# A survey of Bayesian Data Mining - Part I: Discrete and semi-discrete Data Matrices

Stefan Arnborg *
Swedish Institute of Computer Science

SICS TR T99:08, ISSN 1100-3154, ISRN:SICS-T–99/08-SE

**Abstract**

This tutorial summarises the use of Bayesian analysis and Bayes factors for finding significant properties of discrete (categorical and ordinal) data. It overviews methods for finding dependencies and graphical models, latent variables, robust decision trees and association rules.

## 1 Introduction

Data mining is complementary to Bayesian data analysis. Whereas data mining is often seen as the problem of grinding through massive data sets for the purpose of finding unexpected dependencies in the form of correlations, association rules and segmentations, Bayesian data analysis is typically seen as an activity of evaluating detailed models for small data sets. We are interested in the middle ground, where data is scarce enough to pose delicate questions of validity and significance of our findings, but where we do not yet have detailed mathematical models. We are developing tools and methodology for exploratory analysis of small and fragile data sets, as a preparatory step for a more detailed analysis, as can be performed in the Bayesian framework with, e.g., the BUGS system [33].

The application area is human brain research. Here, many different types of data are recorded for patients and for healthy control persons. Besides results of established and well standardized tests and background data, many results from imaging investigations (measuring cell structure, blood flow, receptor presence, etc.) are entered as extracted features of images mapped to brain atlases. Genetic data related to brain development is also emerging. Some data entered are uncertain, others are being standardized. We seldom have a complete data set for any individual, since the data collection process is costly and often infeasible for patients in bad condition. The objective of data mining on these data are deeper understanding of the interplay between physiological and psychiatric conditions, and also improved procedures for diagnosing patients and choosing therapies.

The purpose of this report is to explain the advantage of the Bayesian approach in the present application, and how the Bayes factor can be an almost

*email: stefan@nada.kth.se; mail: NADA, KTH, SE-100 44 Stockholm, Sweden

universal tool for choosing between models, and to show how models can display the information or knowledge we are after in an application. It is also our intention to give a full account of the computations required. It can serve as a survey of the area, although it focuses on techniques being investigated in the present project. Several of the computations we describe have been analysed at length, although not exactly in the way and with the same conclusions as found here. The contribution here is a systematic treatment that is confined to pure Bayesian analysis and puts several established data mining methods in a joint Bayesian framework. We do not want to enter the discussion of why the Bayesian approach is superior to its alternatives, but some background material is included. We will see that, although many computations of Bayesian data-mining are straightforward, one soon reaches problems where difficult integrals have to be evaluated, and presently only Markov Chain Monte Carlo (MCMC) methods are available. There are several recent books describing the Bayesian method from both a theoretical[3], an ideological[19, 32] and an application oriented[7] perspective. A main historic influence leading to increased interest in Bayesian methods is Harold Jeffreys, who wrote particularly two books on scientific inference and probability theory from a Bayesian perspective[21, 20]. A current survey of MCMC methods, which can solve some complex evaluations required in Bayesian modeling, can be found in the book[17]. Books explaining theory and use of graphical models are Lauritzen[22], Cox and Wermuth[10], and Whittaker[35]. A tutorial on Bayesian network approaches to data mining is found in (Heckermann[18]). This present report describes data mining in a relational data structure with discrete data (discrete data matrix) and the simplest generalizations to numerical data. A second part will describe general real valued data matrices, raster data representing, e.g., scalar and/or vector fields, as well as time series and strings.

# 2 Data model

We consider a data matrix where rows are cases and columns are variables. In our application, the row is associated with a person or an investigation (patient and date). The columns describe a large number of variables that could be recorded, such as background data (occupation, sex, age, etc), and numbers extracted from investigations made, like sizes of brain regions, receptor densities and blood flow by region, etc. Categorical data can be equipped with a confidence (probability that the recorded datum is correct), and numerical data with an error bar. Every datum can be recorded as missing, and the reason for missing data can be related to patients condition or external factors (like equipment unavailability or time and cost constraints). Only the latter type of missing data is (at least approximately) unrelated to the domain of investigation. On the level of exploratory analysis we confine ourselves to discrete and multivariate normal distributions, with Dirichlet and inverse Wishart priors. In this way, no delicate and costly MCMC methods will be required until missing data and/or segmentation is introduced. If the data do not satisfy these conditions (e.g., normality for a real variable), they may do so after suitable transformation and/or segmentation. Another approach is to ignore the distribution over the real line and regard a numerical attribute as an ordinal one, i.e., considering only the ordering between values. Such ordinal data also appear

naturally in applications where subjects are asked to grade a quantity, like their appreciation of a phenomenon in organized society or their valuation of their own emotions.

## 2.1 Multivariate data models

Given a data matrix, the first question that arises concerns the relationships between its variables(columns). Could some pairs of variables be considered independent, or do the data indicate that there is a connection between them - either directly causal, mediated through another variable, or introduced through sampling bias? These questions are analyzed using graphical models, directed or decomposable[24]. As an example, in figure 1 $M_1$ indicates a model where $A$ and $B$ are dependent, whereas they are independent in model $M_2$. In figure 2, we describe a directed graphical model $M_4''$ indicating that variables $A$ and $B$ are independently determined, but the value of $C$ will be dependent on the values for $A$ and $B$. The similar decomposable model $M_4$ indicates that the dependence of $A$ and $B$ is completely explained by the mediation of variable $C$. We could think of the data generation process as determining $A$, then $C$ dependent on $A$ and last $B$ dependent on $C$, or equivalently, determining first $C$ and then $A$ dependent on $C$ and $B$ dependent on $C$.

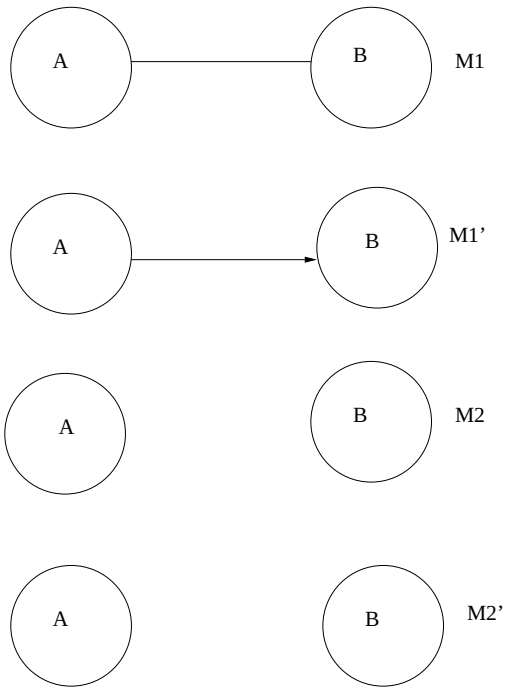


Figure 1: Graphical models, dependence or independence?

Bayesian analysis of graphical models involves selecting all or some graphs on the variables, dependent on prior information, and comparing their posterior probabilities with respect to the data matrix. A set of highest posterior probability models usually gives many clues to the data dependencies[23, 24], although

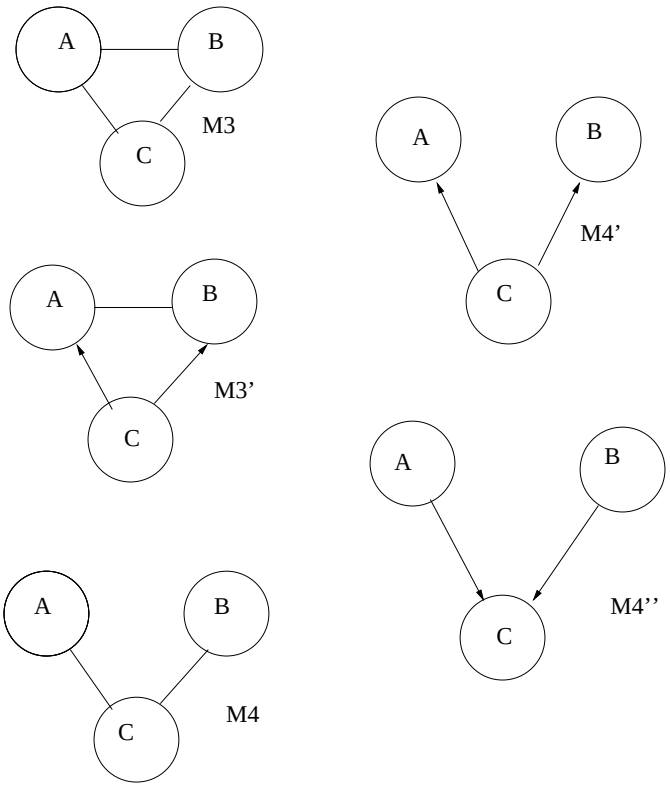


Figure 2: Graphical models

one must - as always in statistics - constantly remember that dependencies are not necessarily causalities.

