parameter is related to, for example, a non-negligible change in the outcome.

This index is a continuous index of significance. It can be said that if Rope covers 99% of the highest density region (HDI), ie Rope covers most of the credible values then H0 can be accepted. If Rope covers 97.5% of the HDI then the chance of the null hypothesis being rejected is probably negligible. If rope covers between 2.5 and 97.5 of HDI no conclusion can be made about significance., If Rope covers less than 2.5% of the HDI it is probably safe to reject the null hypothesis whereas if Rope covers less than 1% of HDI H0 can be rejected.

The Bayes factor (BF) index is a multipurpose index that can be used to compare different models. BF is a ratio that gives us information about the probability of the observed data under the two compared models (model with effect) versus (model without effect). Its interpretation depends upon whether the BF is the ratio of the posterior probability of the model with the effect to the model without the effect or vice-versa.

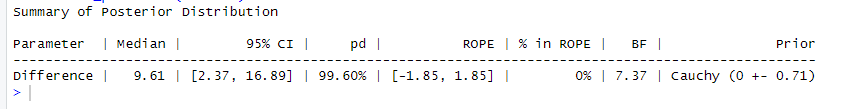
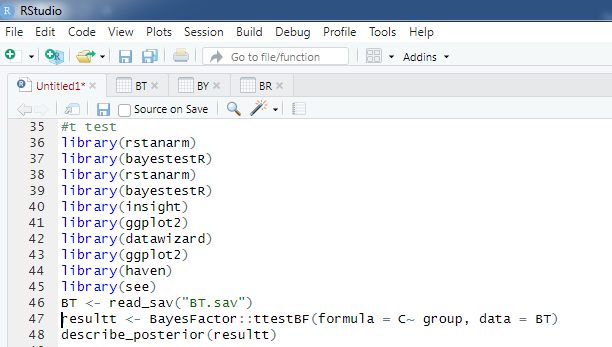
BF can be used both in the context of significance and in the context of the existence of an effect. To interpret BF based on Jeffreys (1961) criterion, which is used by default in R packages, it can be said that 3<BF≤10 indicates moderate evidence, 10<BF≤30 strong evidence, 30<BF≤100 very strong evidence and a BF>100 represents extreme evidence.

### Practical example using R

An artificial school psychologist wants to know whether the level of calm continuity of introverted (Class=1) and extroverted (Class=2) students is different. In a test, he measures the amount of calm persistence (the patience of students when they are faced with very difficult questions in an exam). By using a t-test, two independent samples are compared with Bayesian statistics. The relevant codes are shown in the figure 8.2.

The results of the Bayesian t analysis are given in the table. As can be seen, the median as the central index of the posterior distribution is equal to 12.5. Group membership (extroversion-introversion) shows the median association is positive with a probability of 1. (Pd=100).

1. R codes of Bayesian T test



1. Summary of Bayesian t-test

Rope, supports this with the region corresponding to values associated with a median of zero of 0% showing it covers a negligibly small amount of possible values. Observation of BF shows moderate evidence in favor of introversion being associated with calm continuity (BF=7.37).

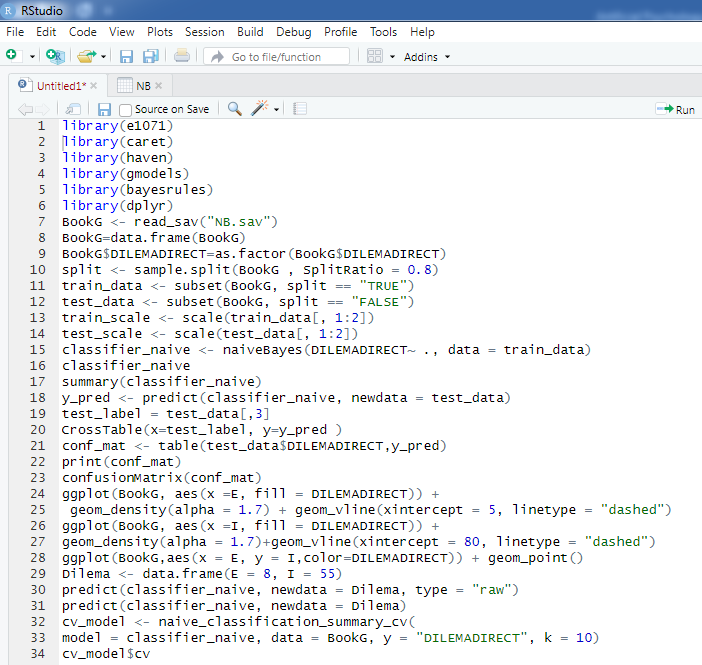
## Naïve Bayes Classifier

Naïve Bayes models are a family of machine learning models which attempt to classify data into groups allocating to the group with the higher or highest posterior probability of group membership based on a set of characteristics. These models utilize distributions such as the Bernoulli (classification into two groups) and Multinomial (classification into more than two groups).

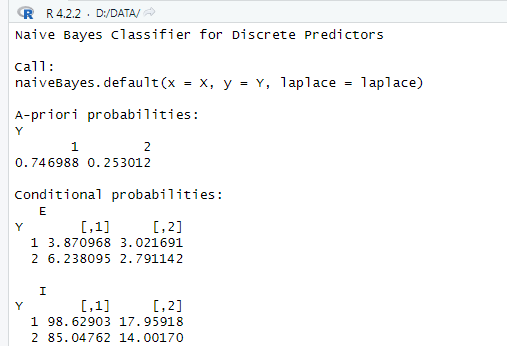
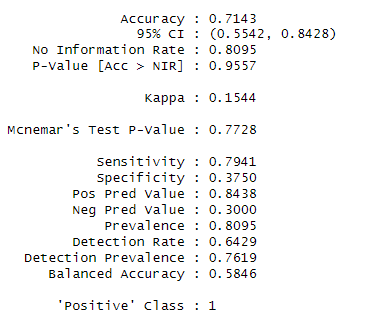
We use a cross-validation approach using 80% of the data to obtain the classification model and the remaining 20% to assess how well it can assign new data into groups.

