



KNOWLEDGE MEDIA INSTITUTE

**Bayesian Methods for Intelligent
Data Analysis**

Marco Ramoni and Paola Sebastiani

KMI-TR-67

July 1998



Bayesian Methods for Intelligent Data Analysis

Marco Ramoni
Knowledge Media Institute
The Open University

Paola Sebastiani
Department of Statistics
The Open University

Abstract

This paper provides an introduction to Bayesian statistics as methodological tool for intelligent data analysis, knowledge discovery, and machine learning. The paper starts from the basic concepts of Bayesian statistics and reaches the presentation of most recent developments of the field, such as Bayesian Belief Networks. Knowledge of basic statistical concepts is required.

Keywords: Bayesian Statistics, Bayesian Belief Networks, Knowledge Discovery Machine Learning, Intelligent Data Analysis.

Reference: KMi Technical Report KMi-TR-67, July 1998.

Address: Marco Ramoni, Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom MK7 6AA. PHONE: +44 (1908) 655721, FAX: +44 (1908) 653169, EMAIL: m.ramoni@open.ac.uk, URL: <http://kmi.open.ac.uk/people/marco>.

Contents

1	The Bayesian Paradigm	2
2	Bayesian Inference	5
2.1	Prior Elicitation	5
2.2	Estimation	9
2.3	Credibility Intervals	11
2.4	Hypothesis Testing	12
3	Bayesian Modeling	13
3.1	Multiple Linear Regression	13
3.1.1	Parameter Estimation	15
3.1.2	Model Selection	19
3.2	Generalized Linear Models	21
3.3	Approximate Methods	22
3.3.1	Asymptotic Posterior Distributions	22
3.3.2	Stochastic Methods	24
4	Bayesian Networks	25
4.1	Foundations	25
4.2	Learning Bayesian Networks From Data	30
4.2.1	Parameter Estimation	30
4.2.2	Model Selection	36
4.3	Advanced Topics	39

1. The Bayesian Paradigm

Chapters 2 and 3 have shown that classical statistical methods are usually focused on the distribution $p(\mathbf{y}|\boldsymbol{\theta})$ of data \mathbf{y} , where $p(\cdot|\boldsymbol{\theta})$ denotes either the probability mass function or the density function of the sample of n cases $\mathbf{y} = (y_1, \dots, y_n)$, which is known up to a vector of parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. The information conveyed by the sample is used to refine this probabilistic model by estimating $\boldsymbol{\theta}$, by testing hypotheses on $\boldsymbol{\theta}$ and, in general, by performing statistical inference. However, classical statistical methods do not give the user the possibility of incorporating exogenous knowledge about the problem at hand, as arising for instance from past experience. Consider an experiment that consists of tossing a coin n times. If the results can be regarded as values of independent binary random variables Y_i taking values 1 and 0, with $\theta = p(Y_i = 1)$, where $Y_i = 1$ corresponds to the event “head in trial i ”, the likelihood function $L(\theta) = p(\mathbf{y}|\boldsymbol{\theta})$ (see Chapter 2) is

$$L(\theta) = \theta^{\sum_i y_i} (1 - \theta)^{n - \sum_i y_i}$$

and the MLE of θ is

$$\hat{\theta} = \frac{\sum_i y_i}{n}$$

which is the relative frequency of heads in the sample. This estimate of the probability of head is only a function of the sample information.

Bayesian methods, on the other hand, are characterized by the assumption that it is also meaningful to talk about the conditional distribution of $\boldsymbol{\theta}$, given the information I_0 currently available. Thus, a crucial aspect of Bayesian methods is to regard $\boldsymbol{\theta}$ as a random quantity whose *prior* density $p(\boldsymbol{\theta}|I_0)$ is known, before seeing the data. The prior distribution can arise from data previously observed, or it can be the *subjective* assessment of some domain expert. In the example of tossing the coin, we could for instance represent the prior belief that the coin is fair by adopting a prior density for θ , defined in $[0, 1]$, and with expectation 0.5.

The available information changes as new data \mathbf{y} are observed, and so does the conditional distribution of $\boldsymbol{\theta}$. This operation of revising the conditional distribution of $\boldsymbol{\theta}$ is done by making use of Bayes' Theorem:

$$p(\boldsymbol{\theta}|\mathbf{y}, I_0) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, I_0)p(\boldsymbol{\theta}|I_0)}{p(\mathbf{y}|I_0)}, \quad (1)$$

which updates $p(\boldsymbol{\theta}|I_0)$ into the *posterior* density $p(\boldsymbol{\theta}|\mathbf{y}, I_0)$ as the amount of information changes. Hence, regarding $\boldsymbol{\theta}$ as a random quantity gives Bayesian methods the ability to incorporate exogenous information into the inferential process so that the posterior distribution of $\boldsymbol{\theta}$ is conditional on the total information available:

$$\begin{array}{rcccc} \text{total information} & = & \text{prior information} & + & \text{data information} \\ I_1 & = & I_0 & + & \mathbf{y} \end{array}$$

Thus, inference is a continuous, dynamic process in which new data are used to revise the current knowledge. The probability density $p(\mathbf{y}|I_0)$ in (1) is computed by using the Total Probability Theorem:

$$p(\mathbf{y}|I_0) = \int p(\mathbf{y}|\boldsymbol{\theta}, I_0)p(\boldsymbol{\theta}|I_0)d\boldsymbol{\theta} \quad (2)$$

and it is also called the *marginal* density of data, to point out the fact that it is no longer conditional on $\boldsymbol{\theta}$.

The posterior distribution is the inference produced by Bayesian methods, and it can be used to compute a point estimate of the parameters, to test a hypothesis or, in general, to make probability statements as credibility intervals, or to predict future data \mathbf{y} , conditional on the available information I_1 , by computing

$$p(\mathbf{y}|I_1) = \int p(\mathbf{y}|\boldsymbol{\theta}, I_1)p(\boldsymbol{\theta}|I_1)d\boldsymbol{\theta}. \quad (3)$$

Note that (3) is essentially (2) with I_0 replaced by I_1 which is now the information currently available. If interest is in a subset of the parameters, e.g. $\boldsymbol{\theta}_1$, then the conditional density of $\boldsymbol{\theta}_1$ given I_1 can be obtained from the posterior density $p(\boldsymbol{\theta}|I_1)$ by integrating out the *nuisance* parameters $\boldsymbol{\theta}_2 = \boldsymbol{\theta} \setminus \boldsymbol{\theta}_1$:

$$p(\boldsymbol{\theta}_1|I_1) = \int p(\boldsymbol{\theta}|I_1)d\boldsymbol{\theta}_2.$$

In particular, inference on a single parameter, say θ_1 , is based on its marginal posterior density:

$$p(\theta_1|I_1) = \int p(\boldsymbol{\theta}|I_1)d\theta_2 \dots d\theta_k.$$

Similarly, inference on any function of the parameters can be performed.

Let now I denote the information currently available and \mathbf{y} be future data. In some circumstances, it is reasonable to assume that, conditional on $\boldsymbol{\theta}$, knowledge of I is irrelevant to \mathbf{y} , and hence

$$p(\mathbf{y}|\boldsymbol{\theta}, I) = p(\mathbf{y}|\boldsymbol{\theta}).$$

In this case \mathbf{y} and I are said to be *conditionally independent* given $\boldsymbol{\theta}$, and we will write $i(I, \mathbf{y}|\boldsymbol{\theta})$. The conditional independence assumption is reasonable when $\boldsymbol{\theta}$ specifies completely the current state of knowledge, so that I cannot add any relevant information if $\boldsymbol{\theta}$ is known.

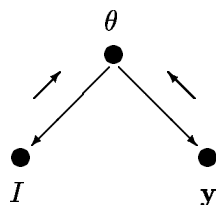


Figure 1: Graphical representation of conditional independence assumptions $i(I, \mathbf{y}|\boldsymbol{\theta})$.

The stochastic dependency among I , $\boldsymbol{\theta}$ and \mathbf{y} , together with the conditional independence of I and \mathbf{y} given $\boldsymbol{\theta}$ can be graphically represented using the Directed Acyclic Graph (DAG) in Figure 1. From a qualitative viewpoint, the two directed links pointing from node $\boldsymbol{\theta}$ to I and \mathbf{y} represent the stochastic dependence of I and \mathbf{y} on $\boldsymbol{\theta}$, so that $\boldsymbol{\theta}$ is called a *parent* of I and \mathbf{y} , and they are both *children* of $\boldsymbol{\theta}$. There is not a directed link between \mathbf{y} and I , which can “communicate” only via $\boldsymbol{\theta}$: in other words, $\boldsymbol{\theta}$ separates I from \mathbf{y} . Separation is interpreted in terms of conditional independence assumptions, and hence $i(I, \mathbf{y}|\boldsymbol{\theta})$ [22]. The stochastic dependence of I and \mathbf{y} on $\boldsymbol{\theta}$ is quantified by the conditional densities $p(I|\boldsymbol{\theta})$ and $p(\mathbf{y}|\boldsymbol{\theta})$ that are used to sequentially revise the distribution of $\boldsymbol{\theta}$. First, the conditional density $p(\boldsymbol{\theta}|I)$ is computed by processing the information I . This updating operation is represented by the arrow from I towards $\boldsymbol{\theta}$ that represents the “flow” of information from I to $\boldsymbol{\theta}$ via application of Bayes’ Theorem. After this first updating, the probability density associated with $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta}|I)$. When new data arrive, whose probability density $p(\mathbf{y}|\boldsymbol{\theta}, I) = p(\mathbf{y}|\boldsymbol{\theta})$ is known, the new piece of information is processed locally, along the path $\boldsymbol{\theta} \rightarrow \mathbf{y}$, by applying Bayes’ Theorem again, so that the conditional density of $\boldsymbol{\theta}$, after this second updating, is $p(\boldsymbol{\theta}|I, \mathbf{y})$.

Bayes’ Theorem and the Total Probability Theorem are the fundamental tools of Bayesian methods to produce the posterior distribution of the parameters. When coupled with the conditional independence of the data given the parameters, the inference procedure can process data as a whole (*in batch*), as classical methods do, but it can also process data one at the time (*sequentially*), and the final result will be the same. This incremental nature is a further advantage of Bayesian methods: the acquisition of new data does not require reprocessing data considered so far.

In the next section we will describe the foundations of Bayesian methods and their applications to estimation, model selection, and reliability assessment. We will use simple examples with the aim of explaining the techniques described. More complex models are considered in Section 3 in which Bayesian methods are applied to the statistical analysis of multiple linear regression models and Generalized Linear Models. Computational methods

to effectively implement these methods will be described in Section 3.3. Section 4 will describe a powerful formalism known as *Bayesian Belief Networks* (BBNS) and its applications to prediction, classification and modeling tasks.

2. Bayesian Inference

Suppose we have a sample of n cases $\mathbf{y} = \{y_1, \dots, y_n\}$, generated from a density function $p(\mathbf{y}|\boldsymbol{\theta})$. The density $p(\cdot|\boldsymbol{\theta})$ is known, up to a vector of unknown parameters. We assume that the cases are independent given $\boldsymbol{\theta}$, and hence the joint probability density of the sample is

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^n p(y_i|\boldsymbol{\theta}).$$

The likelihood function $L(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$ plays a central role in classical methods. For Bayesian methods, the likelihood function is the instrument to pass from the prior density $p(\boldsymbol{\theta}|I_0)$ to the posterior density $p(\boldsymbol{\theta}|I_0, \mathbf{y})$ via Bayes' Theorem. Compared to classical methods, Bayesian methods involve the use of a further "ingredient": the prior density of $\boldsymbol{\theta}$. The first step of a Bayesian data analysis is therefore the assessment of this prior density.

2.1 Prior Elicitation

The prior density of $\boldsymbol{\theta}$ can arise either as posterior density derived from processing past data or it can be the *subjective* assessment elicited by some domain expert. Several methods for eliciting prior densities from experts exist, and [21, Ch 6] reports a comprehensive review.

A common approach chooses a prior distribution with density function similar to the likelihood function. In doing so, the posterior distribution of $\boldsymbol{\theta}$ will be in the same class and so the prior is said to be *conjugate* to the likelihood. Conjugate priors play an important role in Bayesian methods, since their adoption can simplify the integration procedure required by the marginalization. If the prior is conjugate to the likelihood, then the posterior density is in the same class, and the computational problem reduces to finding the updated parameters. A list of important conjugate distributions is given in [2, Ch. 5].