A second question that arises concerns the relationships between rows (cases) in the data matrix. Are the cases built up from distinguishable classes, so that each class has its data generated from a simpler graphical model than that of the whole data set? In the simplest case these classes can be directly read off in the graphical model. In a data matrix where inter-variable dependencies are well explained by the model $M_4$, if $C$ is a categorical variable taking only few values, splitting the rows by the value of $C$ could give a set of data matrices in each of which $A$ and $B$ might be independent. However, the interesting cases are where the classes cannot be directly seen in a graphical model because then the classes are not trivially derivable. If the data matrix of the example contained only variables $A$ and $B$, because $C$ was unavailable or unknown to interfere with $A$ and $B$, the highest posterior probability graphical model might be one with a link from $A$ to $B$. The classes would still be there, but since $C$ would be latent or hidden, the classes would have to be derived from the $A$ and $B$ variables only. A different case of classification is where the values of one numerical variable are drawn from several normal distributions with different means and variances. The full column would fit very badly to any single normal distribution, but after classification, each class could have a set of values fitting well to a normal distribution. The problem of identifying classes is known as

unsupervised classification. One comprehensive system for classification based on Bayesian methodology is described by Cheeseman and Stutz[8].

A third question - often the one of highest practical concern - is whether some designated variable can be reliably predicted in the sense that it is well related to combinations of values of other variables, not only in the data matrix, but also with high confidence in new cases that are presented. This question leads to another concept that has been extensively studied, namely association rules. Consider a data matrix well described by model $M_4$ in figure 2. It is conceivable that the value of $C$ is a good predictor of variable $B$, and better than $A$. It also seems likely that knowing both $A$ and $C$ is of little help compared to knowing only $C$, because the influence of $A$ on $B$ is completely mediated by $C$. On the other hand, if we want to predict $C$, it is well conceivable that knowing both $A$ and $B$ is better than knowing only one of them.

Finally, it is possible that a data matrix with many categorical variables with many values gives a scattered matrix with very few cases compared to the number of potentially different cases. Generalization is a technique by which a coarsening of the data matrix can yield better insight, such as replacing the age and sex variables by the categories kids, young men, adults and seniors in a car insurance application. The question of relevant generalization is clearly related to the problems of finding association rules and to classification. For ordinal variables, this line of inquiry leads naturally to the concept of decision trees, that can be thought of as a recursive splitting of the data matrix by the size of one of its ordinal variables.

# 3 Bayesian analysis, uninformative priors, and over-fitting

A natural procedure for estimating dependencies among categorical variables is by means of conditional probabilities estimated as frequencies in the data matrix. Likewise, correlations can be used to find dependencies among real valued variables. Such procedures usually lead to selection of the more detailed models and give poor generalizing performance, in the sense that new sets of data are likely to have completely different dependencies. Various penalty terms have been tried to avoid over-fitting. However, the Bayesian method has a built-in mechanism that favors the simplest models compatible with the data, and also selects more detailed models as the amount of data increases. The procedure is to compare posterior model probabilities, where the posterior probability of a model is obtained by combining its prior distribution of parameters with the probability of the data as a function of the parameters, using Bayes rule. Thus, if $p_1(\Theta_1)$ is the prior pdf of the parameter (set) $\Theta_1$ of model $M_1$ and the probability of obtaining the case (row of data matrix) $d$ is $p(d|M_1\Theta_1)$, then the probability in model $M_1$ of the data matrix $D$ containing the ordered cases $\{d_i\}_{i\in I}$ is:

$$p(D|M_1) = \int \prod_{i\in I} p(d_i|M_1\Theta_1)p(\Theta_1)d\Theta_1, \tag{1}$$

and the posterior probability of model $M_1$ given the data $D$ is, by Bayes rule:

$$p(M_1|D) = p(D|M_1)p(M_1)/p(D). \quad (2)$$

From a frequentist or orthodox statistical point of view it is questionable to do this interchange and consider the probability of a model given the data. This is exactly what makes the difference between Bayesian and frequentist methods. If the data matrix is unordered, one should multiply with a multinomial coefficient, but this is often not done - whether or not this is done does not matter for computation of Bayes factors, see below. Two models $M_1$ and $M_2$ can now be related with respect to the data by the Bayes factor $p(D|M_1)/p(D|M_2)$. This is a factor which is multiplied with the prior odds between the two models, $p(M_1)/p(M_2)$, to get the posterior odds $p(M_1|D)/p(M_2|D)$. The posterior odds can now take the place of a new prior for the next data batch, and the procedure can be repeated. It should be noted, however, that the model averaging is done for each batch - whether this is appropriate or not depends on the application, and often it is not.

A high value of the Bayes factor, say more than 100, speaks strongly in favor of model $M_1$, like a value below .01 gives strong support for $M_2$. Values closer to one (i.e. in the range .3 to 3 ), however, tell us that the data are insufficient to decide between the models, and this is unavoidable - methods that decide in those cases cannot be well designed. This appears to be a significant difference between the Bayesian approach and many analyses occurring in AI and data mining - we do not consider our data as an imperfect image of an ideal underlying and completely precise probability model. On the contrary, we ask which imperfect underlying models best serve to describe our data. If we tried to get much more data than we have, we would not necessarily become wiser, since the data collection process may well be such that cases are not independent and the data collection process may change the nature of the data through the sampling process.

A disturbing feature of the Bayesian methodology is that it requires prior distributions. Priors give an impression of subjectivity, which they should not do. The prior is an assessment of a state of information, and is not related to a subject except that the information state is possessed by a subject. Often the information state is difficult to deal with since its form is fairly open-ended - just imagine information related to an open mathematical problem, or even an $NP$-hard optimization problem. However, every well-founded choice between alternatives must involve the prior beliefs of - objectively the state of information held by - the decision maker in some way, and the Bayesian method is one (in fact the only) consistent way of doing this. Bayesian methodology provides an expedient for the case where no strong prior beliefs should influence the conclusion, namely uninformative or weakly informative priors. For such prior distributions, more data is typically needed to reach a definite conclusion than for cases where there is distinct prior information to include in the analysis. With the Bayesian method there is no need to penalize more detailed models to avoid over-fitting - if $M_2$ is more detailed than $M_1$ in the sense of having more parameters to fit, then the parameter dimension is larger in $M_2$ and $p(\Theta_1)$ is larger than $p(\Theta_2)$, which automatically penalizes $M_2$ against $M_1$. This automatic penalization has been found appropriate in many application cases, and should be complemented by explicit prior model probabilities only when there is concrete prior information that justifies it, or when the data is too abundant to

select a model simple enough to comprehend. Asymptotically, the penalization of detailed models implicit in the Bayes factor approach is a factor $n2^{(p_1-p_2)}$, where $n$ is the number of data points (cases) and $p_i$ is the number of parameters in model $M_i$. This estimate was first found by Schwarz[31], and is known, when used to penalize more detailed models in a likelihood based model comparison, as the Bayesian information criterion (BIC). So deciding between the models using the likelihood ratios with the BIC as a penalizing factor is an approximation to the 'orthodox Bayesian' procedure of comparing posterior probabilities, and it is useful when the integration required for posterior determination is infeasible or otherwise unwanted. Some discussions of this point can be found in (Ch 24 of Jaynes[19]) and also in Neal[25].

The discussion above relates to choosing one of two models. Clearly, there is a possibility that the data discredits both these models, or that we have a whole family of models to choose from.

Consider the problem of comparing models in a family, $\{M_1, \ldots, M_k\}$, and having no prior preference for any of them. If the models do not overlap, we should choose the probabilities $\{p(M_i|D)/\sum_j p(M_j|D)\}$ as the probabilities of these models given the data. By overlapping we mean that parameter sets of prior non-zero probability exist which give the same distribution in two models. We usually do not have overlap, since, e.g., in the case of nested models the region of overlap, the whole 'less specific' model, has prior probability zero in the more specific model. Typically, a nested family forming a tree or directed acyclic graph structure is chosen, where the dimension of the parameter space increases as one descends in the tree, and where the root is associated with the fewest parameters. The root model is the least specific one in the family.

In the modeling effort, the analyst must decide on grounds of what is known in general terms about the application and the purpose of the analysis, which model family to consider. Here we must remember that inference is not an idle activity, but should normally be used to make decisions. Clearly, it is not adequate to select a model from its posterior probability without considering the consequences of decisions. In Bayesian decision theory (see, e.g., (Berger[1]), we introduce actions and expected utility of actions given a 'state of the world', which could be a model or a model with its parameter. However, in Bayesian decision theory, the rational decision making follows from only the posterior and the utility functions (statisticians seem to be a pessimistic breed and usually talk about loss functions, but this is of course really the same thing). For this reason we do not introduce loss functions in this report.

## 3.1 The Bayesian debate and the unavoidability of Bayesian analysis

There was a quite heated debate among statisticians on the proper application of mathematical tools in the interpretation of experimental data. This debate started between Fisher and Pearsson and continued between Fisher and Jeffreys. What is most remembered is the discussion between Bayesians and 'frequentists' (as traditional statisticians were called by Bayesians). For a trained pure mathematician the controversy between frequentist and Bayesian views does simply not appear. He is interested in abstract spaces with probability measures, and leaves interpretation of real world phenomena to others.