An artificial cognitive psychologist tries to check the probability of utilitarian-duty-oriented group membership with NB based on emotion regulation (E) and (I) introversion and to know how likely it is that an individual, assuming utilitarianism (1) has high emotion regulation and high introversion. It was already mentioned that the number of predictor variables should be large and independent. Here, for illustrative purposes, the number of predictor variables is 2 and they are assumed to be independent of each other. In this research, he examines 125 people and, based on a cognitive task, separates 94 people as utilitarian and 31 people as task oriented. He also uses the Laplace method for smoothing (R codes are shown in the figure). 80% of the sample is selected as the training sample and 20% as the test sample. The prior for this analysis was equal to 75% (utility oriented) and 25% (task oriented) based on the original data.

1. R codes of Naïve Bayesian Classifier

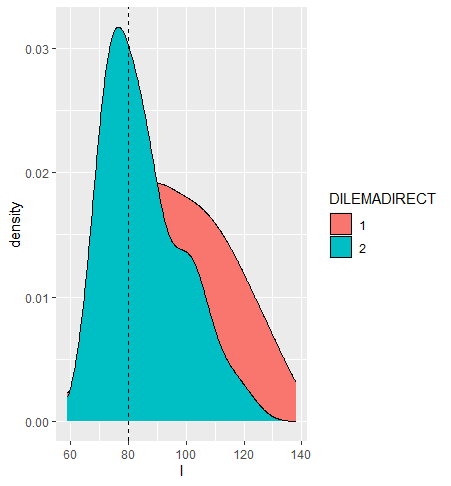
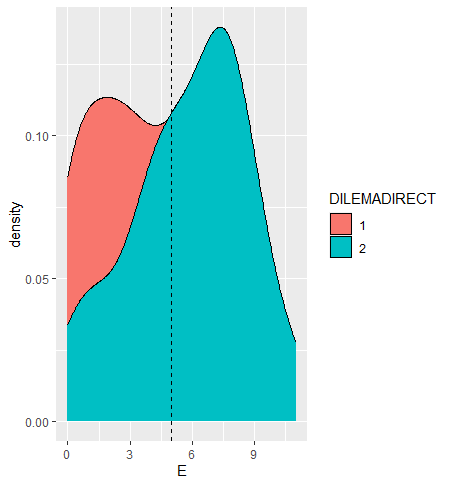


The results showed that the accuracy of this analysis is 71% and the Kappa coefficient is equal to 0.15, which is a small value. Of course, this is just an example. This analysis has a sensitivity and specificity of 0.79 and 0.38, respectively. Based on the resulting model, which is also validated with 10-fold cross-validation, for a person whose emotional regulation is 8 and whose introversion is 55, the predicted probability of his group membership equals 0.66 for the utilitarian group and 0.34 for the duty-oriented group(Figure 8.3).



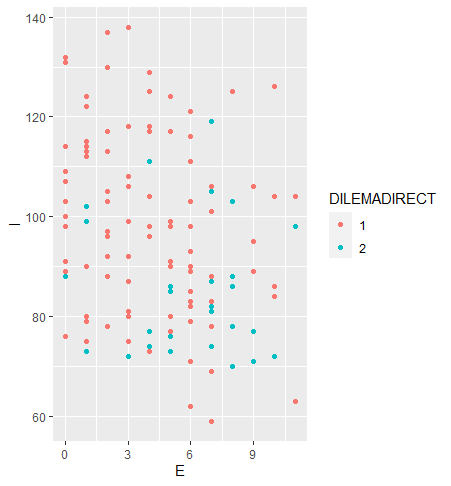
1. Summary of Naïve Bayesian Classifier

Plots in Figure 8.4 show density, variables E and I in two utilitarian and task oriented groups based on Posterior distribution.



1. Density of posterior of E and I variables.

The last diagram shows the shape of the relationship between E and I, separated by the duty-oriented and utilitarian groups. The utilitarian group is marked with small red circles and the duty-oriented group with small blue circles. It should be noted that here the aim is only to use NB in classification. Obviously, the sample size is not large enough and more data is needed (Figure8.5).



1. Relationships of E and I .

## Cross validation

The Accuracy of the model is 71.14 percent, the kappa, comparing predicted group membership to actual group membership, is weak and McNemar’s test is not significant suggesting a poor rate of classification.

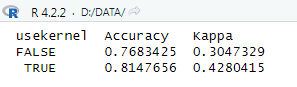
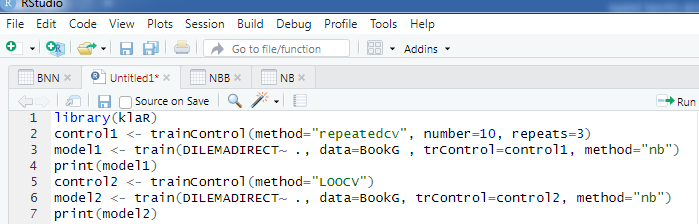
Although the hold-out method is very important in obtaining a valid model, it may lead to over-fitting or under-fitting, therefore, we should use some other methods for this purpose.

**Repeated K-fold Cross-Validation**

The process of splitting the data into k folds can be repeated a number of times. This is when the data is split into k groups with k-1 groups used to estimate the machine learning model and the remaining group used to assess its predictive accuracy. This is repeated for each of the k groups and called repeated k-fold cross-validation. The final model accuracy is the mean of the number of repeats.

### 8.6.1. A Practical example using R

1. R codes of Naïve Bayesian Classifier(NBC) using CV and LOOCV

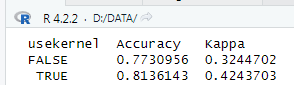


1. The R output of NBC using CV

The mean accuracy for this model is 76.84 percent. The kappa coefficient comparing predicted group with the actual group is low (Figure 8.6).