Example 1 Let $y_1, \dots, y_n|\theta$ be independent variables taking 0 and 1 values, and let $\theta = p(Y_i = 1|\theta)$, $\theta \in [0, 1]$. The likelihood function is therefore

$$L(\theta) \propto \theta^{\sum_i y_i} (1 - \theta)^{n - \sum_i y_i}.$$

The parameter θ is univariate, and constrained to be in the interval $[0, 1]$. This choice constrains the class of possible priors. A prior density of the form

$$p(\theta|I_0) \propto \theta^{a-1} (1 - \theta)^{b-1}, \quad \theta \in [0, 1], \quad a, b > 0$$

will be conjugate, since it has the same functional form as the likelihood $\theta^x(1-\theta)^z$, except for the exponents. This distribution is called a Beta distribution, with hyper-parameters a and b , and it is sometimes denoted by $\text{Beta}(a, b)$. The word hyper-parameters is used to distinguish a and b from the parameter θ of the sampling model. Note that, compared to the likelihood, the hyper-parameters $a-1$ and $b-1$ of $p(\theta|I_0)$ play the roles of $\sum_i y_i$ and $n - \sum_i y_i$ respectively. Thus, $a-1$ and $b-1$ can be chosen by assuming that the expert has an imaginary sample of 0s and 1s, of size $a+b-2$, and he can distribute the imaginary cases between 0 and 1 as his prior knowledge dictates. The size of this imaginary sample can be used to characterize the subjective confidence of the expert in her/his own assessment. Summaries of this distribution are

$$\begin{aligned} \text{E}(\theta|I_0) &= \frac{a}{a+b} \\ \text{mode} &= \frac{a-1}{a+b-2} \\ \text{variance} &= \frac{ab}{(a+b)^2(a+b+1)} = \frac{\text{E}(\theta|I_0)(1-\text{E}(\theta|I_0))}{a+b+1} \end{aligned}$$

The prior expectation $\text{E}(\theta|I_0)$ corresponds to the marginal probability of Y before seeing any data:

$$\text{E}(\theta|I_0) = \int \theta p(\theta|I_0) d\theta = \int p(Y=1|\theta) p(\theta|I_0) d\theta = p(Y=1|I_0).$$

Since the variance of θ is a decreasing function of $a+b$ for a given mean, the sum of the hyper-parameters $a+b$ is also called the precision of the distribution. The posterior density is easily found to be:

$$p(\theta|I_0) \propto \theta^{a+\sum_i y_i-1} (1-\theta)^{b+n-\sum_i y_i-1}, \quad \theta \in [0, 1]$$

which identifies a Beta distribution with hyper-parameters $a + \sum_i y_i$ and $b + n - \sum_i y_i$. Thus, the posterior precision is increased by the sample size n .

Conjugacy restricts the choice of priors to a limited class of distributions and prior information can only be used to choose the hyper-parameters. However, if the class of distributions is large enough, this limitation is not a serious issue. For instance, in Example 1 a choice $a = b = 1$ yields a prior distribution which is uniform in $[0, 1]$, and hence uniform prior uncertainty about θ , so that all values are equally likely. Choosing $a = b$ implies that the distribution of θ is symmetrical about the prior mean and mode that are both equal to 0.5. A choice $a < b$ induces negative skew, so that large values of θ are, a priori, more likely. Positive skew corresponds to $a > b$. Several examples are given in plots (a), (b) and (c) in Figure 2. The corresponding posterior densities derived from a sample of size $n = 10$, and $\sum_i y_i = 6$ are given in plots (d), (e) and (f).

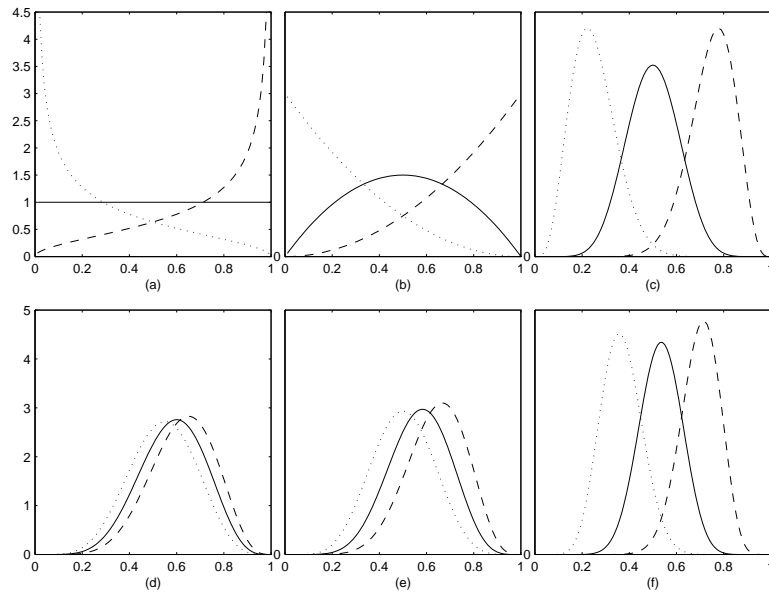


Figure 2: Densities of Beta distributions for different choices of the hyper-parameters a and b . Continuous lines report symmetric densities: (a) $a=b=1$; (b) $a=b=2$; (c) $a=b=10$. Dotted lines report positively skewed densities: (a) $a=1.5$; $b=0.5$; (b) $a=3$; $b=1$; (c) $a=15$; $b=5$. Dashed lines report negatively skewed densities: (a) $a=0.5$; $b=1.5$; (b) $a=1$; $b=3$; (c) $a=5$; $b=15$. Plots (d), (e) and (f) report the corresponding posterior densities derived from a sample of size $n = 10$, and $\sum_i y_i = 6$.

The effect of the prior input is small compared to the data when the sample size n is larger than the prior precision (plots (a), (b) versus (d) and (e)): posterior densities are very different from prior densities. For instance, when the prior hyper-parameters are $a = 0.5$ and $b = 1.5$ (dashed line in plot (a)), the prior distribution assigns larger probability to values of θ smaller than 0.5. The impact of the sample information is to concentrate the posterior distribution in the range $[0.2 \ 0.9]$, with a median around 0.5. The difference is less evident in plots (c) and (f), when the prior precision is larger than the sample size: in this case the posterior densities are slightly shifted to take into account the effect of the sample.

An alternative way to assess prior distributions is based on the concept of *Maximum Entropy*: it requires the expert to specify some summaries of the prior distribution, such as the mean or the variance, and it returns the prior distribution having maximum entropy among the class of distributions identified by the given summary. This assessment method, due to Jaynes [17, 18], was devised in order to provide probability assessments that are subject only to the available information and the prior returned contains as little information as possible, apart from the summaries specified. In this way, bias due to unintentional subjective components included in the elicitation process is removed, and two experts with the same prior information will return the same prior distribution.

Example 2 (Maximum Entropy Priors) *When θ is univariate and takes all real values, and the prior mean and variance are specified, the maximum entropy prior is a Normal distribution with the specified mean and variance.*

An open problem of Bayesian methods is the choice of a prior distribution representing genuine ignorance. When this prior exists, it is called *non-informative*. When a distribution for θ is non-informative, and we make a parameter transformation $\psi = g(\theta)$, then the distribution of ψ must be non-informative. The Jeffreys' rule [19] allows us to find prior distributions that are invariant under reparameterizations. If the likelihood function $L(\theta)$ is known and we define $l(\theta) = \log L(\theta)$, the expected Fisher information matrix is

$$I(\theta) = -E \left\{ \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right\}$$

where the expectation is over the conditional distribution of the data given θ , and the Jeffreys prior has density

$$p(\theta|I_0) \propto \det\{I(\theta)\}^{1/2},$$

with $\det(\cdot)$ denoting the determinant of a matrix. If $\psi = g(\theta)$, then $p(\psi|I_0) \propto \det\{I(\psi)\}^{1/2}$, and the prior distribution is invariant with respect to reparameterization. In most cases, Jeffreys priors are not technically probability distributions, since their density functions do not have finite integrals over the parameter space, and are therefore termed *improper*

prior. It is often the case that Bayesian inference based on improper priors returns proper posterior distributions which

then turn out to be numerically equivalent to the results of classical inference [3].

Example 3 Let $y_1, \dots, y_n | \theta$ be independent, normally distributed variates with mean θ and known variance σ^2 . Then,

$$p(\mathbf{y}|\theta) \propto \exp\{-n(\bar{y} - \theta)^2/2\sigma^2\}$$

and $I(\theta) = n/\sigma^2$, so that the Jeffreys prior for θ is the (improper) uniform distribution over the real numbers. Nonetheless, the posterior distribution is

$$\theta|\mathbf{y} \sim N(\bar{y}, \sigma^2/n)$$

which is a proper distribution.

Problems related to the use of improper prior distributions can be overcome by assigning prior distributions that are as uniform as possible but still remain probability distributions. For instance, in Example 3, the prior distribution can be Normal with a very large variance. This would also be the Maximum Entropy prior, when the prior knowledge is extremely uncertain.

The use of uniform prior distribution to represent uncertainty clearly assumes that equally probable is an adequate representation of lack of information. Recent advances question this assumption and advocate the use of bounds on the set of possible values that θ can take [14], [28]. Finally, it is worth mentioning that prior distributions elicited from several experts can be combined in a single mixture of different distributions with weights representing the reliability of each expert.

2.2 Estimation

Bayesian inference returns the posterior density $p(\boldsymbol{\theta}|I_0, \mathbf{y}) = p(\boldsymbol{\theta}|I_1)$. Marginal inference on parameters of interest is then based on the marginal posterior distribution, and for an individual parameter θ_1 on

$$p(\theta_1|I_1) = \int p(\boldsymbol{\theta}|I_1) d\theta_2 d\theta_3 \dots d\theta_k.$$

A standard point estimate of θ_1 is the posterior expectation:

$$E(\theta_1|I_1) = \int \theta_1 p(\theta_1|I_1) d\theta_1. \tag{4}$$

An alternative point estimate, based on a principle similar to Maximum Likelihood (see Chapters 2 and 3), is the posterior mode:

$$\hat{\theta}_1 = \arg \max_{\theta} p(\theta | I_1). \quad (5)$$

Since $p(\theta | I_1) \propto p(\mathbf{y} | \theta)p(\theta | I_0)$, the vector of posterior modes $\hat{\theta}$ maximizes the *augmented* likelihood $p(\mathbf{y} | \theta)p(\theta | I_0)$, and it is therefore called the Generalized Maximum Likelihood Estimate (GMLE) of θ . When the parameters $\theta_1, \dots, \theta_k$ are, a posteriori, independent, then

$$p(\theta | I_1) = \prod_i p(\theta_i | I_1)$$

and the posterior mode of individual parameters can be computed from the marginal posterior densities.

Example 1 (continued). Since the posterior distribution is still Beta, the posterior mean is

$$\begin{aligned} \mathbb{E}(\theta | I_1) &= \frac{a + \sum_i y_i}{a + b + n} = \frac{a + b}{a + b + n} \frac{a}{a + b} + \frac{n}{a + b + n} \frac{\sum_i y_i}{n} \\ &= \frac{a + b}{a + b + n} \mathbb{E}(\theta | I_0) + \frac{n}{a + b + n} \bar{y} \end{aligned}$$

which is a weighted average of the prior $\mathbb{E}(\theta | I_0)$ and sample mean \bar{y} , with weights depending on the sample size n and prior precision $a + b$. If $n < a + b$, the prior mean has a larger weight than the posterior mean. As the sample size increases, prior information becomes negligible, and the posterior mean approximates the MLE \bar{y} of θ . The posterior mode

$$\hat{\theta}_1 = \frac{a + \sum_i y_i - 1}{a + b + n - 2}$$

reduces to the standard MLE of θ when $a = b = 1$, and hence the prior distribution assumes that all values of θ are equally likely.

Example 4 Let $y_1, \dots, y_n | \theta$ be independent, normally distributed with mean θ and known variance $\sigma^2 \equiv \tau^{-2}$, where τ^2 is called precision. Then,

$$p(\mathbf{y} | \theta) \propto \exp\{-n\tau^2(\bar{y} - \theta)^2/2\}$$

and $\theta | I_0 \sim N(\mu_0, \sigma_0^2)$ is a conjugate prior with hyper-parameters μ_0 and $\sigma_0^2 \equiv \tau_0^{-2}$, (this is also the Maximum Entropy prior for specified mean and variance.) By conjugacy, the posterior distribution is $N(\mu_1, \sigma_1^2)$ and it can be easily shown, for instance [21, page 7], that the posterior hyper-parameters are:

$$\begin{aligned} \sigma_1^{-2} = \tau_1^2 &= \tau_0^2 + n\tau^2 \\ \mu_1 &= \frac{\tau_0^2 \mu_0 + n\tau^2 \bar{y}}{\tau_1^2} . \end{aligned}$$

Thus, the posterior mean is again a weighted average of the prior and sample means, with weights that depend on the sample size n , the datum precision τ^2 and the prior precision τ_0^2 . As in Example 1, there is a trade-off between data and prior information, and for small samples the prior mean has a weight larger than the sample mean. As the sample size increases, the prior input becomes negligible, and asymptotically the Bayesian estimate is the sample mean \bar{y} . By symmetry, the posterior mode and mean are coincident.