The strong feelings among Bayesians on their infallibility has created adverse reactions among statisticians and also recently in the AI community. Bayesians are known for their arrogance and claim to own the truth. It is unfortunate that this claim is not presented in many textbooks, because it is easy to understand, and also quite surprising. It is generally agreed that Bayes original paper is deep and challenging, but it is also too vague and incoherent to be convincing, and many readers have rejected it outright. There is apparently no documented evidence that Laplace actually saw the paper or heard of it, but the work of Laplace is a continuation of the ideas in Bayes work. Unfortunately, he did not succeed to convince his colleagues and successors in the scientific community. His idea of the rule of succession is a clear application of Bayesian analysis, but it was rejected because his readers did not accept his choice of prior information (deciding the number of days, all with sunrise, since creation, by reading the Bible) and discarded the method on the basis of one dubious application. Obviously, if the Bible is reliable on this point, other information on the order of Nature found in it might contradict his application. Other sources of prior information were known by Laplace, but he did not use them for this purpose. Several great 19th century mathematicians have more or less by instinct used the ideas of Bayes and Laplace when performing computations on experimental data (typically in astronomy), but these efforts were more or less ignored when the discipline of statistics was created in the early 20th century.

The first derivation of the necessity of Bayesian methods was done by R. T. Cox in 1946[11], and has been repackaged by Jaynes with a lot of motivating discussion. Basically, the analysis investigates which family of rules for reasoning with the plausibility of statements about the world is permissible in the sense that they satisfy the following criteria:

- I: The plausibility of a statement is a real number and dependent on information we have on the plausibility of other statements.
- II: Consistency - If the plausibility of a statement can be derived in two ways, the two results must be equal.
- III: Common sense - Some properties of statements known to be true or known to be false, and continuity rules.

From these criteria follows that any permissible way to reason with plausibility is equivalent to Bayesian analysis.

A very short outline follows, were we do not in fact show that the Bayesian method satisfies the criteria (this is not usually questioned):

Let $A$, $B$, $C$, ... be statements, combinable with the invisible logical**and** operator: $AB$ means $A$ and $B$. The negation of a statement $A$ is written $\overline{A}$. Statements must in some way be considered objective and relate to states of the world, and have an agreed interpretation. Let $A|C$ be the plausibility of $A$ given the additional information that $C$ is true. $C$ is thus the *context* in which we consider the plausibility of $A$. That such a notation must be present in every calculus to derive plausibility is clear - there must for example be a way to relate a measured value$(A|C)$ to the reality behind it $(B|C)$ using background information on the measurement process and its accuracy $(C)$. Numerical values - parameters, measured values, etc. - enter this framework by a limit process. We cannot start with infinite domains and directly put plausibility measures on them.

Consider the possible ways to compute $AB|C$: it must be a function of two or more of the plausibilities $A|C$, $B|AC$, $B|C$ and $A|BC$. It can be shown that we must consider either $B|AC$ and $A|C$ or $A|BC$ and $B|C$ - any other alternative can be shown inadequate by violating common sense in some situation. As an example we cannot derive the plausibility of $AB|C$ from only the plausibilities of $A|C$ and $B|C$, since that gives us no means to consider how $A$ and $B$ relate to each other- it would force us to assume, for example, that the plausibility of a person having a left blue and a right brown eye would depend only on the plausibilities of left blue and right brown eye, and not allowing us to consider the dependency between these two statements.

Thus, we can assume that the plausibility of $AB|C$ is a function of the plausibilities of $A|C$ and $B|AC$, the other case being a natural consequence of the commutativity of the and operator:

$$AB|C = F(A|C, B|AC). \tag{3}$$

The common sense requirement tells us that the function $F$ must be continuous, and monotonously increasing in both its arguments. It can have a stationary point for its first argument only if the second argument represents impossibility and vice versa. We assume it twice continiously differentiable, although there exists a fairly complex proof that this is not necessary for our conclusions[19].

Now we consider the consistency requirement. Since the and operator is not only commutative but also associative, $ABC = (AB)C = A(BC)$, we can derive a consistency requirement for $F$:

$$ABC|D = F(AB|CD, C|D) = F(A|BCD, BC|D). \tag{4}$$

Expanding once more, we get:

$$F(F(A|BCD, B|CD), C|D) = F(A|BCD, F(B|CD, C|D)). \tag{5}$$

This must hold for any statements $A, B, C, D$, and thus $F$ must satisfy the following functional equation in its range of definition:

$$F(x, F(y, z)) = F(F(x, y), z). \tag{6}$$

The above is called the equation of associativity. The trivial constant solution is clearly useless. Which non-trivial solutions are there? We can differentiate equation (6) with respect to $x$, $y$ and $z$, and see that the following equality holds, i.e., the left side is independent of $z$ (we use the notation $F_1(x, y) = \frac{\partial F(x,y)}{\partial x}$):

$$F_2(x, F(y, z))F_1(y, z)/F_1(x, F(y, z)) = F_2(x, y)/F_1(x, y). \tag{7}$$

Let $G(x, y) = F_2(x, y)/F_1(x, y)$ and we find $F_1(y, z)G(x, F(y, z)) = G(x, y)$, and the left side of this (which is algebraically independent of $z$) we denote $U$. Likewise, after a little algebra: $G(x, f(y, z))F_2(y, z) = G(x, y)G(y, z)$, and the left side we denote by $V$. Now $\partial V/\partial y$ is identical to $\partial U/\partial z$ and thus zero, since $U$ is independent of $z$. But then $V$ which can be written $G(x, y)G(y, z)$, is independent of $y$. This can only happen if $G(y, z)$ and $1/G(x, y)$ have a common factor dependent on $y$, and no other dependence on $y$. So we must have $G(y, z) = H(y)E(z)$ and $G(x, y) = E'(x)/H(y)$ so we also have, by substituting

$y$ for $x$ and $z$ for $y$ in the latter: $G(y,z) = E'(y)/H(z)$. In other words, $G$ must have the form $G(x,y) = rH(x)/H(y)$, and this is also by definition equal to $F_2(x,y)/F_1(x,y)$. This is what we need to separate variables and put the diferential of $v = F(x,y)$ on an integrable form:

$$\frac{dv}{H(v)} = \frac{dx}{H(x)} + r\frac{dy}{H(y)} \tag{8}$$

which can be integrated using $w(x) \equiv \exp(\int^{\infty} \frac{dx}{H(x)})$ to:

$$w(F(x,y)) = w(x)w^r(y) \tag{9}$$

but the equation of associativy also gives us

$$w(F(x,y),z) = w(x)w^r(y)w^r(z) = w(F(x,F(y,z))) = w(x)(w(y)w(z))^r \tag{10}$$

and in every non-trivial and useful case we must have $r = 1$. We can now investigate what $w(x)$ must be when $x$ represents truth or falsity, and we get $w(x) = w(x)w(T)$, $w(F) = w(x)w(F)$ and some more conditions we do not have to use. It is possible that the values $\infty$ and $-\infty$ are obtained since truth and falsity might be considered a limit case. The first condition yields $w(T) = 1$, the other could mean either $w(F) = 0$ or $w(F) = \infty$ ($-\infty$ is ruled out since we cannot allow $w(x)$ to pass zero in its way from $w(T)$ to $w(F)$). But the solution going from 1 to $\infty$ can be replaced by its inverse, which goes from 1 to 0. We are now very close to probability rules, since the function $w$ goes from 0 for impossibility to 1 for truth, and our rule for the conjunction of statements can be written

$$w(AB|C) = w(A|BC)w(B|C) \tag{11}$$

It now remains to find out how plausibilities of complements must be treated. Since $A\overline{A}$ is always false and either of $A$ or $\overline{A}$ must be true, the plausibility of $\overline{A}$ must be a function of the plausibility of $A$. Introduce the function $S$ on the unit interval: $S : [0,1] \to [0,1]$, such that $w(\overline{A}|C) = S(w(A|C))$. By considering Aristotelian logic, and our choice of $w(T) = 1$ and $w(F) = 0$, and reasonable common sense, we find that $S$ is a monotone and continuous function decreasing from 1 to 0 on the unit interval. We will assume that $S$ is differentiable - again this is not necessary but it is almost required by common sense and simplifies the argument. Also, since $\overline{\overline{A}} = A$, we have $S(S(x)) = x$. This is not all, however, because $S$ must also be consistent with the product rule:

$$w(AB|C) = w(A|C)w(B|AC) = w(A|C)S(w(\overline{B}|AC)) \tag{12}$$

$$w(A\overline{B}|C) = w(A|C)w(\overline{B}|AC) = w(A|C)S(w(B|AC)) \tag{13}$$

Rearranging these constraints and using the commutativity $AB = BA$ we find $w(AB|C) = w(A|C)S(w(\overline{B}|AC)) = w(A|C)S(w(A\overline{B}|C)/w(A|C))$, and

$$w(A|C)S(\frac{w(A\overline{B}|C)}{w(A|C)}) = w(B|C)S(\frac{w(B\overline{A}|C)}{w(B|C)}). \tag{14}$$

Equation (14) must hold for all statements $A$, $B$, and $C$. In particular, choose $B$ such that $\overline{B} = AD$, and now $A\overline{B} = \overline{B}$ and $B\overline{A} = \overline{A}$. Introducing

the abbreviations $x \equiv w(A|C)$ and $y \equiv w(B|C)$. After a little work we find the fundamental equation governing the possible functions $S$:

$$xS(\frac{S(y)}{x}) = yS(\frac{S(x)}{y}), S(y) \leq x \tag{15}$$

The analysis of this equation is not entirely trivial, but it can readily be verified that among its solutions are the (easily obtainable) solutions to the simple equation:

$$S(x)^m + x^m = 1, m > 0. \tag{16}$$

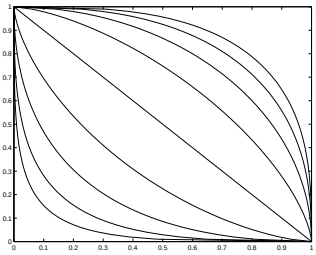

Figure 3: Sample solutions to (16).