**Leave-One-Out Cross-Validation**

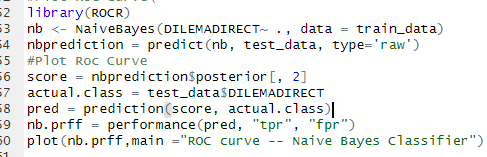
Leave-one-out cross-validation, or LOOCV, is the cross-validation technique in which the size of the fold is “1” with “k” being set to the number of observations in the data. This validation is useful when the training data is of limited size and the number of parameters to be tested is not high(Figure 8.7).

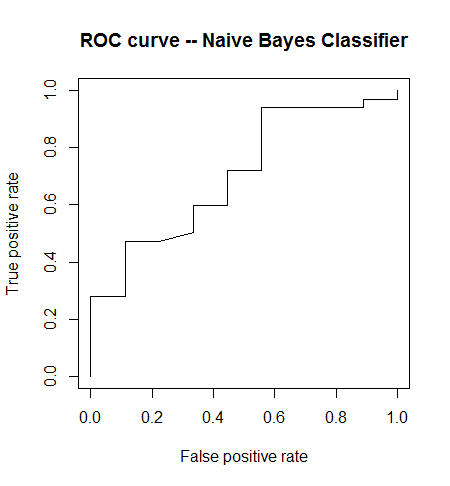


1. The R output of NBC using LOOCV

The mean accuracy for this model is 77.31 percent. This accuracy in Holdout Validation Approach is 71.14 percent.

1. R codes of ROC curve of NBC





1. The ROC curve for NBC

As figure 8.8 shows the curve is above the 50 percent and it means the classification accuracy is more than chance. Good classification is shown in a ROC curve which goes up to the top left hand corner which is not the case in this example.

## Bayesian Binary Logistic Regression

We can use prior information for model parameters in a logistic regression specified in the example below using the user.prior.density subcommand in the Bayesian logistic regression program, MCMClogit, in R. These prior distributions are then incorporated with information from the raw data to yield posterior distributions for model parameters. The fitting of these models in R is performed using sampling from posterior distributions (Monte Carlo) based upon Markov Chains which can be viewed as a network of paths where each path relates the parameters (nodes) in the model to one another and to the outcome variable.

Bayesian regression and Bayesian logistic regression are based on Bayes theory. A theory that contains Conditional Probability in its core. Regardless of the mathematical basis and attention to statistical details, the most important distinction between traditional Maximum Likelihood models such as logistic regression and models based on Bayes theory such as Bayesian regression can be summarized as follows.

1) The regression models have slope, intercept and sigma parameters and each parameter has an associated prior.

2) The estimated parameters have a normal distribution, while in the classic or frequentist-based models, the estimated parameters are fixed and have a probability distribution function based on the same probability distribution function.

3) In logistic regression, each parameter is separate and described by a different distribution.

4) In Bayesian regression analysis, the posterior distribution is made from the prior distribution and likelihood. The mean or other central indicators of a posterior distribution are considered as the coefficient of interest of the variable under consideration.

5) In logistic regression analysis, when the posterior distribution is highly skewed or bimodal or multimodal, equal tailed credible sets are used, which are defined as the outer 0.025 quantiles of the posterior distribution.

Credible confidence indicates that there is a 95% probability that this Credible confidence interval includes the posterior mean or the true posterior mean of the posterior distribution.

6) The last main difference is the existence of additional or prior information. The posterior distribution that is defined to estimate the parameters in Bayesian analysis can be combined with additional or initial information, which is our initial knowledge about the variable or parameter, which is separate from the data used in the analysis.

The main formula used in Bayesian analysis is as follows:

**(8.10)**

In this formula, p(x │θ) is the likelihood function and p(θ) represents the prior distribution. The denominator of the fraction means that p (x) is the probability of x in all x.

As can be seen in the case of subtraction, the probability and initial distributions are multiplied together. Usually, the denominator of the fraction, which is the normalization term, is left out of the calculations in such a way that the posterior distribution or a predictive model includes the product of likelihood and prior. As mentioned, each posterior variable between predictors can have its own posterior distribution. If the artificial psychologist finds that there is no important information outside the obtained data that has an effect on the prior variables, then a uniform prior is usually considered and in this case it is said that the prior in the analysis is non-informative.

It should be noted that if we consider the prior distribution as having a normal distribution with a mean of zero and a very high variance, Bayesian analysis is non-informative and if all predictor variables in the model are non-informative, then the result is traditional logistic regression based on maximum likelihood which will give the same, or very nearly the same, results as using Bayesian regression. In fact, the purpose of the prior distribution is to reflect the information that does not exist in the existing data from the sample, and therefore, if the prior information is weak, this does not have much effect on the Bayesian analysis.

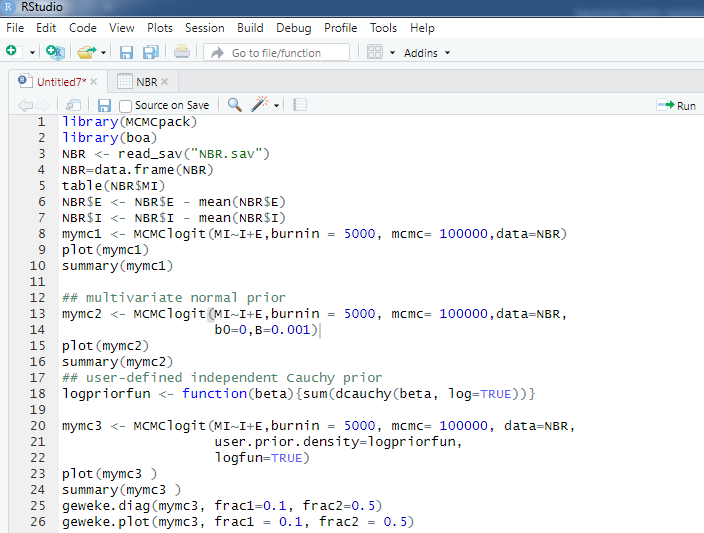
It should be noted that the prior information can not only provide significant quantitative information, but represents a distribution with parameters that is combined with the posterior probability distribution. Markov Chain Monte Carlo (MCMC) methods are used to obtain posterior results. MCMC is a set of algorithms that are used for various purposes such as optimization, dynamic simulation, and sampling. Due to ease of implementation and numerical stability, statisticians prefer the MCMC method, although some consider it a black box of sampling and posterior estimation (Brooks, 2011). After the popularization of Bayesian methods in applied problems in the 1990s, the main idea of creating approximate samples from the posterior distribution of interest was expanded by the Markov chain.