2.3 Credibility Intervals

The posterior mean and mode provide simple summaries of the posterior distribution that can be further used to evaluate the probability that θ is in some given region R , or to find a region R that contains θ with a specified probability $1 - \alpha$. The latter is called a $(1 - \alpha)\%$ *credibility region*, i.e.

$$p(\theta \in R|I_1) = 1 - \alpha.$$

When R is the region of smallest volume, it is also called the *Posterior Highest Density* (PHD) region. When

$\theta = \theta$, the PHD region is an interval, and if the posterior density of θ is unimodal, the PHD region is given by two values l_1 and l_2 such that

$$\begin{aligned} \int_{l_1}^{l_2} p(\theta|I_1)d\theta &= 1 - \alpha \\ p(l_1|I_1) &= p(l_2|I_1) \end{aligned}$$

Computational methods for finding the PHD region in particular problems can be found in [1, page 140].

Example 3 (continued). Given that the posterior distribution of θ is $N(\bar{y}, \sigma^2/n)$, a $(1 - \alpha)\%$ PHD interval is easily found to be:

$$\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \quad p(Z > z_{\alpha/2}) = \alpha/2; \quad Z \sim N(0, 1)$$

which is identical to the classical $(1 - \alpha)\%$ confidence interval for the mean of a normal population when the variance is known (see Chapter 2). However, the meaning of the latter is different. The frequentist interpretation of the $(1 - \alpha)\%$ confidence interval is based on the repeatability of the sampling process, so that if we could take, say, 100 samples, we would expect that in $(1 - \alpha)\%$ of cases the interval $\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ contains the true value of θ . The $(1 - \alpha)\%$ PHD interval returned by the Bayesian method is a credibility statement, conditional on the information I_1 currently available: we believe that, with probability $(1 - \alpha)$, θ belongs to the interval $\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

2.4 Hypothesis Testing

The Bayesian approach to hypothesis testing is based on the computation of the conditional probability of a hypothesis H given the information currently available I_0 , that is, $p(H|I_0)$. Thus, when the null hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ and the alternative hypothesis $H_1 : \boldsymbol{\theta} \in \Theta_1$, with $\Theta_0 \cap \Theta_1 = \emptyset$, are formulated, there are prior beliefs on both of them, say $p(H_0|I_0)$ and $p(H_1|I_0)$, with $p(H_0|I_0) + p(H_1|I_0) = 1$. By the Total Probability Theorem (applied to the discrete case), the prior density of $\boldsymbol{\theta}$ is then:

$$p(\boldsymbol{\theta}|I_0) = p(\boldsymbol{\theta}|H_0, I_0)p(H_0|I_0) + p(\boldsymbol{\theta}|H_1, I_0)p(H_1|I_0)$$

where $p(\boldsymbol{\theta}|H_i, I_0)$ are the prior densities of $\boldsymbol{\theta}$, conditional on each hypothesis. The sample information is then used to compute from the *prior odds*:

$$\frac{p(H_0|I_0)}{p(H_1|I_0)}$$

the *posterior odds* in favor of H_0 :

$$\frac{p(H_0|I_1)}{p(H_1|I_1)} = \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)} \frac{p(H_0|I_0)}{p(H_1|I_0)},$$

from which the following decision rule is derived:

$$\begin{array}{ll} \text{if } p(H_0|I_1) < p(H_1|I_1) & \text{Reject } H_0 \\ \text{if } p(H_0|I_1) > p(H_1|I_1) & \text{Accept } H_0 \\ \text{if } p(H_0|I_1) = p(H_1|I_1) & \text{Undecidability.} \end{array}$$

Compared to classical methods in which the sampling variability is taken into account in the definition of the rejection region of the test (see Chapter 2), the Bayesian method to hypothesis testing is to accept as true the hypothesis whose posterior probability is the largest, since it is the most likely given the information available. The ratio $p(\mathbf{y}|H_0)/p(\mathbf{y}|H_1)$ is called the *Bayes factor*, and when the prior probabilities of H_0 and H_1 are equal, the Bayes factor determines the decision rule. The evaluation of the Bayes factor involves the computation of:

$$\begin{aligned} p(\mathbf{y}|H_0) &= \int p(\mathbf{y}|H_0, \boldsymbol{\theta})p(\boldsymbol{\theta}|H_0, I_0)d\boldsymbol{\theta} \\ p(\mathbf{y}|H_1) &= \int p(\mathbf{y}|H_1, \boldsymbol{\theta})p(\boldsymbol{\theta}|H_1, I_0)d\boldsymbol{\theta} \end{aligned}$$

and hence the evaluation of the marginal densities of the data on the two parameter spaces specified in H_0 and H_1 . When the two hypotheses are simple: $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, then the Bayes factor reduces to the classical likelihood ratio test to discriminate between two simple hypotheses. Further details can be found in [1, Chapter 4].

Example 5 Let $y_1, \dots, y_n | \theta$ be independent, identically distributed Poisson $P(\theta)$ variates. Thus

$$p(y_i | \theta) = \frac{\theta^{y_i}}{y_i!} e^{-\theta}; \quad \theta > 0 \quad y_i = 0, 1, 2, \dots$$

Let $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ be two simple hypotheses, with $p(H_0 | I_0) = p(H_1 | I_0)$. The Bayes factor is

$$\left(\frac{\theta_0}{\theta_1} \right)^{\sum_i y_i} e^{n(\theta_1 - \theta_0)}$$

and hence, since the prior odds are equal to 1, the decision rule is to accept H_0 if the Bayes factor is greater than 1.

3. Bayesian Modeling

The examples discussed in Sections 2.1–2.4 are “toy” examples used to explain the Bayesian approach. In this section, we will focus on more realistic models in which a response variable Y is a function of some covariates $X_1, \dots, X_c \in \mathcal{X}$. We begin by considering the multiple linear regression model, in which data have a Normal distribution whose expectation is a linear function of the parameters. We then consider Bayesian methods for the analysis of Generalized Linear Models, which provide a general framework for cases in which normality and linearity are not viable assumptions. These cases point out the major computational bottleneck of Bayesian methods: when normality and linearity assumptions are removed, there is no closed form solution to compute the posterior distribution. We will then discuss some computational methods to approximate this distribution.

3.1 Multiple Linear Regression

We begin by considering the standard multiple linear regression model, in which data are assumed to have the distribution

$$Y_i | (\mu_i, \tau^2) \sim N(\mu_i, \tau^{-2}) \tag{6}$$

conditional on μ_i and τ^2 . The expectation μ_i is a linear function of the regression parameters $\boldsymbol{\beta}$, with coefficients that are functions of the regression variables X_1, \dots, X_c :

$$\mu_i = f(\mathbf{x}_i)^T \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^p \beta_j f_j(\mathbf{x}_i) \equiv \beta_0 + \sum_{j=1}^p \beta_j t_{ij}.$$

The function $f(\cdot)$ is defined in \mathcal{X} and takes values in $\mathcal{T} \subset \mathcal{R}^{p+1}$. This definition allows us to consider general regression models as polynomial regression, e.g. with $c = 1$ and

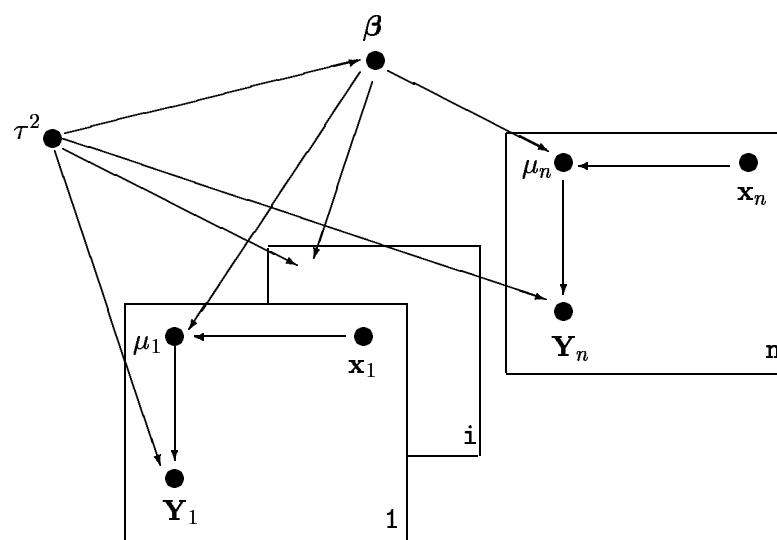


Figure 3: Graphical representation of a regression model with independent observations given $\theta = (\beta, \tau^2)$.

$f(x_i) = (1, x_i, x_i^2)$ we have a quadratic regression model $\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. The linear model can be written in matrix form as

$$\mathbf{Y}|\boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \tau^{-2}I_n)$$

where X is the $n \times (p + 1)$ design matrix, whose i th row contains the coefficients $f(\mathbf{x}_i)$ of the regression parameters, and I_n is the $n \times n$ identity matrix.

3.1.1 Parameter Estimation

The parameter vector $\boldsymbol{\theta}$ is given by $(\boldsymbol{\beta}, \tau^2)$. We further suppose $i(Y_1, \dots, Y_n | \boldsymbol{\theta})$, so that the likelihood function is:

$$L(\boldsymbol{\theta}) \propto \tau^n \exp(-\tau^2 \sum_i (y_i - f(\mathbf{x}_i)^T \boldsymbol{\beta})^2 / 2).$$

When both $\boldsymbol{\beta}$ and τ^2 are unknown, the simplest analysis is obtained by assuming a conjugate prior for $\boldsymbol{\theta}$, which is specified in two steps:

- (i) Conditional on τ^2

$$\boldsymbol{\beta} | \tau^2, I_0 \sim N(\boldsymbol{\beta}_0, (\tau^2 R_0)^{-1})$$

where $\tau^2 R_0$ is the (conditional) prior precision.

- (ii) The datum precision τ^2 is assigned a prior distribution:

$$\tau^2 | I_0 \sim \chi_{\nu_0}^2 / (\nu_0 \sigma_0^2)$$

which corresponds to assigning the error variance σ^2 an *Inverse Gamma* distribution [2, page 119] with density function:

$$p(\sigma^2 | I_0) \propto (\sigma^2)^{-(\nu_0+2)/2} \exp(-\nu_0 \sigma_0^2 / (2\sigma^2)).$$

The specification of the hyper-parameters ν_0 , σ_0^2 , $\boldsymbol{\beta}_0$ and R_0 allows the encoding of the prior information I_0 . For instance, the expectation and variance of τ^2 are respectively σ_0^{-2} and $2\sigma_0^{-4}/\nu_0$, so that the expert's information on the variability of the data can be used to define σ_0^{-2} , while the choice of ν_0 may represent the expert's assessment of his ability in terms of size of an imaginary sample used to elicit such prior distribution [21, Ch. 9]. The prior hyper-parameters and the distribution chosen imply that

$$E(\sigma^2 | I_0) = \frac{\nu_0 \sigma_0^2}{\nu_0 - 2}, \quad V(\sigma^2 | I_0) = \frac{2\nu_0^2 \sigma_0^4}{(\nu_0 - 2)^2 (\nu_0 - 4)}$$

provided that $\nu_0 > 4$. For $2 < \nu_0 \leq 4$ the variance does not exist, and for $\nu_0 \leq 2$ the mean does not exist.

Similarly, β_0 represents some expert's belief about the regression model, while R_0 is a measure of the precision of his assessment, for a fixed value of τ^2 . In this way, the prior elicitation of the distribution of β can be done independently of the sampling variability. The marginal variance of β can be easily found to be:

$$V(\beta|I_0) = E\{V(\beta|I_0, \tau^2)\} + V\{E(\beta|I_0, \tau^2)\} = \frac{\nu_0\sigma_0^2}{\nu_0 - 2}R_0^{-1}.$$

The joint prior distribution is known as a *Normal-Inverse-Gamma* prior. The elicitation of the prior distribution in two steps and the conditional independence assumptions are represented by the DAG in Figure 3. The link from τ^2 to β represents the two-step prior elicitation process described above, so that the distribution of τ^2 depends only on the information I_0 currently available, and then conditional on τ^2 the distribution of β is elicited. For consistency, the node τ^2 should be child of a root node I_0 representing the prior information collected from past experience. For simplicity, this node has been removed from the graph. The links from τ^2 and β to Y_1, \dots, Y_n specify the sampling model, and hence the stochastic dependence of Y on β , via μ , and on τ^2 . The conditional independence $i(Y_1, \dots, Y_n|\theta)$ is represented by the lack of directed links among Y_1, \dots, Y_n that can communicate only via τ^2 and β . The conditional independence of the observations given θ is encoded by representing the associations among Y_i, μ_i and \mathbf{x}_i on different plateaux, one for each observation.

The quantification of the dependencies is done by associating the prior distribution $\chi_{\nu_0}^2/(\nu_0\sigma_0^2)$ to the root node τ^2 , and the distribution $N(\beta_0, (\tau^2 R_0)^{-1})$ to the node β . The joint information of β and values of the covariates are summarized into the nodes μ_1, \dots, μ_n that represent linear functions of β with coefficients $f(\mathbf{x}_i)$ and hence they inherit their variability from the "stochastic" parents β . The sampling models $N(\mu_i, \tau^{-2})$ are attached to the nodes Y_1, \dots, Y_n .