For the different values of $m$, the curve family will cover the interior of the unit square (see figure 3). It is also easy, by consideriung the choice $y = S(x) + \epsilon$ as $\epsilon \to 0$, to see that $S$ is governed by a first order differential equation. Therefore, there are no more solutions than these. It might seem odd that the solution $S(x) = 1 - x$ is not the only one, since it would fit well with equation (11) and the choice of $w(A|C)$ as the probability of $A|C$. However, by taking the $m$th power of equation (11), we find that we can still interpret all possible ways to compute with plausibilities as Bayesian analysis, simply by letting probability correspond to $w(A|C)^m$.

The next question in this line of inquiry concerns proper choices of priors - we have no help whatsoever in the preceding discussion. Building a repertoire of methods to assign priors would start with simple symmetry considerations: If I have no background knowledge whatsoever to find differences in plausibility between a set of $n$ exclusive and exhaustive hypotheses, then the prior probability of each should be set to the same value and the probabilities should sum to one. Thus, each hypothesis will have prior probability $1/n$. This leads to the standard assignments for coin tossing and urn drawing experiments considered in basic probability texts. Translating this rule, by limit forming operations, to pdfs with continuous parameter spaces leads naturally to the concept of minimum-information (maximum entropy) priors, which have revolutionized the methods for analyzing physics data and is spreading to other sciences. We do not describe this revolution here, see e.g. Jaynes[19]. A remaining problem is that we simply cannot consider all possible hypotheses. This means that the set of hypotheses we actually consider must in some sense be realistic. This is a problem that is the key problem that must get a convincing solution in every single application before meaningful application related conclusions can

be drawn. The models considered in this report are very general and have been found applicable to many different problems as a first quantitative grinding of the collected data. However, once the big lines have been uncovered, there is usually plenty of scope for investigating more specific and application related models.

## 3.2 An educational example: Tossing a coin

When it comes to the interpretation of experimental outcomes, we can illustrate the controversy with an example that has been discussed frequently by statisticians, first by Lindley (see, e.g.,[7, 32, 19]): Assume we toss a coin 12 times and observe the outcome `ttthhttttttth`, where `t` means tail and `h` means head. We are interested in what this means for our objective of learning whether or not the coin is fair. The probability of this sequence for a fair coin is $0.5^{12}$ as it is for any other sequence of 12 tosses. So it does not seem extraordinary, nor does a sequence consisting of millions of heads only, because it also has the same probability as any other sequence of the same length. The frequentists approach is to define a test. We order the possible outcomes linearly or map them to the real line, and this induces a pdf of a real-valued quantity. If the current outcome lies far out on the tail of this distribution, we reject the hypothesis that the coin is fair. It is accepted that a 5% cutoff can be used, and this gives us a 5% risk of rejecting a true hypothesis. Of course the map of outcomes to the real line must be defined in some impartial way, essentially before we have seen the actual outcome. Typically, at least if we are more concerned with fairness than with independence, we choose the number of tails in the sequence, which has a binomial distribution. The probability of 9 or more tails in 12 tosses of a fair coin is slightly more than 5% ($\sum_{i=9}^{12} \binom{12}{j} 2^{-12} = .075$), so we could reasonably assume that the coin is fair. There is a very fundamental problem with this approach, however, and that is that we made an assumption about the possible unobserved outcomes that is not justified. We just assumed that the outcome is one of the possible outcomes when tossing 12 times. The actual sequence observed does not exclude the possibility that the experimenter tossed the coin until he had 3 heads. If that were the case we should instead compute the distribution of the number of tails seen before the third head. This distribution is different, particularly it admits arbitrarily large values. A rapid calculation shows that with this rule we should reject the null hypothesis at the 5% level for the same outcome of the experiment (the probability of 9 or more tails is $\sum_{i=9}^{\infty} \binom{j+2}{j} 2^{-(j+3)} = .0325$). This dependence on the unknown experimental design violates a fundamental statistical principle saying that only the likelihood of the observed data can influence our belief in a hypothesis. This principle, the Likelihood Principle, was proposed by Fisher and Barnard, but it was first given a detailed analysis by Birnbaum in 1962[5]. In the subsequent debate, frequentists have proposed that the Likelihood Principle is not applicable in this case and that the experimental design could in practise be relevant information. A Bayesian only admits that the probability, under the fairness assumption, of the outcome observed is $0.5^{12} = .000244$ and that the probability of 9 tails is a factor $\binom{12}{9}$ larger. In order to evaluate the experiment he needs prior beliefs. Such prior belief could be an alternative model, defined before the experiment is observed. If the alternative model is that the coin gives tails with probability 2/3, then the probability is .000963 and the probability of 3 heads

under the alternative model is again a factor $\binom{12}{9}$ larger. So a Bayes factor of 3.9 in favor of the alternative hypothesis is observed, and the Bayesian starting out with no preference (probability 1/2 for each alternative) would end up with a preference for the alternative which could be quantified as probability .8 for the unfair alternative and .2 for the fair alternative. This preference should not be regarded as a rejection of the less believed alternative, but can easily be reversed by more information. There is a tempting alternative hypothesis in this case, namely that the true probability is the observed frequency, .75 for tails. This model has the highest probability (.001173) of those alternatives assuming independent outcomes. Even higher (probability 1) we reach if we assume that the observed sequence is the only possible outcome and that the tosses were not independent - but now we have definitely used the data too much, since we would probably not designate this hypothesis as a major alternative before the experiment.

Now, let the alternative hypothesis be: The probability of tails is a number $\Theta$. Figure 4 shows the probability of the outcome as a function of $\Theta$. We do not know anything about $\Theta$, but we must assume some distribution of it. One obvious alternative is the uniform distribution. This gives the model probability $\int_0^1 \Theta^9(1-\Theta)^3 d\Theta = .00035$. The resulting Bayes factor is 1.4 in favor of the hypothesis of unfairness, much weaker than 4.9 for the maximum likelihood hypothesis ($\Theta = 0.75$). A Bayesian with no prior preference of the hypotheses fair against unfair would end up by assigning probability 0.411 to the fair and .589 to the unfair hypothesis.

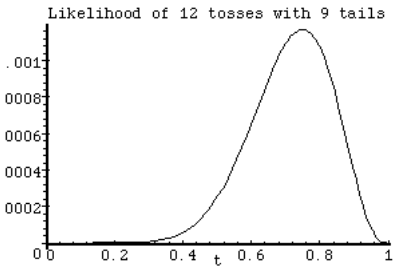


Figure 4: Posterior frequency distribution

It might seem unreasonable to let the probabilities less than 1/2 water out our belief in unfairness when data clearly suggest that the probability, if it is not 1/2, is greater. Let us split the unfairness case into two and consider three models: $M_l$ - bias for heads; $M_f$ - fair; $M_h$ - bias for tails. We again assume a uniform distribution of $\Theta$ in the intervals 0 to 1/2 for $M_l$ and in 1/2 to 1 for $M_h$. A similar calculation leads to the posterior probabilities 0.034, 0.259 and 0.706, respectively. Clearly, by separating the unfairness hypothesis into low and high bias, we decreased our belief in the fairness alternative. Unfortunately, this is to some extent an illusion. The real reason why our posterior belief in the fairness decreased is that our prior belief in fairness decreased when we replaced two equally believable hypothesis (prior probability 1/2 each) by three equally believable hypotheses (prior probability 1/3 each). It would be equally possible to split the 1/2 belief in unfairness into two unfairness alternatives with

1/4 prior probability each, and then the posterior belief in fairness would not change. This is one example of non-robustness problems appearing when doing Bayesian analyses with weak priors.

In any case, this is a result that seems much weaker than the frequentists ability to reject the fairness assumption given the information that the experimenter tossed the coin until 3 heads were observed. In 'fairness' it should be noted that the two views would yield similar results if 120 tosses were made with 30 observed heads: The Bayes factor would be $10^6$ in favor of unfairness and the level of the frequency test would be $10^{-7}$ – two equally convincing reasons to reject the fairness assumption.