MCMC has two traditional approaches which are called Gibbs sampling (GS) and the Metropolis-Hastings (MH) algorithm. It can be said that the Gibbs sampling algorithm is a special case of MH and can only be used in some conditions such as when the discrete distribution is discrete or normal, while the MH algorithm is used in a wide range of distributions and is based on the possible candidate values of the proposed sample distribution. To achieve a valid inference from the posterior distribution using MCMC, the MCMC chain must converge.

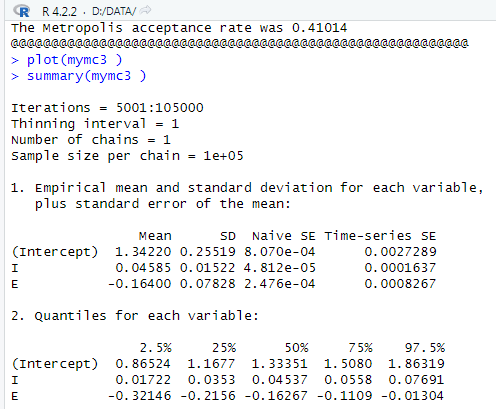
### 8.7.1. A Practical example using R

An artificial psychologist tries to predict the probability of marital infidelity based on emotion regulation (E) and cyber introversion (I). He selects a sample of 125 people including two groups (a group with experience of marital infidelity (31 people) and a group of 94 people without experience) and questionnaires on emotion regulation (E) and virtual introversion (I) implemented on them, then he decides to determine the prior distribution by examining the distribution of the variables. He concludes that Cauchy considers the prior distribution of these two variables. This distribution is similar to the normal distribution. (listing 5).

1. R codes of Bayesian logistic Regression



In this example, the result is interpreted with Cauchy's prior distribution. The prior distribution can be considered in a similar way to a normal or t distribution. The result showed that the two variables of emotion regulation and cyber introversion significantly predict group membership and 95% internal credibility is significant for both variables (Figure 8.9). The R codes are in listing 5.



1. The R output of Bayesian logistic regression

The above figure gives the estimated parameter values from the 2.5% to 97.5% quantiles of their posterior distributions providing a 95% credibility interval for each variable. Parameters of Introversion(I), and Extroversion(E) are 0.04585 and -0.16400 respectively (Figure 8.9) which are significant at the 5% significance level in that their 95% credible regions do not contain zero which would indicate no relationship.

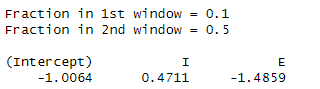
This index is a proportion of the posterior distribution that has the same sign as its median. This index is very similar to a p-value in classical statistics and is known as the Maximum Probability of Effect (MPE).This method is especially useful for models in which there are a large number of variables and categorical input variables have a large number of states and values. This is a simple and interesting method that memorizes how each variable in the training phase is related to the outcome. and then makes a prediction by multiplying the effects of each variable.

To easily understand this method, let's consider a non-statistical and intuitive example, suppose we want to predict whether a person is psychologically healthy based on his education level, attachment style, and affective mentalization. In Naive Bayes, this logic is reversed and the question is asked that if a person is psychologically healthy, what is the probability that his emotional metallization is healthy and his attachment is secure? .

In short, the question here is that based on the features we have, what is the probability that a person with these features is in a certain group? Naive Bayes answers this question under a very bold naïve assumption, and that assumption is the one that says that all predictor variables are independent of each other. Of course, in the real world such an assumption is difficult to verify, using NB significantly reduces the complexity of the model..

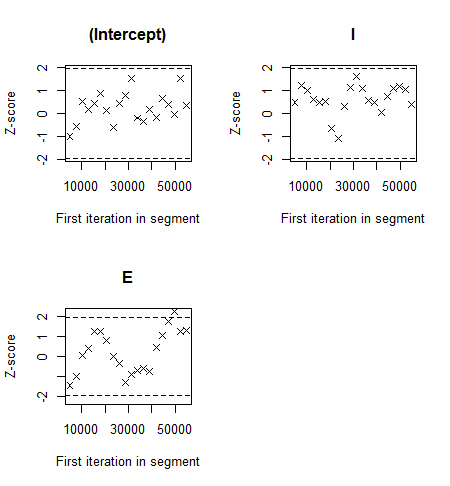
As we know, NB is also based on Bayes law. Bayesian probability is discussed in different parts in this chapter. In short, Bayesian statistics is based on the product of probabilities.

It should be noted that the choice of statistical inference method is not a matter of taste, but a direct result of the problem that the artificial psychologist seeks to answer. If he is worried about the sensitivity of his statistical results to variation in data and modeling procedures he should use frequentist statistics but if he is worried about the sensitivity of his statistical findings to possible variation in the unknown quantity that needs to be modeled, he should take steps in the framework of Bayesian statistics.

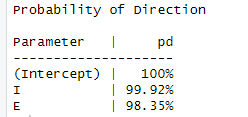


1. Z values of the Geweke diagnostic comparing 2 fractions of the data

The artificial psychologist used the dCauchy function. d is the density indicator.. The Geweke diagnostic can be used to assess if the parameter estimates have converged and it compares the values at the start of the repeated sampling from the posterior distributions with those at the end with a z value less than 2 in absolute value indicating no difference. If these values do not differ the parameter estimates are stable and are deemed to have converged. The output shows that the regression slopes for both the variables of Introversion(I), and Extroversion(E), have |z| < 2. (Figure 8. 10). Z values for Geweke are less than 2, indicating that the MCMC posterior distributions of these estimates have converged.