Data are then processed by applying Bayes' Theorem in an order which is opposite to the order of elicitation of the prior density, so that first there is a flow of information from Y_1, \dots, Y_n to β and the conditional posterior distribution is found to be multivariate normal with updated hyper-parameters:

$$\beta|I_1, \tau^2 \sim N(\beta_1, (\tau^2 R_1)^{-1}) \tag{7}$$

$$R_1 = (R_0 + X^T X) \tag{8}$$

$$\beta_1 = R_1^{-1}(R_0\beta_0 + X^T \mathbf{y}). \tag{9}$$

Thus, for fixed τ^2 , the posterior precision of β is increased by the datum precision $X^T X$, that is the Fisher information matrix, and the posterior expectation is an average of prior

expectation β_0 and data \mathbf{y} . Thus, a point estimate of β (conditional on τ^2) is

$$E(\beta|I_1) = R_1^{-1}R_0\beta_0 + R_1^{-1}X^T\mathbf{y} = R_1^{-1}R_0\beta_0 + (R_0 + X^T X)^{-1}X^T\mathbf{y}$$

and compared to the MLE $\hat{\beta} = (X^T X)^{-1}X^T\mathbf{y}$ there is an adjustment to take into account prior information. As the weight of the prior information becomes negligible compared to the sample information, the Bayesian estimate approximates $\hat{\beta}$. By conjugacy, the marginal posterior distribution of τ^2 is found to be

$$\tau^2|I_1 \sim \chi_{\nu_1}^2/(\nu_1\sigma_1^2) \quad (10)$$

$$\nu_1 = \nu_0 + n \quad (11)$$

$$\nu_1\sigma_1^2 = \nu_0\sigma_0^2 + \mathbf{y}^T\mathbf{y} + \beta_0^T R_0\beta_0 - \beta_1^T R_1\beta_1. \quad (12)$$

in which the degrees of freedom are increased by the sample size n . Denote by $\hat{\mathbf{y}} = X\hat{\beta}$ the fitted values of the classical regression model and let rss denote the residual sum of squares $(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$, so that $\hat{\sigma}^2 = rss/(n - p - 1)$ is the classical unbiased estimate of the error variance (see Chapters 2 and 3.) Then, it can be shown [21, page 249] that

$$\nu_1\sigma_1^2 = \nu_0\sigma_0^2 + rss + (\beta_0 - \hat{\beta})^T(R_0^{-1} + (X^T X)^{-1})^{-1}(\beta_0 - \hat{\beta})$$

so that the expected posterior variance is

$$E(\sigma^2|I_1) = \frac{\nu_0 - 2}{\nu_0 + n - 2}E(\sigma^2|I_0) + \frac{n - p - 1}{\nu_0 + n - 2}\hat{\sigma}^2 + \frac{p + 1}{\nu_0 + n - 2}d$$

where $d = (\beta_0 - \hat{\beta})^T(R_0^{-1} + (X^T X)^{-1})^{-1}(\beta_0 - \hat{\beta})/(p + 1)$. Therefore $E(\sigma^2|I_1)$ provides an estimate of the error variance, which combines prior information $E(\sigma^2|I_0)$, the standard unbiased estimate of the error variance $\hat{\sigma}^2$ and d , that is a weighted discrepancy between β_0 and $\hat{\beta}$. Thus, a large discrepancy between β_0 and $\hat{\beta}$ represents the fact that the prior information is scarce and hence the variance is large.

Inference on the regression parameters is performed by computing the marginal posterior distribution of β :

$$\beta|I_1 = \beta_1 + \frac{N(\mathbf{0}, \sigma_1^2 R_1^{-1})}{\sqrt{\chi_{\nu_1}^2/\nu_1}}$$

which is a non-central multivariate t -distribution [2, page 139], and it can be used to provide credibility regions or PHD regions. For instance, with t denoting the upper $\alpha/2$ quantile of a Student's t distribution on ν_1 degrees of freedom, the $(1 - \alpha)100\%$ PHD interval for β_i is given by:

$$\beta_{1i} \pm t\sqrt{\sigma_1^2 v_i}$$

where v_i denotes the i th diagonal element of $(R_0 + X^T X)^{-1}$. Further details can be found for instance in [21, Ch 9]. Prior uncertainty can be modeled by assuming $R_0 \approx O$, $\nu_0 = -(p+1)$ and $\sigma_0^2 = 0$, so that

$$\tau^2 | I_1 \sim \chi_{n-p-1}^2 / rss \quad (13)$$

$$\boldsymbol{\beta} | \tau^2, I_1 \sim N(\hat{\boldsymbol{\beta}}, (\tau X^T X)^{-1}) \quad (14)$$

and

$$\nu_1 = n - p - 1 \quad R_1 = X^T X \quad E(\boldsymbol{\beta} | I_1) = \hat{\boldsymbol{\beta}} \quad E(\sigma^2 | I_1) = rss / (n - p - 1).$$

In this way, the Bayesian estimates of σ^2 and $\boldsymbol{\beta}$ reduce to the classical MLE.

Example 6 (Simple Linear Regression) Suppose $c = 1$, and that we are interested in a simple linear regression model for Y on X . We assume that

$$Y | (\boldsymbol{\beta}, \tau) \sim N(\mu = \beta_0 + \beta_1 x, \tau^{-2})$$

and the parameters are assigned the prior distribution:

$$\tau^2 | I_0 \sim \chi_{\nu_0}^2 / (\nu_0 \sigma_0^2)$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} | \tau^2, I_0 \sim N \left(\begin{pmatrix} \beta_{00} \\ \beta_{01} \end{pmatrix}, \left(\tau^2 \begin{pmatrix} r_{00} & r_{01} \\ r_{01} & r_{11} \end{pmatrix} \right)^{-1} \right).$$

With a sample of n independent observations given $\boldsymbol{\theta} = (\tau^2, \beta_0, \beta_1)$, the posterior distribution of the parameters is

$$\tau^2 | I_1 \sim \chi_{\nu_1}^2 / (\nu_1 \sigma_1^2); \quad \boldsymbol{\beta} | \tau^2, I_1 \sim N \left(\boldsymbol{\beta}_1, (\tau^2 R_1)^{-1} \right)$$

with

$$R_1 = \begin{pmatrix} r_{00} + n & r_{01} + n\bar{x} \\ r_{01} + n\bar{x} & r_{11} + \sum_i x_i^2 \end{pmatrix}.$$

A choice $R_0 \approx O$, $\nu_0 = -2$ and $\sigma_0^2 = 0$ yields

$$\tau^2 | I_1 \sim \chi_{n-2}^2 / r_{ss} \quad \boldsymbol{\beta} | \tau^2, I_1 \sim N(\hat{\boldsymbol{\beta}}, (\tau^2 X^T X)^{-1})$$

and inference on the regression parameters is based on the Student's t distributions:

$$\frac{\beta_0 - \beta_{10}}{\sqrt{r_{ss}v_1/(n-2)}} \sim t_{n-2} \quad \frac{\beta_1 - \beta_{11}}{\sqrt{r_{ss}v_2/(n-2)}} \sim t_{n-2}$$

where v_1, v_2 are the diagonal elements of $R_1^{-1} = (X^T X)^{-1}$.

The posterior distribution of the parameters can also be used to predict new cases $\tilde{\mathbf{y}} | \boldsymbol{\theta} \sim N(\tilde{X}\boldsymbol{\beta}, \tau^{-2}I_m)$. The analysis is based on the evaluation of the conditional

distribution of $\tilde{\mathbf{y}} | \mathbf{y}$, which again involves the use of a multivariate Student's t distribution. Details are given for instance in [21, Ch 9], and application to real data-sets are provided by [11]. Generalization of the approach described in this section to models for correlated observations involve the use of “matrix-form” distributions as Wishart and can be found in [2, 3].

3.1.2 Model Selection

Our task becomes more challenging when we are interested in discovering the statistical model best fitting the available information. Let $\mathcal{M} = \{M_0, M_1, \dots, M_m\}$ be the set of models that are believed a priori to contain the true model of dependence of \mathbf{y} on X_1, \dots, X_c . For instance with $c = 1$, the set \mathcal{M} can contain the nested models:

$$\begin{aligned} M_0 : \mu &= \beta_0 \\ M_1 : \mu &= \beta_0 + \beta_1 x \\ M_2 : \mu &= \beta_0 + \beta_1 x + \beta_2 x^2. \end{aligned}$$

Each model induces a parameterization $M_i \rightarrow \boldsymbol{\theta}^{(i)} = (\boldsymbol{\beta}^{(i)}, \tau^2)$, e.g. $\boldsymbol{\beta}^{(0)} = \beta_0$, $\boldsymbol{\beta}^{(1)} = (\beta_0, \beta_1)^T$, $\boldsymbol{\beta}^{(2)} = (\beta_0, \beta_1, \beta_2)^T$ in the example above. Prior information I_0 allows the elicitation of prior probabilities of M_0, \dots, M_m , and, conditional on M_i , of prior distributions for $\boldsymbol{\theta}^{(i)}$. Graphically, this is equivalent to assuming a further node corresponding to M in Figure 3, with links towards τ^2 and $\boldsymbol{\beta}^{(i)}$, and the parameterization induced by each model is represented by a different plateau, as in Figure 4. Note that the parameters $\boldsymbol{\beta}^{(i)}$ are conditionally independent given M and τ^2 . Suppose that we wish to use the prior and sample information to select a regression model M_i from \mathcal{M} . We can use the sample information to compute the posterior probability of M_i given the data and the prior information I_0 :

$$p(M_i | I_1) = \frac{p(M_i | I_0)p(\mathbf{y} | M_i)}{p(\mathbf{y} | I_0)},$$

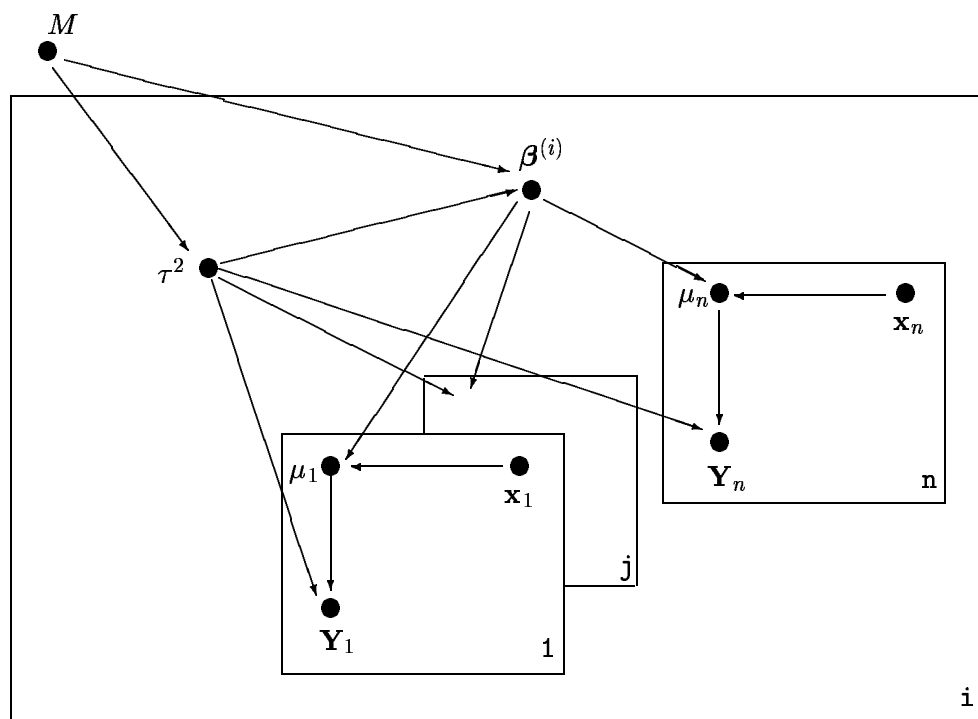


Figure 4: Graphical representation of a regression model with independent observations given $\theta^{(i)} = (\beta, \tau^2)$, conditional on a model.

and then we choose the model with the largest posterior probability. When the comparison is between two rival models M_1 and M_2 with $p(M_1) = p(M_2)$, this is equivalent to choose M_1 if the Bayes factor:

$$\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

is greater than one. The quantity $p(\mathbf{y}|M_i)$ is the *marginal likelihood* (marginal with respect to $\boldsymbol{\theta}^{(i)}$) of the data given the model M_i , which is computed as:

$$p(\mathbf{y}|M_i) = \int p(\mathbf{y}|\boldsymbol{\beta}^{(i)}, \tau^2) p(\boldsymbol{\beta}^{(i)}, \tau^2|I_0) d\boldsymbol{\beta}^{(i)} d\tau^2$$

and the integral has a closed form solution:

$$p(\mathbf{y}|M_i) = \frac{\det(R_0^{(i)})^{1/2} (\nu_0 \sigma_0^2)^{\nu_0/2} \Gamma(\nu_1/2)}{\det(R_1^{(i)})^{1/2} (\nu_1 \sigma_1^2)^{\nu_1/2} \Gamma(\nu_0/2) \pi^{n/2}}$$

where the indexes 0 and 1 specify the prior and posterior hyper-parameters of $\boldsymbol{\beta}^{(i)}$ and τ^2 . A complete treatment of this problem can be found in [21, Ch 9].