The process of dividing the unfairness case into two can be continued, and in the limit we obtain the concept of a posterior distribution for $\Theta$ over the unit interval. This analysis is carried out, with a number of nice graphical results, by Sivia[32]. The resulting posterior with a uniform prior, $t$ tails and $h$ heads is the normalized likelihood function, the Beta distribution, $p(\Theta|h,t) = c\Theta^t(1-\Theta)^h$. In the next section we will perform a generalized derivation, where we allow more than 2 outcomes: we go from a Bernoulli distribution to a general discrete distribution, and we use the more general Dirichlet conjugate family instead of Beta distributions.

There is no mathematical reason to reject one of the frequentist or Bayesian approaches. Bayesians accused frequentists for not accepting probability as dependent on information, whereas frequentists accused Bayesians for putting up with the non-robustness caused by dependence on prior information. Admittedly, it is difficult to translate prior information to prior probability, but Bayesians claim that it is unavoidable. Whether the frequentists reliance on experimental design is worse than the Bayesians reliance on priors is of course impossible to say without a lot of experience. Several other arguments have been put forward in this debate, but those above seem to be the most critical. Today, Bayesian views are gaining ground, perhaps largely due to interest from the AI camp, where several less convincing ways to deal with imprecise information have been tried. Although we promote the pure Bayesian view in this report, it must be remembered that anyone investigating real data must explore it from many angles, in order to avoid being mislead by too constrained or inappropriate models. In practice such explorations are perhaps best performed with various visualization tools. An old saying is that a proper visaulization hits the investigator between his eyes with the truth. There is some truth in this.

# 4 Graphical model choice - local analysis

We will analyze a number of models involving two or three variables of categorical type, as a preparation to the task of determining likely decomposable or directed graphical models. First, consider the case of two variables, $A$ and $B$, and our task is to determine whether or not these variables are dependent. Since we know that Bayes method is the only method that gives us the right answer, we already know how to proceed. We must define one model $M_2$ that captures the concept of independence, and one model $M_1$ that captures the concept of dependence, and ask which one produced our data. The Bayes factor is $P(D|M_2)/P(D|M_1)$ in favor of dependence, and it will be multiplied with

the prior odds (which we, lacking prior information in this general setting, assume is one) to get the posterior odds. There is some latitude in defining the data model for dependence and independence, but they lead us to quite similar computations, as we shall see.

Let $d_A$ and $d_B$ be the number of possible values for $A$ and $B$, respectively. It is natural to regard categorical data as produced by a discrete probability distribution, and then it is convenient to assume Dirichlet distributions for the parameters (probabilities of the possible outcomes) of the distribution.

We will find that this analysis is the key step in determining a full graphical model for the data matrix. Our analysis is analogous to those of Dawid and Lauritzen[12] and Madigan and Raftery[24], but their analyses are in many ways more general and use a likelihood approach with penalization of detailed models using the BIC criterion and other similar techniques.

For a discrete distribution over $d$ values, the parameter set is a sequence of probabilities $\overline{x} = (x_1, \ldots x_d)$, constrained by $0 \leq x_i$ and $\sum_i x_i = 1$ (often the last parameter $x_d$ is omitted - it is determined by the first $d-1$ ones). A prior distribution over $\overline{x}$ is the conjugate Dirichlet distribution with a parameter set $\overline{\alpha} = (\alpha_i)_{i=1}^d$, constrained by $0 \leq \alpha_i$. Then the Dirichlet distribution with parameter set $\overline{\alpha}$ is $\mathrm{Di}(\overline{x}|\overline{\alpha}) = \prod_i x_i^{(\alpha_i - 1)} \Gamma(\sum_i \alpha_i) / \prod_i \Gamma(\alpha_i)$, where $\Gamma(n+1) = n!$ for natural number $n$. The normalizing constant $\Gamma(\sum_i \alpha_i) / \prod_i \Gamma(\alpha_i)$ gives a useful mnemonic for integrating $\prod_i x_i^{(\alpha_i - 1)}$ over the $d-1$-dimensional unit cube (with $x_d = 1 - \sum x_i$). It is very convenient to use Dirichlet priors, for the posterior is also a Dirichlet distribution: After having obtained data with frequency count $\overline{n}$ we just add it to the prior parameter vector $\overline{\alpha}$ to get the posterior parameter vector $\overline{\alpha} + \overline{n}$. It is also easy to handle priors that are mixtures of Dirichlets, because the mixing propagates through and we only need to mix the posteriors of the components to get the posterior of the mixture. We do not need this here, however.

With no specific prior information for $\overline{x}$, it is necessary from symmetry considerations to assume all Dirichlet parameters equal, $\alpha_i = \alpha$. A convenient prior is the uniform prior ( $\alpha = 1$ ). This is, e.g, the prior used by Laplace to derive the rule of succession, see Ch 18 of [19]. Other priors have been used, e.g., $\alpha = 1/2$ in the case $d = 2$, which is a minimum information (Jeffreys) prior. The value $\alpha = 1/2$ has also been used for $d > 2$(Madigan and Raftery[24]). Cheeseman and Stutz[8] report the use of $\alpha = 1 + 1/d$. Experiments have shown little difference between these choices, but it is easy to see that Jeffreys prior promotes $x_i$ close to 0 or 1 somewhat whereas $\alpha = 1 + 1/d$ penalizes extreme probabilities. If we get significant differences between different uninformative priors this warrants a closer investigation on the adequacy of data and modeling assumptions. We will mostly use the uniform prior. In many cases an experts delibered prior information can be expressed as an equivalent sample that is just added to the data matrix, and then this modified matrix can be analyzed with the uniform prior. Likewise, a number of experts can be mixed to form a mixture prior. If the data has occurrence vector $(n_i)_{i=1}^d$ for the $d$ possible data values in a case, and $n = n_\cdot = \sum_i n_i$, then the probability for these data given the discrete distribution parameters $\overline{x}$, is

$$p(\overline{n}|\overline{x}) = \binom{n}{n_1, \ldots, n_d} \prod_i x_i^{n_i}. \tag{17}$$

You should note that many derivations found in the literature drop the multinomial coefficient. This would give the probability not of getting a particular contingency table (data matrix), but a given ordered sample with the frequency counts $n_i$. The difference between these two views disappears when the multinomial coefficients cancel in the division leading to Bayes factors. Integrating out the $x_i$ with the prior gives the probability of the data given model $M$ ($M$ is characterized by a parameterized probability distribution and a prior on its parameters):

$$\begin{aligned} p_J(\overline{n}|M) &= \int p(\overline{n}|\overline{x})p(\overline{x})d\overline{x} \\ &= \int \binom{n}{n_1,\ldots,n_d} \prod_i x_i^{n_i} \prod_i x_i^{\alpha_i - 1} \frac{\Gamma(\alpha_\cdot)}{\prod_i \Gamma(\alpha_i)} d\overline{x} \\ &= \binom{n}{n_1,\ldots,n_d} \frac{\Gamma(d\alpha)}{\Gamma(\alpha)^d} \frac{\prod_i \Gamma(n_i+\alpha)}{\Gamma(n+d\alpha)} \qquad (18) \\ &= \frac{\Gamma(n+1)\Gamma(d\alpha)\prod_i \Gamma(n_i+\alpha)}{\Gamma(\alpha)^d \Gamma(n+d\alpha)\prod_i \Gamma(n_i+1)}. \qquad (19) \end{aligned}$$

As is easily seen, the uniform prior gives a probability for each sample size that is independent of the actual data:

$$p_u(\overline{n}|M) = \frac{\Gamma(n+1)\Gamma(d)}{\Gamma(n+d)}. \qquad (20)$$

Consider now the data matrix over $A$ and $B$. Let $n_{ij}$ be the number of rows with value $i$ for $A$ and value $j$ for $B$. Let $n_{\cdot j}$ and $n_{i\cdot}$ be the marginal counts where we have summed over the 'dotted' index, and $n = n_{\cdot\cdot} = \sum_{ij} n_{ij}$. Let model $M_1$ (figure 1) be the model where the $A$ and $B$ value for a row is combined to a categorical variable ranging over $d_A d_B$ different values, with a Jeffreys or uniform prior. The probability of the data given $M_1$ is obtained by replacing the products and replacing $d$ by $d_A d_B$ in equations (19) and (20):

$$p_J(\overline{n}|M_1) = \frac{\Gamma(n+1)\Gamma(d_A d_B \alpha_{AB})\prod_{ij}\Gamma(n_{ij}+\alpha_{AB})}{\Gamma(\alpha_{AB})^{d_A d_B}\Gamma(n+d\alpha_{AB})\prod_{ij}\Gamma(n_{ij}+1)}, \qquad (21)$$

$$p_u(\overline{n}|M_1) = \frac{\Gamma(n+1)\Gamma(d_A d_B)}{\Gamma(n+d_A d_B)}. \qquad (22)$$

We could also consider a different model $M_1'$, where the $A$ column is generated first and then the $B$ column is generated for each value of $A$ in turn. With uniform priors we get:

$$p_u(\overline{n}|M_1') = \frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)^{d_A}}{\Gamma(n+d_A)} \prod_i \frac{\Gamma(n_{i\cdot}+1)}{\Gamma(n_{i\cdot}+d_B)} \qquad (23)$$

Observe that we are not allowed to decide between the undirected $M_1$ and the directed model $M_1'$ based on equations (22) and (23). This is because these models define the same set of pdf:s involving $A$ and $B$, the difference lying only in the structure of parameter space and parameter priors. They overlap on a

set of prior probability one. Nonetheless, this computation is sometimes done, and it might be useful for seeing how well data fit the two parameterizations and parameter priors. A difference compared to real Bayes factors is that we cannot resolve the hypothesis by taking more data. The factor just measures relative stretch in the parametrization in the high likelihood areas.