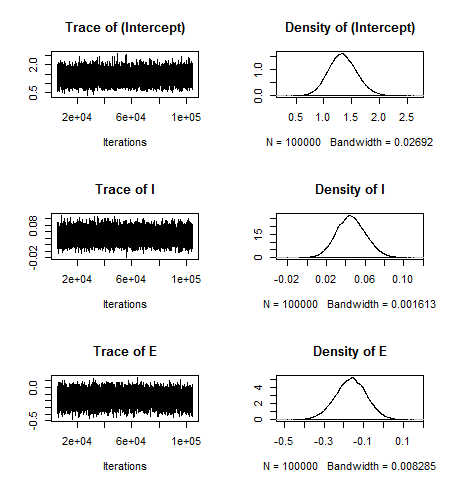


1. Segment plot of iterations



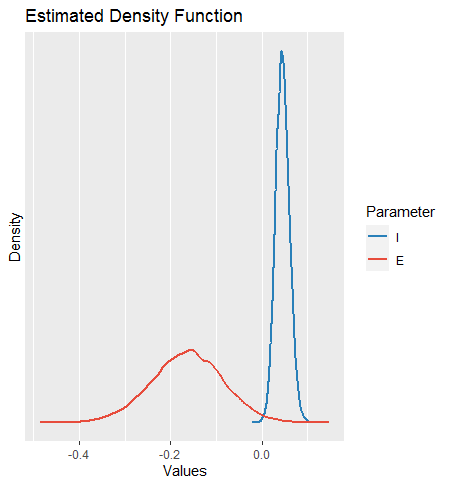
1. PD of the predictors

The probability direction(pd) for cyber introversion is 99.92% and for emotion regulation is 98.35%. (Figure8.12).





1. Trace plots of the corresponding posterior estimates of the Intercept, variables via



1. Posterior plot of I and E

## Bayesian Network Analysis

A Bayesian network model is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a graph. The edges in the graph represent the direction and degree of relationship between pairs of variables (nodes). The outputted network is averaged over all possible networks linking the variables specified as being in the network.

The most important concepts of network analysis as an advanced method are presented in the chapter on network analysis in this book. In this chapter network analysis using Bayesian statistics and its application is presented.

A Bayesian network is a combination of a network structure, specifically a directed acyclic graph (DAG) and a probability distribution associated with it. Therefore, Bayesian networks are graphical methods or models that examine Nodes with Edges in a graphical structure.

Nodes can be considered psychological variables, symptoms of mental disorders, etc., linked by Edges or Arrows indicating probabilistic dependencies. (Kroba & Nicholson, 2004)

The graphic structure G=(V,A) is a Bayesian network of a DAG, where V is a node and A is an arc (or edge). DAG can be considered as a factorization of a joint probability distribution (nodes). In simpler terms, based on the definition of Jensen et al. (2001), the Bayesian network is characterized by 3 components: 1) nodes (which can be an infinite state), 2) a directed edge (which links two nodes) and 3) a conditional probability for each variable or node.

Nodes and edges define the structure of a network and here the structure of the Bayesian network. The direction of the edges in the form of an arrow (A→B) indicates the causal relationship of two variables. The nodes that are placed immediately before a node are called parent, and the nodes that are placed after a node are called child.

1. A simple Bayesian network

In figure 8.15 A and B are parents of C, and D and E are children of C.

Prior information and expert information can be used to build a Bayesian network. In Bayesian network analysis, the first step is to determine the network structure, which we call structure learning, and the next step is to determine the parameters of the Bayesian network structure, which we call parameter learning. Methods of structural learning in a Bayesian network may be either constraint-based or score-based . In the following, a more detailed examination of these two categories of algorithms will be carried out. As already mentioned, the main theoretical core of Bayesian networks is based on Bayes' law, which was presented by Tomas Bayes in 1720, and that is why it is called a Bayesian network. tThe term Bayesian network was coined by Pearl in 1988 and It has been widely used in various fields ever since. In Bayesian networks, if there is a connection between each node and all other nodes in the network, it is called a full Bayesian network, and the main and important feature of Bayesian networks is that each child node matches its parents from the set of nodes and that Non-child and Child nodes are independent.

The goal of structural learning in the Bayesian network is to find the best structure in a way that matches the available data and is optimal in terms of complexity. As mentioned, this learning can be constraint-based or score-based.

Verma and Pearl in 1991 introduced constraint-based algorithms. This algorithm provides a theoretical framework for structural learning of causal models. , 3 steps are required:

Step 1)

In this step, the body of the network, which includes the undirected graph, is learned, generally for ease of implementation of the Markov Blanket, a set of nodes which contain complete information about one another. It is used for each node, including parents, children and co-nodes that have a common child with that particular Node. This structure of a Bayesian network is called the Skeleton of a DAG. Usually, at this stage, to determine the location of edges, Conditional independence tests such as X² and Fisher tests of independence are used.

Step 2)

In the next step, the edges obtained from the previous step are oriented. For this purpose, different algorithms are used such as the Fast Incremental Association (fast.iamb) ... Incremental Association (iamb), Grow-Shrink (gs) and Interleaved Incremental Association (inter.iam). And finally, the max-min parents and children algorithm (MMPC) is used to identify parents and children in the network.

In this book, the last algorithm, MMPC, is explained. This algorithm is a forward selection method for neighborhood deletion based on the maximization of the minimum association value obtained in each subset of nodes selected in previous iterations. This algorithm teaches the basic structure of the Bayesian network Learn. All rcs are non-directional and there is no attempt to direct them.

It should be noted, if thesamples is small or there is a lot of missing data, the use of limit oriented algorithms will cause a lot of error and these methods cannot direct some edges. In these cases, the use of score-based algorithms is suggested. In these methods, a measure is used to determine the matching of the networks with the available data searchingfor a network that matches the data the most. This method is also done in two steps.

In the first step, a search method is specified to make DAG, so that all possible structures of DAG are known, and then in the next step, the matching of each structure with the existing data is determined and evaluated with a suitable measure. These two steps continue until there is no possible structure that has a better match.