When the inference task is limited the prediction of future cases, a weighted average of the models in \mathcal{M} can be used instead of one single model [15].

3.2 Generalized Linear Models

Generalized Linear Models (GLM) provide a unified framework to encompass several situations which are not adequately described by the assumptions of normality of the data and linearity in the parameters. The components of a GLM are (i) the linear predictor $\eta = f(\mathbf{x}_i)^T \boldsymbol{\beta}$, (ii) the link function $g(\mu) = \eta$, relating the expectation of the data μ given the parameters to the linear predictor, and (iii) the distribution of $Y|\boldsymbol{\theta}$ which belongs to the exponential family (see Chapter 3). The parameter vector is made up of $\boldsymbol{\beta}$ and of the dispersion parameter ϕ . The problem with a Bayesian analysis of GLMs is that the posterior distribution of $\boldsymbol{\theta}$ cannot be calculated exactly, since the marginal density of the data

$$p(\mathbf{y}|I_0) = \int p(\mathbf{y}|I_0, \boldsymbol{\theta}) p(\boldsymbol{\theta}|I_0) d\boldsymbol{\theta}$$

cannot be usually evaluated in closed form, as the next example shows.

Example 7 (Logistic regression) *Suppose that, conditional on the vector of parameters $\boldsymbol{\theta} = \boldsymbol{\beta}$, data have a Binomial distribution with $p(Y = 1|\boldsymbol{\theta}) = \mu$. The dispersion parameter is $\phi = 1$. The logit function is the canonical link (see Chapter 3)*

$$g(\mu_i) = \log \frac{\mu_i}{1 - \mu_i} = \eta_i = f(\mathbf{x}_i)^T \boldsymbol{\beta}.$$

Let $p(\boldsymbol{\beta}|I_0)$ be the prior density. With a sample in which n cases, corresponding to n combinations of values of the covariates and supposed to be conditionally independent given $\boldsymbol{\beta}$, are observed, the likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \mu_i^{y_i} (1 - \mu_i)^{1-y_i} = \prod_{i=1}^n \frac{e^{\eta_i y_i}}{1 + e^{\eta_i}}$$

and the marginal density of the data solves

$$\int \prod_{i=1}^n \frac{e^{\eta_i y_i}}{1 + e^{\eta_i}} p(\boldsymbol{\beta}|I_0) d\boldsymbol{\beta}. \quad (15)$$

To date, there are no known prior distributions which lead to a closed form solution of (15).

Numerical integration techniques [21] can be exploited to approximate (15), from which a numerical approximation of the posterior density of $\boldsymbol{\beta}$ can be found. The basic idea is to select a grid of points and replace the value of the integral by a weighted sum of values of the integrand function evaluated in the grid points $\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_g\}$:

$$\int \prod_{i=1}^n \frac{e^{\eta_i y_i}}{1 + e^{\eta_i}} p(\boldsymbol{\beta}|I_0) d\boldsymbol{\beta} \approx \sum_j w_j \prod_{i=1}^n \frac{e^{\eta_{ij} y_i}}{1 + e^{\eta_{ij}}} p(\boldsymbol{\beta}_j|I_0).$$

However, as the dimension of the parameter space increases, numerical integration becomes infeasible, because it is difficult to select a suitable grid of points and it is hard to evaluate the error of the approximation [8].

3.3 Approximate Methods

When numerical integration techniques are no longer available, we are left with two main ways to perform approximate posterior analysis: provide an asymptotic approximation of the posterior distribution or use stochastic methods to generate a sample from the posterior distribution.

3.3.1 Asymptotic Posterior Distributions

When the sample size is large enough, posterior analysis can be based on an asymptotic approximation of the posterior distribution to a Normal distribution with some mean and variance. This idea generalizes the asymptotic approximate distribution of MLEs when their exact sampling distribution cannot be found or it is too difficult to be used (see Chapter 2.)

Berger [1, page 224] mentions four approximate Normal distributions, which differ in the way the mean and variance are computed and return approximations of decreasing accuracy:

1. $\boldsymbol{\theta}|I_1$ is approximately $N(\boldsymbol{\theta}_1, V(\boldsymbol{\theta}|I_1))$, where $\boldsymbol{\theta}_1$ and $V(\boldsymbol{\theta}|I_1)$ are the exact posterior mean and variance of $\boldsymbol{\theta}$.
2. $\boldsymbol{\theta}|I_1$ is approximately $N(\hat{\boldsymbol{\theta}}_1, (I(\hat{\boldsymbol{\theta}}_1|I_1))^{-1})$, where $\hat{\boldsymbol{\theta}}_1$ is the GMLE of $\boldsymbol{\theta}$, i.e. $\hat{\boldsymbol{\theta}}_1$ maximizes the augmented likelihood $L(\boldsymbol{\theta})p(\boldsymbol{\theta}|I_0) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|I_0)$ and the matrix $I(\hat{\boldsymbol{\theta}}_1|I_1)$ is the $k \times k$ (k is the dimension of the parameter vector) matrix having element i, j :

$$I(\hat{\boldsymbol{\theta}}_1|I_1)_{ij} = - \left| \frac{\partial^2 \log\{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|I_0)\}}{\partial\theta_i\partial\theta_j} \right|_{\hat{\boldsymbol{\theta}}_1}$$

evaluated in the GMLEs.

3. $\boldsymbol{\theta}|I_1$ is approximately $N(\hat{\boldsymbol{\theta}}, (I(\hat{\boldsymbol{\theta}}|\mathbf{y}))^{-1})$, where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, i.e. $\hat{\boldsymbol{\theta}}$ maximizes the likelihood $L(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$ and the matrix $I(\hat{\boldsymbol{\theta}}|\mathbf{y})$ is the *observed* Fisher information matrix, i.e. it is the $k \times k$ matrix having element i, j :

$$I(\hat{\boldsymbol{\theta}}|\mathbf{y})_{ij} = - \left| \frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} \right|_{\hat{\boldsymbol{\theta}}}$$

evaluated in the MLEs.

4. $\boldsymbol{\theta}|I_1$ is approximately $N(\hat{\boldsymbol{\theta}}, (I(\hat{\boldsymbol{\theta}}))^{-1})$, where the matrix $I(\hat{\boldsymbol{\theta}})$ is the *expected* Fisher information matrix, i.e. it is the $k \times k$ matrix having element i, j :

$$I(\hat{\boldsymbol{\theta}})_{ij} = -E \left| \frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} \right|_{\hat{\boldsymbol{\theta}}}$$

evaluated in the MLEs, and the expectation is over the conditional distribution of the data given $\boldsymbol{\theta}$.

The first approximation is the most accurate, as it preserves the exact moments of the posterior distribution. However, this approximation relies on the computation of the exact first and second moments [21, Ch 8]. When exact first and second moments cannot be computed in closed form, we must resort to the second approximation, which replaces them by approximations based on the maximization of the augmented likelihood. Hence, the prior information is taken into account in the calculations of both moments. As the sample size increases and the prior information becomes negligible, the second and third approximation become equivalent. The fourth approximation is the least accurate and relies on the idea that, for exponential family models, expected and observed Fisher information matrices are identical, since the matrix of second derivatives depends on the data only via the MLEs.

Asymptotic normality of the posterior distribution provides notably computational advantages, since marginal and conditional distributions are still Normal, and hence inference on parameters of interest can be easily carried out. However, for relatively small samples, the assumption of asymptotic normality can be inaccurate.

3.3.2 Stochastic Methods

For relatively small samples, stochastic methods (or Monte Carlo methods) provide an approximate posterior analysis based on a sample of values generated from the posterior distribution of the parameters. Posterior analysis requires the evaluations of integrals:

$$E(g(\boldsymbol{\theta}|I_1)) = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|I_1)d\boldsymbol{\theta}.$$

For instance, for $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ we compute the posterior expectation. If exact integration is not possible, Monte Carlo methods replace the exact integral $E(g(\boldsymbol{\theta}|I_1))$ by $\sum_i g(\boldsymbol{\theta}_i)/s$ where $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_s$ is a random sample of size s generated from $\boldsymbol{\theta}|I_1$. Thus, the task reduces to generating a sample from the posterior distribution of the parameters. Here we will describe Gibbs Sampling (GS), a special case of Metropolis-Hastings algorithms [12], which is becoming increasingly popular in the statistical community. GS is an iterative method that produces a Markov Chain, that is a sequence of values

$$\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)} \dots\}$$

such that $\boldsymbol{\theta}^{(i+i)}$ is sampled from a distribution that depends on the current state i of the chain. The algorithm works as follows.

Let $\boldsymbol{\theta}^{(0)} = \{\theta_1^{(0)}, \dots, \theta_k^{(0)}\}$ be a vector of initial values for the parameter vector $\boldsymbol{\theta}$ and suppose that the conditional distributions of $\theta_i | (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k)$ are known for each i . The first value in the chain is simulated as follows:

$$\begin{aligned} \theta_1^{(1)} & \text{ is sampled from the conditional distribution of } \theta_1 | (\theta_2^{(0)}, \dots, \theta_k^{(0)}, \mathbf{y}); \\ \theta_2^{(1)} & \text{ is sampled from the conditional distribution of } \theta_2 | (\theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, \mathbf{y}) \\ & \vdots \\ \theta_k^{(1)} & \text{ is sampled from the conditional distribution of } \theta_k | (\theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)}, \mathbf{y}) \end{aligned}$$

Then $\boldsymbol{\theta}^{(0)}$ is replaced by $\boldsymbol{\theta}^{(1)}$ and the simulation repeated to generate $\boldsymbol{\theta}^{(2)}$, and so forth. In general, the i th value in the chain is generated by simulating from the distribution of $\boldsymbol{\theta}$ conditional on the value previously generated $\boldsymbol{\theta}^{(i-1)}$. After an initial long iteration, called *burn-in*, of say b iterations, the values

$$\{\boldsymbol{\theta}^{(b+1)}, \boldsymbol{\theta}^{(b+2)}, \boldsymbol{\theta}^{(b+3)} \dots\}$$

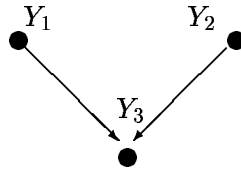


Figure 5: A simple BBN.

will be approximately a sample from the posterior distribution of θ , from which empirical estimates of the posterior means and any other function of the parameters can be computed as

$$\frac{1}{s-b} \sum_{i=b+1}^s g(\theta^{(i)})$$

where s is the total length of the chain. Critical issues for this method are the choice of the starting value $\theta^{(0)}$, the length of the burn-in and the selection of a stopping rule. The reader is referred to [13] for a discussion of these problems. An implementation of GS suitable for problems in which the likelihood function satisfies certain factorization properties is the program BUGS [27]. The program manual [25] contains several interesting illustrative examples.

4. Bayesian Networks

The graphical models that we have used to represent dependencies between data and parameters associated with the sampling model can be further used to describe directed associations among sets of variables. When used in this way, such models are known as Bayesian Belief Networks (BBN) and they result in a powerful knowledge representation formalism based on probability theory widely use in Artificial Intelligence. In this section, we will outline the foundations of BBNS and we will describe how to use BBNS to analyze and model data.

4.1 Foundations

Formally, a BBN is defined by a set of variables $\mathcal{Y} = \{Y_1, \dots, Y_T\}$ and a DAG defining a model M of conditional dependencies among the elements of \mathcal{Y} . We will consider discrete variables and denote by c_i the number of states of Y_i and by y_{ik} a state of Y_i . A conditional dependency links a variable Y_i to a set of *parent* variables Π_i , and it is defined by the conditional distributions of Y_i given each *configuration* $\pi_{i1}, \dots, \pi_{iq_i}$ of the parent variables.

The directed graphical structure encodes the assumption that a variable Y_i is conditionally independent of all the variables $\{Y_1, Y_2, \dots, Y_{i-1}\}$ given its parents, and hence, it yields a factorization of the joint probability of a set of values $\mathbf{y}_k = \{y_{1k}, \dots, y_{Ik}\}$ of the variables in \mathcal{Y} as

$$p(\mathbf{y}_k) = \prod_{i=1}^I p(y_{ik} | \pi_{ij}), \quad (16)$$

where π_{ij} denotes the configuration of states of Π_i in \mathbf{y}_k .