In the next model $M_2$ we assume that the $A$ and $B$ columns are independent, each having its own discrete distribution. There are two different ways to specify prior information in this case. We can either consider the two columns separately, each being assumed to be generated by a discrete distribution with its own prior. Or we could follow the style of $M_1'$ above, with the difference that each $A$ value has the same distribution of $B$-values. Now the first approach: Assuming parameters $\overline{x}^A$ and $\overline{x}^B$ for the two distributions, a row with values $i$ for $A$ and $j$ for $B$ will have probability $x_i^A x_j^B$. For discrete distribution parameters $\overline{x}^A, \overline{x}^B$, the probability of the data matrix $\overline{n}$ will be:

$$\begin{aligned} p(\overline{n}|\overline{x}^A, \overline{x}^B) = \\ & \binom{n}{n_{11}, \ldots, n_{d_A d_B}} \prod_{i,j=1}^{d_A, d_B} (x_i^A x_j^B)^{n_{ij}} \\ & = \binom{n}{n_{11}, \ldots, n_{d_A d_B}} \prod_{i=1}^{d_A} (x_i^A)^{n_{i.}} \prod_{j=1}^{d_B} (x_j^B)^{n_{.j}}. \end{aligned}$$

Integration over the priors for $A$ and $B$ gives the data probability given model $M_2$:

$$\begin{aligned} p_J(\overline{n}|M_2) = \\ & \int p(\overline{n}|\overline{x}^A x^B) p(\overline{x}^A) p(\overline{x}^B) d\overline{x}^A d\overline{x}^B \\ & = \int \binom{n}{n_{11}, \ldots, n_{d_A d_B}} \prod_{i=1}^{d_A} (x_i^A)^{n_{i.}} \prod_{j=1}^{d_B} (x_j^B)^{n_{.j}} \times \\ & \prod_i (x_i^A)^{\alpha_A - 1} \frac{\Gamma(d_A \alpha_A)}{\Gamma(\alpha_A)^{d_A}} \prod_i (x_i^B)^{\alpha_B - 1} \frac{\Gamma(d_B \alpha_B)}{\Gamma(\alpha_B)^{d_B}} d\overline{x}^A d\overline{x}^B \\ & = \frac{\Gamma(n+1)}{\prod_{ij} \Gamma(n_{ij}+1)} \frac{\Gamma(d_A \alpha_A)}{\Gamma(\alpha_A)^{d_A}} \frac{\Gamma(d_B \alpha_B)}{\Gamma(\alpha_B)^{d_B}} \times \\ & \frac{\prod_i \Gamma(n_{i.} + \alpha_A)}{\Gamma(n + d_A \alpha_A)} \frac{\prod_j \Gamma(n_{.j} + \alpha_B)}{\Gamma(n + d_B \alpha_B)}. \end{aligned}$$

If we select the uniform prior we obtain less canceling of terms than we did for $M_1$ in equation (20):

$$p_u(\overline{n}|M_2) = \frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)\Gamma(n+d_B)} \frac{\prod_i \Gamma(n_{i.}+1) \prod_j \Gamma(n_{.j}+1)}{\prod_{ij} \Gamma(n_{ij}+1)}. \qquad (24)$$

From equations (22) and (24) we obtain the Bayes factor for the undirected data model:

$$\frac{p_u(M_2|D)}{p_u(M_1|D)} = \frac{p_u(\overline{n}|M_2)}{p_u(\overline{n}|M_1)} =$$
$$\frac{\Gamma(n+d_Ad_B)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)\Gamma(n+d_B)\Gamma(d_Ad_B)} \frac{\prod_j \Gamma(n_{.j}+1) \prod_i \Gamma(n_{i.}+1)}{\prod_{ij} \Gamma(n_{ij}+1)}. \quad (25)$$

The second approach to model independence between $A$ and $B$ gives the following:

$$p_u(\overline{n}|M_2') =$$
$$\frac{\Gamma(n+1)\Gamma(d_A)}{\Gamma(n+d_A)} \int (\prod_i \binom{n_{i.}}{n_{i1} \ldots n_{id_B}} \prod_j x_j^{n_{ij}}) \Gamma(d_B) d\overline{x}^B =$$
$$\frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)} (\prod_i \binom{n_{i.}}{n_{i1} \ldots n_{id_B}}) \prod_j x_j^{n_{.j}} d\overline{x}^B =$$
$$\frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)\Gamma(n+d_B)} \frac{\prod_i \Gamma(n_{i.}+1) \prod_j \Gamma(n_{.j}+1)}{\prod_{ij} \Gamma(n_{ij}+1)}. \quad (26)$$

We can now find the Bayes factor relating models $M_1'$ (equation 23) and $M_2'$ (equation 26), with no prior preference of either:

$$\frac{p_u(M_2'|D)}{p_u(M_1'|D)} = \frac{p_u(\overline{n}|M_2')}{p_u(\overline{n}|M_1')} =$$
$$\frac{\prod_j \Gamma(n_{.j}+1) \prod_i \Gamma(n_{i.}+d_B)}{\Gamma(d_B)^{d_A-1}\Gamma(n+d_B) \prod_{ij} \Gamma(n_{ij}+1)} \quad (27)$$

Consider now a data matrix with three variables, $A$, $B$ and $C$ (figure 2). The analysis of the model $M_3'$ where full dependencies are accepted is very similar to $M_1$ above (equation 22). For the model $M_4$ without the link between $A$ and $B$ we should partition the data matrix by the value of $C$ and multiply the probabilities of the blocks with the probability of the partitioning defined by $C$.

Since we are ultimately after the Bayes factor relating $M_4$ and $M_3$ respectively $M_4'$ and $M_3'$, we can simply multiply the Bayes factors relating $M_2$ and $M_1$ (equation 25) respectively $M_2'$ and $M_1'$ (equation 27) for each block of the partition to get the Bayes factors sought:

$$\frac{p_u(M_4|D)}{p_u(M_3|D)} = \frac{p_u(\overline{n}|M_4)}{p_u(\overline{n}|M_3)} =$$
$$\frac{\Gamma(d_A)^{d_C}\Gamma(d_B)^{d_C}}{\Gamma(d_Ad_B)^{d_C}} \prod_c \frac{\Gamma(n_{..c}+d_Ad_B) \prod_j \Gamma(n_{.jc}+1) \prod_i \Gamma(n_{i.c}+1)}{\Gamma(n_{..c}+d_A)\Gamma(n_{..c}+d_B) \prod_{ij} \Gamma(n_{ijc}+1)} \quad (28)$$

and in the directed case we have:

$$\frac{p_u(M_4'|D)}{p_u(M_3'|D)} = \frac{p_u(\overline{n}|M_4')}{p_u(\overline{n}|M_3')} =$$
$$\Gamma(d_B)^{(d_A+1)d_C} \prod_c \frac{\prod_j \Gamma(n_{.jc}+1) \prod_i \Gamma(n_{i.c}+d_B)}{\Gamma(n_{..c}+d_B) \prod_{ij} \Gamma(n_{ijc}+1)}.$$

For analysis of directed graphical models in the next section, we must also be able to compare models $M_5$ and $M_6$ of figure 5:

$$\frac{p_u(M_5|D)}{p_u(M_6|D)} = \frac{p_u(\overline{n}|M_5)}{p_u(\overline{n}|M_6)} = \frac{1}{\Gamma(d_B)^{(d_A-1)d_C}} \prod_c \frac{\prod_j \Gamma(n_{.jc}+1) \prod_i \Gamma(n_{i.c}+d_B)}{\Gamma(n_{..c}+d_B) \prod_{ij} \Gamma(n_{ijc}+1)}.$$