An important algorithm, from the score-based category,which is the greedy search algorithmcalled Hill-climbing (HC) that greedily searches the space of directed graphs (Daly and Shen, 2007).

The PCstructure learning algorithm is one of the earliest and the most popular algorithms, introduced by (Spirtes et al., 1993). It uses independences observed in data (established by means of classical independence tests) to infer the structure that has generated them.

Another algorithm is Tabu Search, which Glover introduced in 1990. This algorithm is a modified HC algorithm that does not stop after the first DAG, but continues until any addition, deletion, and reversal does not improve the score.

The structure of the Bayesian network may change during several executions of the learning algorithm. Therefore, in cross-sectional data, it is necessary to check the stability of the network structure, that is, to obtain a robust set of edges and directions..

For partial correlation networks, bootstrapping methods can be used to evaluate the stability of network estimates (See the network analysis chapter). Such a method can be used to check the stability of structural learning in Bayesian networks. For this reason, the Bayesian network learns the structure from a large set of samples obtained from bootstrapping and based on the criteria (Briganti, Scutra, and et al., 2022) that the edges that can be seen in more than 85% of the networks, that their direction appears in more than 50% of networks and that they remain in the network.

The first criterion is called strength and the second criterion is called minimum direction. It is suggested that the number of samples in the bootstrap method be 100 to 200 samples. And the proportion of the number of times that the edges are entered in the Bayesian network can also be reported (Briganti, et al ,2022). It should be remembered that after learning the structure, it is necessary to determine the intensity of the edges by parameter estimation. The purpose of parameter learning is to estimate network parameters. If the structure of the network is known and all the variables are observable, Maximum likelihood (ML), Maximum a posteriori (MAP) and posterior mean (PM) estimation can be used to estimate the network.

### 8.8.1. A Practical example using R

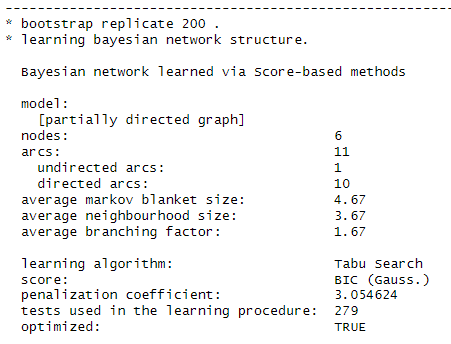
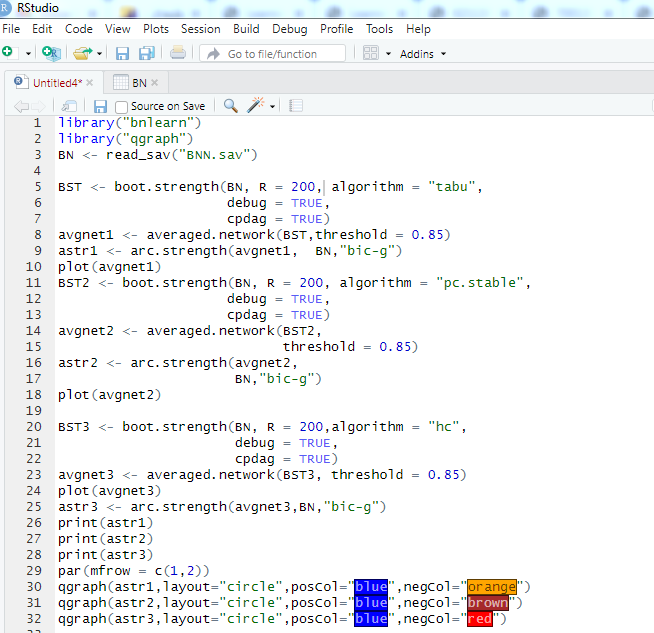
The researcher tries to model the exploratory Bayesian network based on variables related to marital satisfaction. He measures security (AAIS), avoidance (AAIV), ambivalence (AAIAM), positive affect (PANASP), negative affect (PANASN) and marital satisfaction (GRIMS) with valid and reliable questionnaires. His sample was a sample of 450 married men and women. In this research, he first builds a stable network by using bootstrapping with 200 samples, then analyzes the Bayesian network by using the structural learning PC algorithm, Tabu and HC, and finally estimates the parameters using the average strength of the edges. Listing6 and the Figures indicates the R codes for implementing the Bayesian network and the last step of bootstrapping with network specifications and finally the Bayesian network diagram and intensity of edges based on different methods of network learning are shown in the figures.

The R code for implementing the Bayesian network is in listing 6. Table1 shows the variable names and labels for running the Bayesian network.

Table1 The variables and variables names

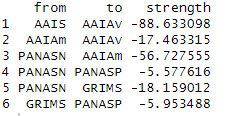
|  |  |
| --- | --- |
| **Variable name** | **Variable label** |
| AAIS | Secure attachment style |
| AAIAv | Avoidance attachment style |
| AAIAm | Ambivalence attachment style |
| PANASN | Negative Affect |
| PANASP | Positive Affect |
| GRIMS | Marital Satisfaction |

1. R codes for Bayesian Network implement



1. Learning the Bayesian network structure using the score-based method (Tabu Search)

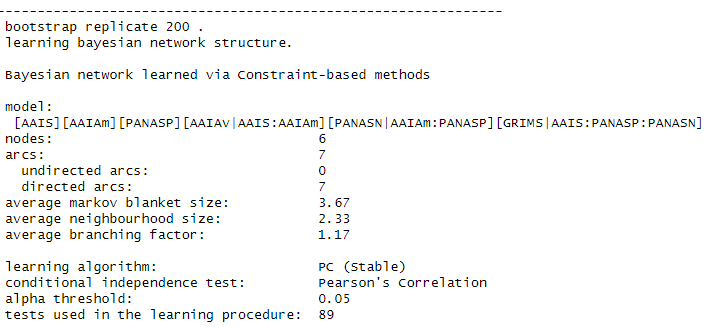
Bootstrap re-sampling can also be used to estimate a level of confidence on the learned edges (Friedman et al., 1999). In this case, one network can be learned from each bootstrap sample and the resulting PDAGs can be aggregated in a weighted PDAG (WPDAG), where the confidence on each edge is estimated as the fraction of bootstrap samples from which the edge can be learned. Bnstruct can also infer the estimated probability distribution of some variables, given evidence on the values of other observed variables. In this case, a junction tree (Koller and Friedman, 2009) is used. The Expectation-Maximization algorithm (Dempster et al., 1977) is also implemented, which exploits a BN structure to iteratively estimate conditional probabilities from a dataset with missing values in order to impute these missing values.