Example 8 (A simple BBN) Consider the BBN in Figure 5, in which the set \mathcal{Y} is $\{Y_1, Y_2, Y_3\}$ and $c_i = 2$ for $i = 1, 2, 3$. The graph encodes the marginal independence of Y_1 and Y_2 , which in turn are both parents of Y_3 . Thus

$$\Pi_1 = \Pi_2 = \emptyset, \quad \Pi_3 = (Y_1, Y_2).$$

Note that Π_3 takes four values π_{ij} corresponding to the four combinations of states of Y_1 and Y_2 . We will denote these four states as $\pi_{31} = (y_{11}, y_{21})$, $\pi_{32} = (y_{11}, y_{22})$, $\pi_{33} = (y_{12}, y_{21})$ and $\pi_{34} = (y_{12}, y_{22})$. The joint probability of a case $\mathbf{y}_k = \{y_{11}, y_{21}, y_{32}\}$ can then be written as

$$p(\mathbf{y}_k) = p(y_{11})p(y_{21})p(y_{32} | y_{11}, y_{21}) = p(y_{11})p(y_{21})p(y_{32} | \pi_{31}).$$

If Y_3 has a child variable, say Y_4 , then the separation of Y_4 from Y_1, Y_2 implies that $i(Y_4, (Y_1, Y_2) | Y_3)$, and the joint probability of $\mathbf{y}_k = \{y_{1k}, y_{2k}, y_{3k}, y_{4k}\}$ factorizes into

$$p(\mathbf{y}_k) = p(y_{1k})p(y_{2k})p(y_{3k} | \pi_{3j})p(y_{4k} | \pi_{4j}).$$

Thus, the graphical component of a BBN provides a simple way to describe and communicate stochastic dependencies among variables. There is a tendency, see for instance the review in [15], to interpret the directed links in terms of causal association between variables, although risks related to this interpretation are well known. The conditional independence assumptions encoded by a BBN have the further advantage that computations with probabilities are greatly simplified, so that tasks as prediction, explanation, classification, decision making can be done very efficiently, as shown in the next two examples.

Example 9 (Representation) Consider the BBN in Figure 6, in which the variables are all binary. Directed links identify parent-child dependencies, and hence:

$$\begin{aligned} \Pi_1 &= \emptyset & \Pi_2 &= Y_1 \\ \Pi_3 &= Y_2 & \Pi_4 &= Y_2 \\ \Pi_5 &= (Y_3, Y_4) & \Pi_6 &= (Y_3, Y_5) \end{aligned}$$

The graph encodes the following conditional independence assumptions:

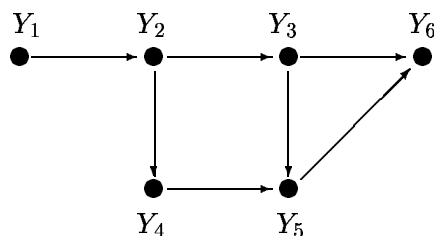


Figure 6: A BBN with 6 binary variables.

1. $i(Y_3, Y_1|Y_2)$
2. $i(Y_4, Y_1|Y_2)$
3. $i(Y_5, (Y_1, Y_2)|(Y_3, Y_4))$
4. $i(Y_6, (Y_1, Y_2, Y_4)|(Y_3, Y_5))$.

Thus, the joint probability of one case

$$\mathbf{y}_k = (y_1, y_{2k}, y_{3k}, y_{4k}, y_{5k}, y_{6k})$$

can be decomposed into a set of independent parent-child contributions as:

$$p(\mathbf{y}_k) = p(y_1)p(y_{2k}|y_{1k})p(y_{3k}|y_{2k})p(y_{4k}|y_{2k})p(y_{5k}|y_{3k}, y_{4k})p(y_{6k}|y_{3k}, y_{5k}).$$

The advantage of this description is that the joint probability of the 6 variables would require $2^6 - 1 = 63$ independent numbers, that are reduced to $1 + 2 + 2 + 2 + 4 + 4 = 15$ when the conditional independence assumptions 1–4 are exploited.¹

This graphical representation provides a compact and powerful formalism to efficiently extract information of interest from a statistical (graphical) model (see Chapter 3.) In a more general sense, BBNS provide a knowledge representation and a reasoning method for problem solving based on probabilistic information. Using BBNS, we can easily make predictions about the value of a variable in a given situation by computing the conditional probability distribution of the variable of interest given the values of a set of some other variables in the network describing the situation at hand. Suppose, for instance, that we are interested in the value of variable Y_6 when the variables Y_3 and Y_4 are observed to be in the states y_{31} and y_{41} . By the Total Probability Theorem

¹Since the variables are binary, each conditional distribution is identified by one number, and hence $1 + 2 + 2 + 2 + 4 + 4$ is the sum of conditional distributions defined by the parent configurations.

$$\begin{aligned} p(y_{61}|y_{31}, y_{41}) &= \sum_j p(y_{61}, y_{5j}|y_{31}, y_{41}) \\ &= \sum_j p(y_{5j}|y_{31}, y_{41})p(y_{61}|y_{5j}, y_{31}). \end{aligned}$$

Thus, the conditional probability of interest is expanded to include all variables between the conditioning variables Y_3 and Y_4 , and Y_6 , and then factorized to account for the conditional independence assumptions encoded by the DAG. Within this framework, the marginal probability $p(y_{61})$ that would be computed by marginalizing the joint probability:

$$p(y_{61}) = \sum_{ijkmnr} p(y_{61}, y_{5j}, y_{4m}, y_{3k}, y_{2n}, y_{1r})$$

turns out to be nothing but the conditional probability of y_{61} given the empty set, and it can be computed as:

$$\begin{aligned} p(y_{61}) &= \sum_{jk} p(y_{61}|y_{5j}, y_{3k}) \sum_m p(y_{5j}|y_{3k}, y_{4m}) \sum_n p(y_{3k}|y_{2n})p(y_{4m}|y_{2n}) \times \\ &\quad \sum_r p(y_{2n}|y_{1r})p(y_{1r}) \end{aligned}$$

by taking advantage of the locality structure of the parent-child configurations.

A network structure of particular interest for data analysis is displayed in Figure 7 and it is known as the Naive Bayesian Classifier.

Example 10 (Naive Bayesian Classifier) *The BBN in Figure 7 represents a Naive Bayesian Classifier, in which a set of mutually exclusive classes is represented as the root node, and the attributes of the objects to be classified depend on this variable. The simplifying assumptions made by this model which brings it the name of "naive" are that the classes are mutually exclusive and that the attributes are independent given the class. The meaning of this assumption is that, once we know the class to which an object belongs, the relations between its attributes become irrelevant. In the example depicted in Figure 7, the root node Y_1 represent the set of mutually exclusive classes and the leaf nodes Y_2 and Y_3 are the attributes. As the model assumes $i(Y_2, Y_3|Y_1)$, the joint probability distribution is decomposed into*

$$p(y_{1k}, y_{2k}, y_{3k}) = p(y_{1k})p(y_{2k}|y_{1k})p(y_{3k}|y_{1k}).$$

Although the underlying structure is so simple to be termed "naive", this model performs extremely well in classification problems with a large number of attributes [9] in which the

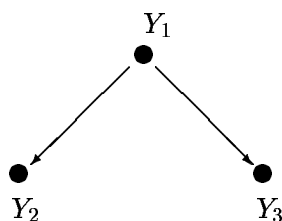


Figure 7: A Naive Bayesian Classifier.

task is to classify a unit presenting a combination of attribute values into one of the states of the variable Y_1 .

The conditional independence assumptions represented by the model allow the classification to be performed in a very efficient way. Suppose that a unit with attributes y_{2j} and y_{3k} is to be assigned to one of the states of the variable Y_1 . The solution is to compute the posterior probability $p(y_{1i}|y_{2j}, y_{3k})$ for all i , and then the state with the largest posterior probability is chosen. Thus, the problem reduces to computing $p(y_{1i}|y_{2j}, y_{3k})$ which is given by:

$$p(y_{1i}|y_{2j}, y_{3k}) = \frac{p(y_{1i})p(y_{2j}|y_{1i})}{\sum_r p(y_{2j}|y_{1r})p(y_{1r})} \frac{p(y_{3k}|y_{1i})}{\sum_r p(y_{3k}|y_{1r})p(y_{1r}|y_{2j})}.$$

The factorization into terms that depend on the associations Y_1, Y_2 , and Y_1, Y_3 is an advantage of Bayesian methods described in section 1. We first have a flow of information from Y_2 to Y_1 by applying Bayes' Theorem:

$$\frac{p(y_{1i})p(y_{2j}|y_{1i})}{\sum_r p(y_{2j}|y_{1r})p(y_{1r})} = p(y_{1i}|y_{2j}).$$

After this first updating, the probability distribution of the class node is $P(Y_1|y_{2j})$, and this is used as prior distribution in next step, to process the information incoming from node Y_3 :

$$\frac{p(y_{3k}|y_{1i})p(y_{1i}|y_{2j})}{\sum_r p(y_{3k}|y_{1r})p(y_{1r}|y_{2j})}.$$

This can be clearly extended to the case of several attributes, and the computation of the conditional probability of Y_1 given a combination of attribute values is found by processing the information incoming from one attribute node at a time.

In this special case, the network structure allows an efficient computation of the posterior distribution of the variable of interest, in linear time with respect to the number of

attributes. Unfortunately, this is not the case for any DAG. Nonetheless, more general (but less efficient) algorithms are available to compute a conditional probability distribution in a generic DAG. The interested reader can find a review of some of these algorithms in [22, 6].

4.2 Learning Bayesian Networks From Data

In their original concept, BBNS were supposed to rely on domain experts to supply the required information: their conditional independence assumptions, when coupled with the subjective assessment of the conditional probability distributions associated to the dependencies, produces a sound and compact probabilistic representation of the domain knowledge. However, the statistical roots of BBNS soon led to the development of learning methods to extract BBNS directly from databases of cases rather than from the insight of human domain experts [7, 5, 16], thus turning BBNS into a powerful tool for the analysis of data.

Suppose we are given a sample of n cases $\mathbf{y} = \{y_1, \dots, y_n\}$ from which we wish to induce a BBN. Note that the sample is now multivariate, since each case in the sample

$$\mathbf{y}_k = (y_{1k}, \dots, y_{Ik})$$

corresponds to a combination of states of the I variables. Thus, \mathbf{y} is a $n \times I$ matrix. Two components of a BBN can be learned: the graphical structure M , specifying the conditional independence assumptions among the variables in \mathcal{Y} , and, given a graph M , the conditional probabilities associated to the remaining dependencies in the graph. We first suppose the graphical structure M given and we focus attention on the second task. The selection of the graphical model M will be considered later.

4.2.1 Parameter Estimation

Given a DAG M , the conditional probabilities defining the BBN are the parameters $\boldsymbol{\theta} = (\theta_{ijk})$, where $\theta_{ijk} = p(y_{ik}|\pi_{ij}, \boldsymbol{\theta})$, that we wish to estimate from \mathbf{y} . We shall denote by $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \dots, \theta_{ijc_i})$ the parameter vector associated to the conditional distribution of $Y_i|\pi_{ij}$, to be inferred from \mathbf{y} . Graphically, we can expand the BBN by adding, to each variable Y_i , new parent variables representing the parameters that quantify the conditional distribution of $Y_i|\pi_{ij}$.

Example 8 (continued). The BBN in Figure 5 can be expanded into the BBN in Figure 8 by adding the parameters that quantify the dependencies. Thus, six parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_{31}, \theta_{32}, \theta_{33}, \theta_{34})$ are needed. Since the variables Y_1 and Y_2 are binary, their marginal distributions are defined by two parameters: $\theta_1 = p(y_{11}|\boldsymbol{\theta})$ and $\theta_2 = p(y_{21}|\boldsymbol{\theta})$. From the two distributions of Y_1 and Y_2 , we can define the joint distribution of the parent variable Π_3 . The distribution of Π_3 is specified by θ_1 and θ_2 as $p(\pi_{31}|\boldsymbol{\theta}) = \theta_1\theta_2$; $p(\pi_{32}|\boldsymbol{\theta}) = \theta_1(1 - \theta_2)$; $p(\pi_{33}|\boldsymbol{\theta}) = (1 - \theta_1)\theta_2$ and $p(\pi_{34}|\boldsymbol{\theta}) = (1 - \theta_1)(1 - \theta_2)$. Finally, each parent configuration π_{3j} defines a conditional distribution $Y_3|\pi_{3j}$. Since Y_3 is binary, each of these conditional

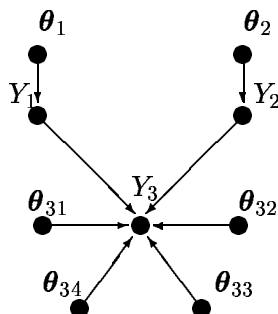


Figure 8: A Simple BBN augmented by parameters.

distributions is identified by one parameter: $\theta_{3j1} = p(y_{31}|\pi_{3j}, \boldsymbol{\theta})$ for $j = 1, 2, 3, 4$. From these parameters, we obtain the parameter vectors $\boldsymbol{\theta}_1 = (\theta_{11}, 1 - \theta_{11})$ associated to the distribution of Y_1 , $\boldsymbol{\theta}_2 = (\theta_{21}, 1 - \theta_{21})$ associated to the distribution of Y_2 , $\boldsymbol{\theta}_{31} = (\theta_{311}, 1 - \theta_{311})$ associated to the distribution of $Y_3|\pi_{31}$, $\boldsymbol{\theta}_{32} = (\theta_{321}, 1 - \theta_{321})$ associated to the distribution of $Y_3|\pi_{32}$, $\boldsymbol{\theta}_{33} = (\theta_{331}, 1 - \theta_{331})$ associated to the distribution of $Y_3|\pi_{33}$ and $\boldsymbol{\theta}_{34} = (\theta_{341}, 1 - \theta_{341})$ associated to the distribution of $Y_3|\pi_{34}$. We have assumed mutual independence among $\boldsymbol{\theta}_{ij}$, although by no means this is a necessary condition. Mutual independence, however, leads to a great simplification of the estimation procedure.