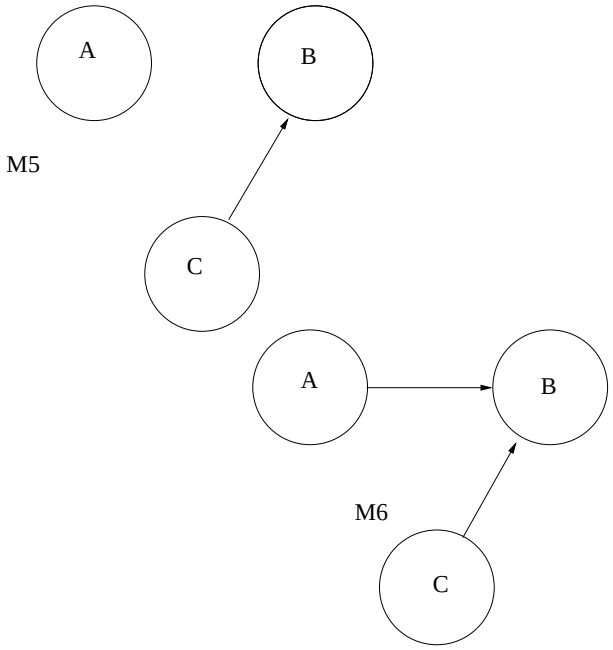


Figure 5: Directed models

# 5 Graphical model choice - global analysis

If we have many variables, their interdependencies can be modeled as a graph with vertices corresponding to the variables. The example of figure 6 is from [23], and shows the dependencies in a data matrix related to heart disease. Of course, a graph of this kind can give a data probability to the data matrix in a way analogous to the calculations in the previous section, although the formulae become rather involved, and the number of possible graphs increases dramatically with the number of variables. It is completely infeasible to list and evaluate all graphs if there is more than a handful of variables. An interesting possibility to simplify the calculations would use some kind of separation, so that an edge in the model could be given a score independent of the inclusion or exclusion of most other potential edges. Indeed, the derivations of last section show how this works. Let $C$ in that example be a compound variable, obtained by merging columns $\{c_1, \ldots c_d\}$. If two models $G$ and $G'$ differ only by the

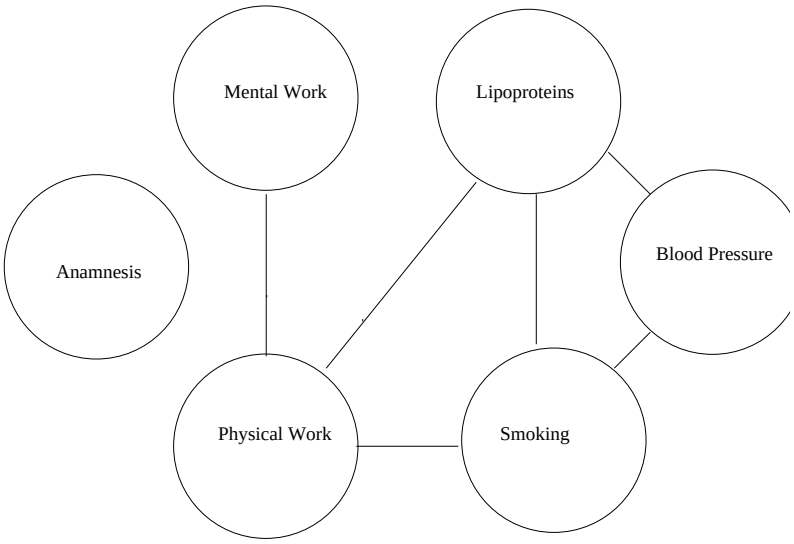


Figure 6: Symptoms and causes relevant to heart problems

presence and absence of the edge $\{A, B\}$, and if there is no path between $A$ and $B$ except through vertex set $C$, then the expressions for $p(\overline{n}|M_4)$ and $p(\overline{n}|M_3)$ above will become factors of the expressions for $p(\overline{n}|G)$ and $p(\overline{n}|G')$, respectively, and the other factors will be the same in the two expressions. Thus, the Bayes factor relating the probabilities of $G$ and $G'$ is the same as that relating $M_4$ and $M_3$. This result is independent of the choice of distributions and priors of the model, since the structure of the derivation follows the structure of the graph of the model - it is equally valid for Gaussian or other data models, as long as the parameters of the participating distributions are assumed independent in the prior assumptions. A beautiful abstract analysis of this phenomenon can be found in (Dawid and Lauritzen[12]).

We can now think of various 'greedy' methods for building high probability interaction graphs relating the variables (columns in the data matrix). It is convenient and customary to restrict attention to either decomposable(chordal) graphs or directed acyclic graphs. Chordal graphs are fundamental in many applications of describing relationships between variables (typically variables in systems of equations or inequalities). They can be characterized in many different but equivalent ways, see (Rose [29], Rose, Lueker and Tarjan[30]). One simple way is to consider a decomposable graph as consisting of the union of a number of maximal complete graphs (cliques, or maximally connected subgraphs), in such a way that (i) there is at least one vertex that appears in only one clique (a *simplicial vertex*), and (ii) if an edge to a simplicial vertex is removed, another decomposable graph remains, and (iii) the graph without any edges is decomposable. A characteristic feature of a simplicial vertex is that its neighbors are completely connected. This recursive definition can be reversed into a generation procedure: Given a decomposable graph $G$ on the set of vertices, find two vertices $s$ and $n$ such that (i): $s$ is simplicial, *i.e.*, its neighbors are completely connected, (ii): $n$ is connected to all neighbors of $s$. Then the graph

$G'$ obtained by adding the edge between $s$ and $n$ to $G$ is also decomposable. We will call such an edge a *permissible edge* of $G$. This procedure describes a generation structure (a directed acyclic graph whose vertices are decomposable graphs on the set of vertices) containing all decomposable graphs on the variable set. An interesting feature of this generation process is that it is easy to compute the Bayes factor comparing the posterior probabilities of the graphs $G$ and $G'$ as graphical models of the data: Let $s$ correspond to $A$, $n$ to $B$ and the compound variable obtained by fusing the neighbors of $s$ to $C$ in the analysis of section 5. Without explicit prior model probabilities we have:

$$\frac{p(G|D)}{p(G'|D)} = \frac{p(\overline{n}|M_3)}{p_u(\overline{n}|M_4)}. \tag{29}$$

A search for high probability graphs can now be organized as follows:

1. Start from the graph $G_0$ without edges.

2. Repeat: find a number of permissible edges that give the highest Bayes factor, and add it if the factor is greater than 1. Keep a set of highest probability graphs encountered.

3. Then repeat: For the high probability graphs found in the previous step, find simplicial edges whose removal increases the Bayes factor the most (or decreases it the least).

For each graph kept in this process, its Bayes factor relative to $G_0$ can be found by multiplying the Bayes factors in the generation sequence. A procedure similar to this one is reported by (Madigan and Raftery[24]), and its results on small variable sets was found good, in that it found the best graphs reported in other approaches. It must be noted, however, that we have now passed into the realm of approximate analysis, since we cannot (yet) know that we will find all high probability graphs. One splendid example of this is where we have many binary categorical columns, all generated randomly and independently of each other except the last one which is the parity function of the other ones. If we start searching from the empty graph, we will never find this relationship since the intermediate graphs will have low probability. Likewise, if some arbitrary subset of the columns are interrelated by a parity constraint it seems unlikely although possible that we will find it even if we start the search from the saturated model (graph with all edges).

Another family of graphical models are the directed acyclic models. They can be treated similarly, since here we check locally for a variable $B$ that has been found dependent on a set $C$, whether it can be inferred also to depend on variable $A$. We compare thus models $M_5$ and $M_6$ of figure 4. The inclusion or exclusion of the arrow from $A$ to $B$ can be inferred independent of all arrows not going to $B$. A problem with directed graphical models is that different acyclic graphs can represent the same family of probability distributions, and this requires some careful argumentation.

# 6 Graphical model choice - categorical, ordinal and Gaussian variables

We now consider data matrices made up from ordinal and real valued data, and then matrices consisting of both ordinal, real and categorical data. The standard

choice for a real valued data model is the univariate or multivariate Gaussian or normal distribution. It has nice theoretical properties manifesting themselves in such forms as the central limit theorem, the least squares method, principal components, etc. However, it must be noted that it is also unsatisfactory for many data sets occurring in practice, because of its narrow tail and because many real life distributions deviate terribly from it. Several approaches to solve this problem are available. One is to consider a variable as being obtained by mixing several normal distributions. This is a special case of the classification or segmentation problem discussed below. Another is to disregard the distribution over the real line, and considering the variable as just being made up of an ordered set of values. This leads naturally to the recursive splitting of the data set by a decision tree, also discussed below.