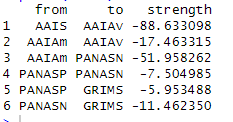
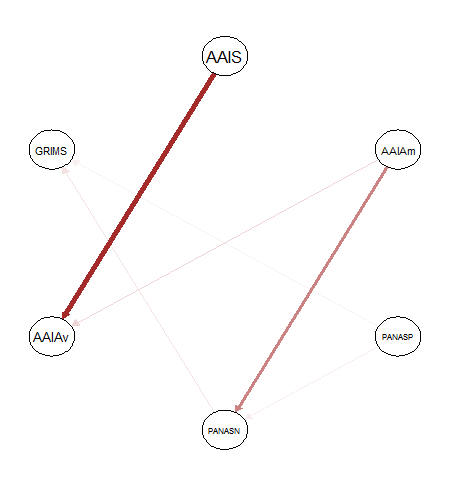
\*

1. The plot of learning the Bayesian network structure (Tabu Search).

Figure 8.16 shows the edge between AAIS and AAIAV has the strongest negative connection and GRIMS and PANAS has the weakest negative edge based on the obtained Bayesian network using Tabu search.

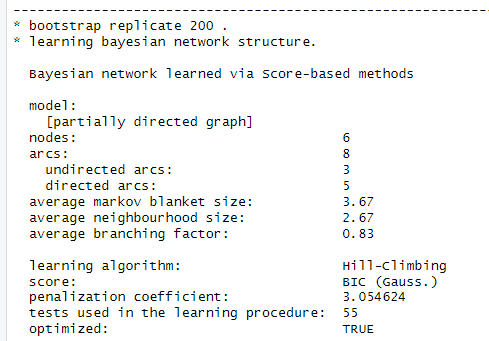


1. Learning the Bayesian network structure using the constraint –based algorithm (PC)

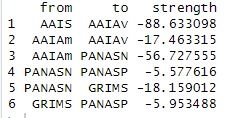
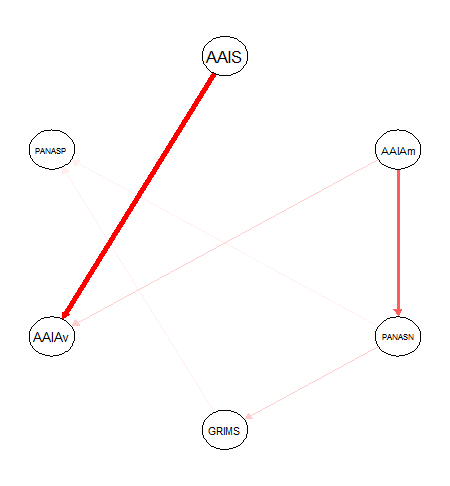


1. The nodes and edges of the Bayesian Network using PC

As the Figure 8.18 indicates AAIs has the strongest negative effect on AAIAv in the network obtained using PC.



1. Learning the Bayesian network structure using the scored –based (Hill-climbing)



1. The nodes and edges of the Bayesian Network using PC the scored –based (Hill-climbing)

As the Figure 8.20 indicates AAIs has the strongest negative effect on AAIAv in the network obtained using Hill-climbing.

## Bayesian Model Averaging

Bayesian model averaging (BMA) works out the posterior probabilities associated with various competing linear regression models comprising different subsets of features and aggregates over them with the models containing features having the highest posterior probability being most influential in giving the weighted average of parameter estimates representing associations of predictors with life satisfaction.

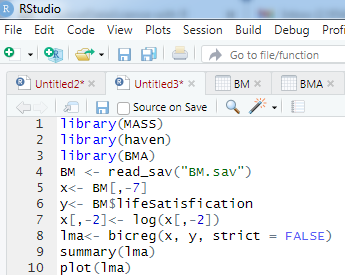
One of the most important points in data analysis and modeling is to find a way to choose a better model. Standard statistical analyzes ignore model uncertainty. In data analysis, a model is usually selected from among a class of models, which leaves this approach of uncertainty in model selection. For this reason, this approach leads to overconfident inferences. BMA provides a mechanism to consider model uncertainty when deriving parameter estimates. Therefore, in short, this method provides the uncertainty of the model that exists in the problem of variable selection by averaging the most important models according to the approximate estimation of the posterior probability of the model. For this purpose, in this method, competing models are examined through two indicatorsto select the best model (the most appropriate model). One of these indicators is the Bayesian Information Criterion (BIC) presented by Schwarz in 1978, and the other is the posterior probability of each model. The most appropriate model is the model that has the lowest BIC and the highest posterior probability. This method is suitable when there are a large number of predictor variables to predict an outcome, for example (20, 30 or 40). In this method, P!=0 is calculated for each variable in each model. See the output of the R software in the example. P! =0 indicates how likely it is that the regression coefficient for any particular predictor variable is non-zero among the resulting models. Also, in the "EV" column, the average of the posterior distribution for each coefficient is presented, and SD indicates the standard deviation of the posterior distribution for each coefficient. To be precise, only the best 5 models are presented. In the output of each BIC model, the posterior probability, the number of variables in each model, P!=0 and R² are presented.