The standard Bayesian method to estimate $\boldsymbol{\theta}$ uses conjugate analysis. Let $n(y_{ik}|\pi_{ij})$ be the frequency of (y_{ik}, π_{ij}) in the sample, and let $n(\pi_{ij}) = \sum_k n(y_{ik}|\pi_{ij})$ be the frequency of π_{ij} . The joint probability of a case \mathbf{y}_k can be written as a function of the unknown θ_{ijk} as

$$p(\mathbf{y}_k|\boldsymbol{\theta}) = \prod_{i=1}^I \theta_{ijk},$$

and hence, if cases are independent, the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{k=1}^n p(\mathbf{y}_k|\boldsymbol{\theta}) = \prod_{i=1}^I \prod_{j=1}^{q_i} \prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})}.$$

Thus, $p(\mathbf{y}|\boldsymbol{\theta})$ factorizes into a product of local parent-child contributions:

$$\prod_{j=1}^{q_i} \prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})}$$

and each of these terms is itself a product of terms that depend on the parent configurations:

$$\prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})}.$$

A common assumption [26] which matches the factorization of the likelihood into parent-child contributions and leads to local computation is that the parameter vectors $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\theta}_{i'j'}$ associated to different variables Y_i and $Y_{i'}$ are independent for $i \neq i'$ (*global independence*). If the parameters $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\theta}_{i'j'}$ associated to the distributions of Y_i , given different parent configurations π_{ij} and $\pi_{i'j'}$ ($j \neq j'$), are further assumed to be independent (*local independence*), then the joint prior density $p(\boldsymbol{\theta}|I_0)$ factorizes into the product

$$p(\boldsymbol{\theta}|I_0) \propto \prod_{ij} p(\boldsymbol{\theta}_{ij}|I_0).$$

If the sample \mathbf{y} is complete, and hence there are not entries reported as unknown, local and global independence induce an equivalent factorization of the posterior density of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|I_1) \propto \prod_{ij} \{p(\boldsymbol{\theta}_{ij}|I_0) \prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})}\}$$

and this allows us to independently update the distribution of $\boldsymbol{\theta}_{ij}$, for all i, j . A further saving in computation is achieved if, for all i and j , the prior distribution of $\boldsymbol{\theta}_{ij}$ is the conjugate *Dirichlet* distribution with *hyper-parameters* $\{\alpha_{ij1}, \dots, \alpha_{ijc_i}\}$ with $\alpha_{ijk} > 0$. We use the notation

$$\boldsymbol{\theta}_{ij}|I_0 \sim D(\alpha_{ij1}, \dots, \alpha_{ijc_i}).$$

In this case, the prior density of $\boldsymbol{\theta}_{ij}$ is, up to a proportionality constant,

$$p(\boldsymbol{\theta}_{ij}|I_0) \propto \prod_k \theta_{ijk}^{\alpha_{ijk}-1}$$

which is conjugate. Note that the density function generalizes that of a Beta distribution in Example 1, and indeed a Beta distribution is a bivariate Dirichlet. The prior hyper-parameters α_{ijk} encode the observer's prior belief and, since $\alpha_{ijk} - 1$ plays the role of $n(y_{ijk}|\pi_{ij})$ in the likelihood, they can be regarded as frequencies of imaginary cases needed to formulate the prior distribution. The quantity $\alpha_{ij} - c_i = \sum_{k=1}^{c_i} (\alpha_{ijk} - 1)$ represents the frequency of imaginary cases observed in the parent configuration π_{ij} and hence α_{ij} is the *local precision*. The effect of the quantities α_{ijk} s is to specify the marginal probability of $(y_{ik}|\pi_{ij})$ as

$$E(\theta_{ijk}|I_0) = \frac{\alpha_{ijk}}{\alpha_{ij}}$$

and the prior variance is

$$V(\theta_{ijk}|I_0) = \frac{E(\theta_{ijk})\{1 - E(\theta_{ijk})\}}{\alpha_{ij} + 1}.$$

Note that, for fixed $E(\theta_{ijk})$, $V(\theta_{ijk})$ is a decreasing function of α_{ij} , so that a small value of α_{ij} will denote great uncertainty. Furthermore, $\alpha = \sum_{ij} \alpha_{ij}$ is the *global precision* on θ . The situation of initial ignorance can be represented by assuming $\alpha_{ijk} = \alpha/(c_i q_i)$ for all i, j and k , so that the prior probability of $(y_{ik}|\pi_{ij})$ is simply $1/c_i$ [10]. An important property of the Dirichlet distribution is that it is closed under marginalization, so that if

$$\theta_{ij}|I_0 \sim D(\alpha_{ij1}, \dots, \alpha_{ijc_i})$$

then any subset of parameters $(\theta_{ij1}, \dots, \theta_{ijs}, 1 - \sum_{k=1}^s \theta_{ijk})$ will have a Dirichlet distribution $D(\alpha_{ij1}, \dots, \alpha_{ijs}, \alpha_{ij} - \sum_{k=1}^s \alpha_{ijk})$. In particular, the parameter θ_{ijk} will have a Beta distribution with hyper-parameters $\alpha_{ijk}, \alpha_{ij} - \alpha_{ijk}$. Thus, marginal inference on parameters of interest can be easily carried out. Spiegelhalter and Lauritzen [26] show that the assumptions of parameter independence and prior Dirichlet distributions imply that the posterior density of θ is still a product of Dirichlet densities and

$$\theta_{ij}|I_1 \sim D(\alpha_{ij1} + n(y_{i1}|\pi_{ij}), \dots, \alpha_{ijc_i} + n(y_{ic_i}|\pi_{ij}))$$

so that local and global independence are retained after the updating. The information conveyed by the sample can be therefore captured by simply updating the hyper-parameters of the distribution of θ_{ij} by increasing them of the frequency of cases (y_{ijk}, π_{ij}) observed in the sample. Thus, the sample information can be summarized into the contingency tables representing parent-child configurations. The posterior expectation of θ_{ijk} :

$$E(\theta_{ijk}|I_1) = \frac{\alpha_{ijk} + n(y_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})}$$

and the posterior mode is

$$\frac{\alpha_{ijk} + n(y_{ik}|\pi_{ij}) - 1}{\alpha_{ij} + n(\pi_{ij}) - c_i}.$$

The posterior variance is given by:

$$V(\theta_{ijk}|I_1) = \frac{E(\theta_{ijk}|I_1)\{1 - E(\theta_{ijk}|I_1)\}}{\alpha_{ij} + n(\pi_{ij}) + 1}$$

with a local precision on θ_{ij} which is increased by the frequency of parents observed in the configuration π_{ij} .

Example 10 (continued). Data in the contingency table 1 are extracted from the British General Election Panel Survey (April 1992). The frequencies are displayed according to Sex (Y_2 : 1=male and 2=female), Social Class (Y_3 : 1=low, 2=middle and 3=high), and the class variable Voting Intention (Y_1 : 1=Conservative, 2=Labour, 3=Liberal Democrat and 4=Other). The task is to classify the voters into these four mutually exclusive classes of

Y_2	Y_3	Y_1			
		1	2	3	4
1	1	28	8	7	0
	2	153	114	53	14
	3	20	31	17	1
2	1	1	1	0	1
	2	165	86	54	6
	3	30	57	18	4

Table 1: Data from the British General Election Panel Survey. Voting Intention (Y_1), Sex (Y_2), Social Class (Y_3).

voting intentions on the basis of their attributes (Sex and Social Class). For this purpose, we can therefore define a Naive Bayesian Classifier similar to the one displayed in Figure 7 and estimate the conditional probability distributions associated to its dependency.

Let $\theta_1 = (\theta_{11}, \dots, \theta_{14})$ be the parameter vector associated to the marginal distribution of Y_1 , and let $\theta_{2j} = (\theta_{2j1}, \theta_{2j2})$ and $\theta_{3j} = (\theta_{3j1}, \theta_{3j2}, \theta_{3j3})$ be the parameter vectors associated to the conditional distributions of $Y_2|y_{1j}$ and $Y_3|y_{1j}$, $j = 1, \dots, 4$. A global prior precision $\alpha = 12$ and the assumption of uniform prior probabilities for y_{1k} , $y_{2k}|y_{1j}$ and $y_{3k}|y_{1j}$ induce a prior distribution $D(3, 3, 3, 3)$ for the parameter θ_1 , prior distributions $D(1.5, 1.5)$ for θ_{2j} , and $D(1, 1, 1)$ for the parameters θ_{3j} . The frequencies in Table 1 are then used to update the hyper-parameters, so that after the updating, the posterior distributions of the parameters are:

$$\begin{aligned}
 \theta_1 | I_1 &\sim D(400, 300, 152, 29) \\
 \theta_{21} | I_1 &\sim D(202.5, 197.5) & \theta_{32} | I_1 &\sim D(30, 319, 51) \\
 \theta_{22} | I_1 &\sim D(154.5, 145.5) & \theta_{33} | I_1 &\sim D(10, 201, 89) \\
 \theta_{23} | I_1 &\sim D(78.5, 73.5) & \theta_{34} | I_1 &\sim D(8, 108, 36) \\
 \theta_{24} | I_1 &\sim D(16.5, 12.5) & &
 \end{aligned}$$

from which the probabilities of $y_{1k}|I_1$, $y_{2k}|y_{1j}, I_1$ and $y_{3k}|y_{1j}, I_1$ are computed as posterior expectations, and are reported in Figure 9. It is worth noting that the so defined BBN, when coupled with the classification procedure defined in Example 10, turns out to be a complete classification system: we can train the network with the data using the procedure just described and then classify future cases using the algorithm described in Example 10. Using the same procedure, we can also calculate the distributions of $Y_1|y_{2j}, y_{3k}, I_1$ as shown in Table 2, thus discovering that the fundamental attribute for classification turns out to be the Social Class (Y_3): high and middle class intend to vote for the Conservative party, while the lower social class has a clear preference for the Labour party.

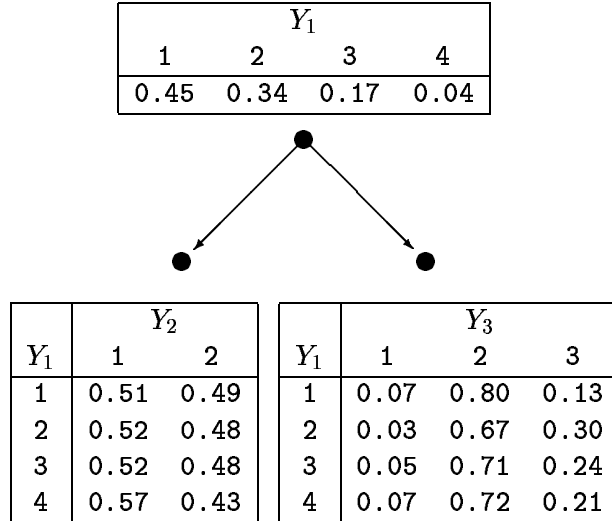


Figure 9: The BBN induced by the data in Table 1.

Y_2	Y_3	$P(Y_1 y_{2j}, y_{3k}, I_1)$				Classify as
		1	2	3	4	
1	1	0.59	0.20	0.16	0.05	1
1	2	0.49	0.31	0.17	0.03	1
1	3	0.28	0.49	0.20	0.03	2
2	1	0.61	0.20	0.16	0.03	1
2	2	0.50	0.31	0.16	0.03	1
2	3	0.28	0.49	0.20	0.03	2

Table 2: Classification of data in Table 1 using the BBN in Figure 9.

4.2.2 Model Selection

Suppose now that the graphical model M has to be induced from the data. As in Section 3, let $\mathcal{M} = \{M_0, M_1, \dots, M_m\}$ be the set of models that are believed a priori to contain the true model of dependence among the variables in \mathcal{Y} . For instance, let $\mathcal{Y} = \{Y_1, Y_2, Y_3\}$ and suppose that Y_1, Y_2 are known to be marginally independent, and that they can be both parents of Y_3 , but Y_3 cannot be parent of Y_1, Y_2 . These assumptions limit the set of possible models to be explored to $\mathcal{M} = \{M_0, M_1, M_2, M_{12}\}$ which are given in Figure 10.