BMA was used to calculate each variable’s relative importance. As can be seen from the above Beck Anxiety, Emotional Processing and Emotional Expression have posterior densities which are centred away from zero and feature in the regression model with the highest posterior probability of occurrence (0.60) and lowest Bayesian Inference Criterion (BIC) with lowest being best. The 4 remaining models share almost the same posterior probabilities and BIC which indicates that these competitive models are only slightly different from each other. The EV column shows the average of model coefficients. It is worth mentioning that if only one model was to be selected, this uncertainty in the model could not be justified. The average of all models’ coefficients, which is calculated by the sum of posterior probabilities’ weights and mentioned in the p!=0 columns, can consider this uncertainty. Beck Anxiety, Emotional Processing and Emotional Expression all have p!=0 equal to 100 showing a very high likelihood of association with life satisfaction.

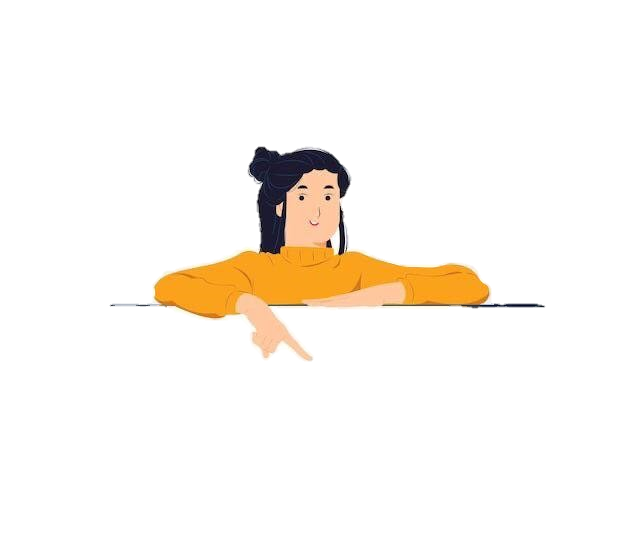
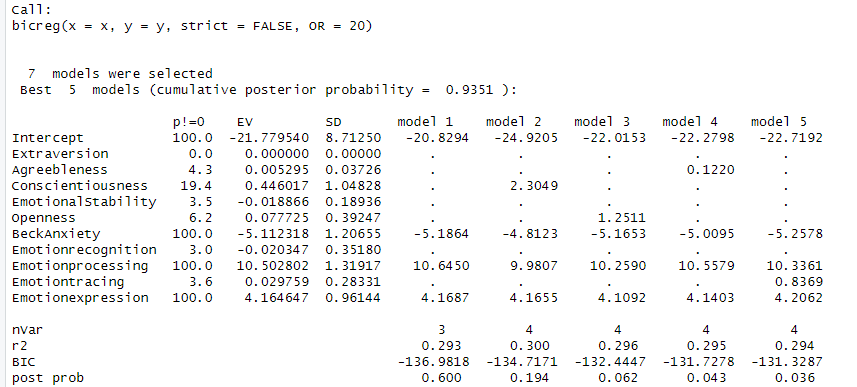
### 8.9.1. Practical example using R

An artificial psychologist is trying to determine which of the 10 variables (Extraversion, Conscientiousness, Agreeableness, Openness, Emotional Stability, Anxiety, Emotion recognition, Emotion processing, Emotional tracing, Emotion expression...) predicts Life Satisfaction. He uses a sample of 447 male and female students, and all of them respond to these variables through the respective scales, then BMA is fitted using the R code inlisting 7..

1. R codes for BMA implement



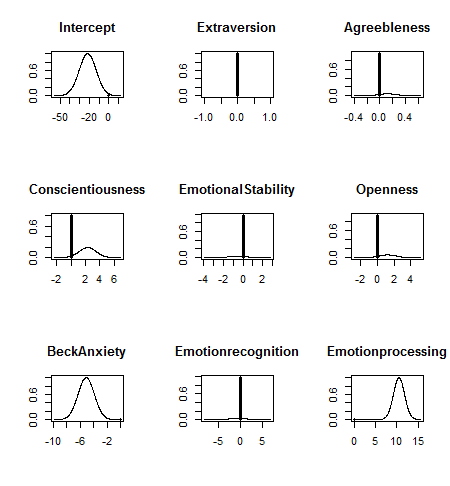
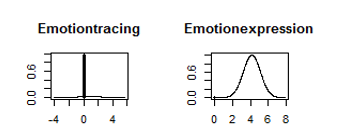
Considering the lower BIC and higher posterior probability, it can be said that the first model is the appropriate model in which 3 variables have been selected, which explains 29.3% of the variance in life satisfaction (Figure 8.21).



1. R output of BMA of the example

Figure 8.21 indicates that anxiety , emotional processing, and emotional expression are able to predict life satisfaction in model 1, which is the most appropriate model. Looking at p! =0 the percentage of models where the coefficients for all three variables are non-zero is 100 percent in this example. The EV column shows the mean of the posterior distribution of the coefficient of each variable, and the SD shows the standard deviation of the mean of each EV coefficient.

Examining the graphs shows that among the density graphs, each variable with the least amount of influence has a spike at the zero point, for example, it can be seen in the graphs that Extraversion, Agreeableness, Emotional Stability, Openness, Emotion recognition, and Emotional tracing have a spike at the zero point and have the least effect. These variables have coefficients that are most likely to be zero, and the variables of anxiety, emotional processing and emotion expression and of course the constant value of the regression equation play an important role in the resulting model.



1. The density of the variables

As we have seen, there are differences between the strength of different paths depending on what algorithm was used to learn the Bayesian network. But in all three Bayesian network algorithms, the path in the Bayesian network between secure attachment style and very strong avoidance is negative but the path between ambivalence and negative affect is in opposite directions using the Tabu search algorithm and the PC algorithm. . In the former, the ambivalent attachment style is affected by negative affect (PANASN→AAIAm), while in the latter, this path is the opposite (AAIAm→PANASN). What makes these models acceptable are the theoretical foundations and the possibility of a theoretical explanation of the resulting findings, and we should not forget that statistical findings are a support for researchers' theories.

Farahani et al. in 2022, used the BMA method to predict the Social-Emotional Competence Based on Childhood Trauma, Internalized Shame, Disability/Shame Scheme, Cognitive Flexibility, Distress Tolerance and Alexithymia in an Iranian Sample. You can find more information about this method in this paper.

.