Each model in \mathcal{M} is assigned a prior probability $p(M_j|I_0)$. Let $\boldsymbol{\theta}^{(j)}$ be the parameter vector associated to the conditional dependencies specified by M_j . The sample information is used to compute the posterior probabilities $p(M_j|I_1)$ from which the most probable model M can be selected. By Bayes' Theorem, we have

$$p(M_j|I_1) = \frac{p(M_j|I_0)p(\mathbf{y}|M_j)}{\sum_{i=1}^m p(M_i|I_0)p(\mathbf{y}|M_i)}$$

where $p(\mathbf{y}|M_j)$ is the marginal likelihood of M_j . Since the denominator is constant for different models, in the comparison between rival models M_i and M_j , M_i is chosen if

$$p(M_i|I_0)p(\mathbf{y}|M_i) > p(M_j|I_0)p(\mathbf{y}|M_j).$$

In order to select the most probable model, it is therefore sufficient to compute $p(\mathbf{y}|M_j)$ as

$$p(\mathbf{y}|M_j) = \int p(\boldsymbol{\theta}^{(j)}|M_j)p(\mathbf{y}|\boldsymbol{\theta}^{(j)})d\boldsymbol{\theta}^{(j)} \quad (17)$$

where $p(\boldsymbol{\theta}^{(j)}|M_j)$ is the prior density of $\boldsymbol{\theta}^{(j)}$ and $p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$ is likelihood function.

It is shown in [7] that (17) has a closed form solution when:

1. *The sample is complete, i.e. there are not cases reported as unknown;*
2. *The cases are independent, given the parameter vector $\boldsymbol{\theta}^{(j)}$ associated to M_j ;*
3. *The prior distribution of the parameters is conjugate to the sampling model $p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$, and hence $\boldsymbol{\theta}_{ij}^{(j)} \sim D(\alpha_{ij1}, \dots, \alpha_{ijc_i})$ and the parameters are locally and globally independent;*

and the marginal likelihood of M_j is:

$$p(\mathbf{y}|M_j) = \prod_{i=1}^I \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(y_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})} \quad (18)$$

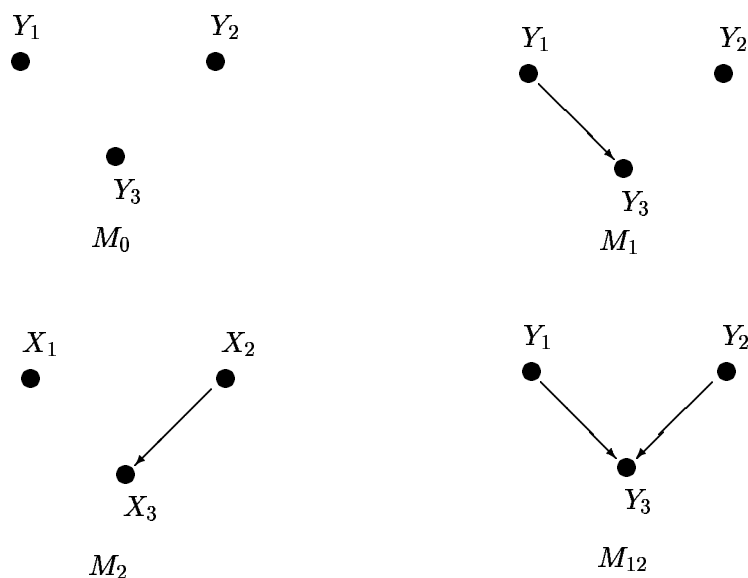


Figure 10: A set of possible models.

where $\Gamma(\cdot)$ is the Gamma function [29]. When the database is complete, (18) can be efficiently computed using the hyper-parameters $\alpha_{ijk} + n(y_{ik}|\pi_{ij})$ and the precisions $\alpha_{ij} + n(\pi_{ij})$ of the posterior distributions of θ_{ij} .

Example 11 (Discrimination between two models) Suppose we have two categorical variables Y_1 and Y_2 , and a random sample of n cases. Both Y_1 and Y_2 are binary variables, and it is known that Y_1 cannot be a parent of Y_2 . This assumption leaves us with two possible models to be explored:

Model M_0 specifies that the two variables are independent and, conditional on M_0 , we can parameterize $p(y_{11}|\theta^{(0)}) = \theta_{11}$ and $p(y_{21}|\theta^{(0)}) = \theta_{21}$ where $\theta^{(0)}$ is the parameter vector associated to M_0 .

Model M_1 specifies that Y_2 is a parent of Y_1 , so that the induced parameterization is $p(y_{21}|\theta^{(1)}) = \theta_{21}$ and $p(y_{11}|y_{2j}, \theta^{(1)}) = \theta_{j11}$.

We assume that, given M_0 , $\theta_2 = (\theta_{21}, \theta_{22}) \sim D(2, 2)$ and $\theta_1 = (\theta_{11}, \theta_{12}) \sim D(2, 2)$, and they are independent. Given M_1 , we assume that $\theta_{j1} = (\theta_{j11}, \theta_{j12}) \sim D(1, 1)$, and they are independent. Thus, a priori, the marginal probabilities of y_{2j} , y_{1k} and $y_{1k}|y_{2j}$ are all

Y ₂	Y ₁		Total
	1	2	
1	$n(y_{11} y_{21})$	$n(y_{12} y_{21})$	$n(y_{21})$
2	$n(y_{11} y_{22})$	$n(y_{12} y_{22})$	$n(y_{22})$
Total	$n(y_{11})$	$n(y_{12})$	n

Table 3: Contingency table.

uniform and are based on a global prior precision 4. Suppose we collect a random sample, and we report the summary statistics in the contingency table given in Table 3.

With complete data, the marginal likelihood under models M_0 and M_1 are found by applying Equation (18) and are:

$$\begin{aligned}
 p(\mathbf{y}|M_0) &= \prod_{j=1}^2 \frac{\Gamma(4)\Gamma(2+n(y_{2j}))}{\Gamma(4+n)\Gamma(2)} \prod_{k=1}^2 \frac{\Gamma(4)\Gamma(2+n(y_{1k}))}{\Gamma(\alpha+n)\Gamma(2)}, \\
 p(\mathbf{y}|M_1) &= \prod_{j=1}^2 \frac{\Gamma(4)\Gamma(2+n(y_{2j}))}{\Gamma(\alpha+n)\Gamma(2)} \prod_{k=1}^2 \frac{\Gamma(2)\Gamma(1+n(y_{1k}|y_{2j}))}{\Gamma(2+n(y_{2j}))\Gamma(1)}
 \end{aligned}$$

Thus, the model choice is based on the value of the ratio

$$r = \frac{p(M_0|I_0)p(\mathbf{y}|M_0)}{p(M_1|I_0)p(\mathbf{y}|M_1)},$$

from which the following decision rule is derived: if $r < 1$ model M_1 is chosen, if $r > 1$ model M_0 is chosen, and if $r = 1$ then the two models are equivalent.

Unfortunately, as the number of variables increases, the evaluation of all possible models becomes infeasible, and heuristic methods have to be used to limit the search process to a subset of models. The most common heuristic search limits its attention to the subset of models that are consistent with a partial ordering among the variables, so that $Y_i \prec Y_j$ if Y_i cannot be parent of Y_j . Furthermore, the fact that (18) is a product of terms that measure the evidence of each parent-child dependence can be exploited to develop search algorithms that work locally. This heuristic method was originally proposed by Cooper and Herskovitz [7]: they describe an algorithm which uses this heuristic to fully exploit the decomposability of (18). Denote the local contribution of a node Y_i and its parents Π_i to the overall joint probability $p(\mathbf{y}|M_j)$ by

$$g(Y_i, \Pi_i) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}. \quad (19)$$

For each node Y_i , the algorithm proceeds by adding a parent at a time and computing $g(Y_i, \Pi_i)$. The set Π_i is expanded to include the parent node that give the largest contri-

Y ₁	Y ₂	Y ₃	
		1	2
1	1	2	10
	2	5	3
2	1	6	3
	2	10	2

Table 4: Artificial sample.

bution to $g(Y_i, \Pi_i)$, and stops if the probability does not increase any longer. In the next example we use synthetic data to show the algorithm at work in a simple application.

Example 12 (Model Search) *Let $\mathcal{Y} = \{Y_1, Y_2, Y_3\}$ and suppose that $Y_3 \prec Y_2 \prec Y_1$ and they are all binary. Suppose further that all models consistent with this ordering have the same prior probability, and that the parameterization induced by each of these models is based on a global precision $\alpha = 6$ which is then distributed uniformly across the parameters. Data collected are reported in the contingency table 4.*

The algorithm starts by exploring the dependencies of Y_3 as child node, and results are

$$\begin{aligned} \Pi_3 = \emptyset \quad \log g(Y_3, \Pi_3) &= -29.09044 \\ \Pi_3 = Y_1 \quad \log g(Y_3, \Pi_3) &= -26.92774 \\ \Pi_3 = Y_2 \quad \log g(Y_3, \Pi_3) &= -27.56373 \end{aligned}$$

so that the node Y_1 is selected as parent of Y_3 . Next, both nodes Y_1, Y_2 are linked to Y_3 and $\log g(Y_3, (Y_1, Y_2)) = -26.46519$. Since this value is larger than -26.92774 , this model of local dependence is selected. Then, the node Y_2 is chosen as child node, and the dependence from Y_1 is explored:

$$\begin{aligned} \Pi_2 = \emptyset \quad \log g(Y_2, \Pi_2) &= -29.34088 \\ \Pi_2 = Y_1 \quad \log g(Y_2, \Pi_2) &= -29.79732 \end{aligned}$$

and since the model with Y_2 independent of Y_1 gives the largest contribution, the model selected is Y_1, Y_2 independent, both parents of Y_3 .

4.3 Advanced Topics

In the description of learning BBNS from data, we have made the assumption that the sample is complete, that is there are not cases reported as unknown. When the sample is incomplete, each missing datum can be replaced by any value of the associated variable. Therefore, an incomplete sample induces a set of possible databases given by the combination of all possible variable values for each missing datum. Exact Bayesian analysis requires computing

the posterior distribution of the parameters θ as a mixture of the posterior distributions of all these possible databases. Clearly, this approach becomes infeasible as the proportion of missing data increases and we must resort to approximate methods. Popular approaches use either asymptotic approximations based on the use of the MLEs or GMLEs, that can be computed using iterative methods as the EM-algorithm, or stochastic approximations as GS (see [15] for a review.) The Bound and Collapse method by [24] provides an efficient deterministic algorithm to approximate the posterior distribution of incomplete samples which can be further used for model selection [23].

In this chapter, we have limited our attention to BBNS with discrete variables. Methods for learning and reasoning with BBNS with continuous, or mixed variables can be found in [20] and a recent review is reported in [4].

References

- [1] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag: New York, 2nd edition, 1985. First Ed 1980.
- [2] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, New York, 1994.
- [3] G E P Box and G C Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, New York, 1973.
- [4] W. Buntine. Graphical models for discovering knowledge. In *Advances in Knowledge Discovery and Data Mining*, pages 59–81. MIT Press, Cambridge, MA, 1996.
- [5] W. L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994.
- [6] E Castillo, J M Gutierrez, and A S Hadi. *Expert Systems and Probabilistic Network Models*. Springer Verlag, New York, 1997.
- [7] G.F. Cooper and E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
- [8] M Evans and T Swartz. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 10:254–272, 1995.
- [9] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 1997.
- [10] D. Geiger and D. Heckerman. A characterization of Dirichlet distributions through local and global independence. *Annals of Statistics*, 25:1344–1368, 1997.

- [11] A. Gelman, J. B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.
- [12] W R Gilks, S Richardson, and D J Spiegelhalter. Introducing Markov chain Monte Carlo. In W R Gilks, S Richardson, and D J Spiegelhalter, editors, *Markov Chain Monte Carlo in practice*, pages 1–19. Chapman and Hall, London, 1996.
- [13] W.R. Gilks and G.O. Roberts. Strategies for improving MCMC. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in practice*, pages 89–114. Chapman and Hall, London, 1996.
- [14] I.J. Good. *The Estimation of Probability: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, MA, 1968.
- [15] D. Heckerman. Bayesian networks for knowledge discovery. In *Advances in Knowledge Discovery and Data Mining*, pages 153–180. MIT Press, Cambridge, MA, 1996.
- [16] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combinations of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.
- [17] E T Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.
- [18] E T Jaynes. Prior probabilities. *IEEE Transactions on Systems, Science and Cybernetics*, SSC-4:227–241, 1968.
- [19] H Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 3rd ed. edition, 1961.
- [20] S. L. Lauritzen. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–108, 1992.
- [21] A O’Hagan. *Bayesian Inference*. Kendall’s Advanced Theory of Statistics. Arnold, London, 1994.
- [22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann, San Mateo, CA, 1988.
- [23] M. Ramoni and P. Sebastiani. Learning Bayesian networks from incomplete databases. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA, 1997. Morgan Kaufmann.
- [24] M. Ramoni and P. Sebastiani. Parameter estimation in bayesian networks from incomplete databases. *Intelligent Data Analysis Journal*, 2(1), 1998.

- [25] D J Spiegelhalter, A Thomas, N Best, and W R Gilks. Bugs bayesian inference using gibbs sampling. available at <http://www.mrc-bsu.cam.ac.uk>.
- [26] D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:157–224, 1990.
- [27] A. Thomas, D.J. Spiegelhalter, and W.R. Gilks. Bugs: A program to perform Bayesian inference using Gibbs Sampling. In *Bayesian Statistics 4*, pages 837–42. Clarendon Press, Oxford, 1992.
- [28] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- [29] S. S. Wilks. *Mathematical Statistics*. Wiley, New York, 1963.