# 7 Missing values and errors in data matrix

Data collected from experiments are seldom perfect. The problem of missing and erroneous data is a vast field in the statistics literature. First of all there is a possibility that 'missingness' of data values are significant for the analysis, in which case missingness should be modeled as an ordinary data value. Then the problem has been internalized, and the analysis can proceed as usual, with the important difference that the missing values are not available for analysis. A more sceptical approach was developed by Ramoni and Sebastiani[27], who consider an option to regard the missing values as adversaries (the conclusions on dependence would then be true no matter what the missing values are). The other possibility is that missingness is known to have nothing to do with the objectives of the analysis. For example, in a medical application, if data is missing because of the bad condition of the patient, missingness is significant if the investigation is concerned with patients. But if data is missing because of unavailability of equipment, it is probably not - unless maybe if the investigation is related to hospital quality. In Bayesian data analysis, the problem of missing or erroneous data creates significant complications, as we will see. As an example, consider the analysis of the two-column data matrix with binary categorical variables $A$ and $B$, analyzed against models $M_1$ and $M_2$ of section 5. Suppose we obtained $n_{00}$, $n_{01}$, $n_{10}$ and $n_{11}$ cases with the values 00, 01, etc. We then have a posterior Dirichlet distribution with parameters $n_{ij}$ for the probabilities of the four possible cases. If we now receive a case where both $A$ and $B$ are unknown, it is reasonable that this case is altogether ignored. But what shall we do if a case arrives where $A$ is known, say 0, but $B$ is unknown? One possibility is to waste the entire case, but this is not orthodox Bayesian, since we are not making use of information we have. Another possibility is to use the current posterior to estimate a pdf for the missing value, in our case the probability that $B$ has value 0 is $p_0 = n_{00}/n_{0\cdot}$. So our posterior is now either a Dirichlet with parameters $n_{00}$, $n_{01}-1$, $n_{10}-1$ and $n_{11}-1$ (probability $p_0$) or one with parameters $n_{00}-1$, $n_{01}$, $n_{10}-1$ and $n_{11}-1$ (probability $1-p_0$). But this means that the posterior is now a weighted average of two Dirichlet distributions, in other terms, is not a Dirichlet distribution at all! As the number of missing values increases, the number of terms in the posterior will increase exponentially, and the whole advantage with conjugate distributions will be lost. So wasting the whole case seems to be a reasonable option unless we find a more

clever way to proceed.

The related case of errors in data is more difficult to treat. How do we describe data where there are known uncertainties in the recording procedure? This is a problem worked on for centuries when it comes to real valued quantities as measured in physics and astronomy, and is one of the main features of interpretation of physics experiments. When it comes to categorical data there is less help in the literature - an obvious alternative is to relate recorded vs actual values of discrete variables as a probability distribution, or - which is fairly expedient in our approach - as an equivalent sample.

# 8 Decision trees

Decision trees are typically used when we want to predict a variable - the class variable - from other - explanatory - variables in a case, and we have a data matrix of known cases. When modeling data with decision trees, we are usually trying to segment the data set into ranges - $n$-dimensional boxes of which some are unbounded - such that a particular variable - the class variable - is fairly constant over each box. If the class variable is truly constant in each box, we have a tree that is consistent with respect to the data. This means that for new cases, where the class variable is not directly available, it can be well predicted by the box into which the case falls. The method is suitable where the variables used for prediction are of any kind (categorical, ordinal or numerical) and where the predicted variable is categorical or ordinal with a small domain. There are several efficient ways to heuristically build good decision trees, and it is a central technique in the field of machine learning. Practical experience has given many cases were the predictive performance of decision trees is good, but also many counter-intuitive phenomena have been uncovered by practical experiments. Recently, several treatments of decision trees have been published where it is discussed whether or not the smallest possible tree consistent with all cases is the best one. This turned out not to be the case, and the argument that a smallest decision tree should be preferred because of some kind of Occam's razor argument is apparently not valid, neither in theory nor in practise[34, 2]. The Bayesian approach gives the right information on the credibility and generalizing power of a decision tree. It is explained in recent papers by (Chipman, George and McCullogh[9]) and by (Paass and Kindermann[26]). A decision tree statistical model is one where a number of boxes are defined on one set of variables by recursive splitting of one box into two by splitting the range of one designated variable into two. Data are assumed to be generated by a discrete distribution over the boxes, and for each box it is assumed that the class variable value is generated by another discrete distribution. Both these distributions are given uninformative Dirichlet prior distributions, and thus the posterior probability of a decision tree can be computed from data. Since larger trees have more parameters, there is an automatic penalization of large trees, but the distribution of cases into boxes also enters the picture, so it is not clear that the smallest tree giving perfect classification will be preferred, or even that a consistent tree will be preferred over an inconsistent one. The decision trees we described here do not give a clear cut decision on the value of the decision variable for a case, but a probability distribution over values. If the probability distribution is not peaked at a specific class value, then this indicates that pos-

sibly more data must be collected before a decision can be made. Also, since the name of this data model indicates its use for decision making, one can get better trees for an application by including information about the utility of the decision in the form of a loss function and by comparing trees based on the expected utility rather than model probability.

For a decision tree $T$ with $d$ boxes data with $c$ classes, and where the number of cases in box $i$ with class value $k$ is $n_{ik}$, and $n = n_{..}$, we have with uniform priors on both the assignment of case to box and of class within box,

$$p(D|T) = \frac{\Gamma(n+1)\Gamma(d)}{\Gamma(n+d)} \prod_i \frac{\Gamma(n_{i.}+1)\Gamma(c)}{\Gamma(n_{i.}+c)} \tag{30}$$

However, in order to compare two trees $T$ and $T'$, we would have to form the set of intersection boxes and ask about the probability of finding the data with a common parameter over the boxes belonging to a common box of $T$ relative to the probability of the data when the parameters are common in boxes of $T'$. For the case where $T$ and $T'$ only differ by splitting of one box $i$ into $i'$ and $i''$, the calculation is easy ($n_{i''j} + n_{i'j} = n_{ij}$):

$$\frac{p(D|T')}{p(D|T)} = \frac{\Gamma(n_{i.}+c)}{\Gamma(n_{i'.}+c)\Gamma(n_{i''.}+c)} \prod_j \frac{\Gamma(n_{i'j}+1)\Gamma(n_{i''j}+1)}{\Gamma(n_{ij}+1)} \tag{31}$$

# 9 Segmentation - Latent variables

Segmentation and latent variable analysis is directed at describing the data set as a collection of subsets, each having simpler descriptions than the full data matrix. Suppose data set $D$ is partitioned into $d_c$ classes $\{D^{(i)}\}$, and each of these has a high posterior probability $p(D^{(i)}|M_i)$ wrt some model set $\{M_i\}$. Then we think that the classification is a good model for the data. However, some problems remain to consider. First, what is it that we compare the classification against, and second, how do we accomplish the partitioning of the cases? The first question is the simplest to answer: we compare a classification model against some other model, based on classification or not. The second is trickier, since the introduction of this section is somewhat misleading. The prior information for a model based on classification must have some information about classes, but it does not have an explicit division of the data into classes available. Indeed, if we were allowed to make this division into classes on our own, seeking the highest posterior class model probabilities, we would probably over-fit by using the same data twice - once for class assignment and once for posterior model probability computation. The statistical model generating segmented data could be the following: A case is first assigned to a class by a discrete distribution obtained from a suitable uninformative Dirichlet distribution, and then its visible attributes are assigned by a class-dependent distribution. This model can be used to compute a probability of the data matrix, and then, via Bayes rule, a Bayes factor relating the model with another one, e.g., one without classes or with a different number of classes. One can also have a variable number of classes and evaluate by finding the posterior distribution of the number of classes. The data probability is obtained by integrating, over the Dirichlet distribution, the sum over all assignments of cases to classes,

of the assignment probability times the product of all resulting case probabilities according to the respective class model. Needless to say, this integration is feasible only for a handful of cases where the data is too meager to permit any kind of significant conclusion on the number of classes and their distributions. The most well-known procedures for automatic classification are built on expectation maximization. With this technique, a set of class parameters are refined by assigning cases to classes probabilistically, with the probability of each case membership determined by the likelihood vector for it in the current class parameters[8]. We can also solve the problem with the MCMC approach[28]. The MCMC approach to classification is the following: Assume that we have a data matrix and want a classification of its cases which makes the attributes independent. Define a class assignment randomly, and compute the probability of data, given the model with independent attributes, as in (24) which is easy to generalize to more attributes. The MCMC will now implement a move function, proposing a changed class for some case. The move is accepted if the posterior probability increases, or otherwise by a probability given by the ratio of new to old data probability (see section 11). This procedure is reasonably efficient, since it is possible to evaluate the class probabilities incrementally, by keeping just the current contingency table for each class and updating it incrementally. Since absolute probabilities are held updated, we also avoid a common complication in MCMC applications arising when the dimension of the parameter space changes. Although it can sometimes be avoided it is not always so. The reversible jump process was designed to cope with this phenomenon[6].

# 10 Association rules

Association rules are special sets of rules used to predict data in data mining. The literature on association rules emphasizes rapid extraction, since typically a data matrix has very many potential association rules and the data matrices considered are very large. An association rule is written $A \to B$, where $A$ and $B$ are conditions on a data case. They can be either defined by giving a predicate on the value of an attribute, or as a conjunction of such conditions for several attributes. In the literature, binary attributes are often assumed. These usefulness of this rule depends on how well it satisfies the intuitive condition of the rule: Whenever $A$ is true for a case, $B$ is also true. The *support* of the rule is the fraction of cases where both $A$ and $B$ are true, whereas the *confidence* is the fraction of cases with $A$ true where also $B$ is true. The *lift* of a rule is the factor by which its confidence exceeds the confidence we would have with in-dependency between $A$ and $B$, computing in a ML framework, i.e., $n_{AB}n_{..}/(n_{A.}n_{.B})$, where the notation is an obvious adaptation of the contingency table notation used previously. Clearly, the concept of lift assumes a large database, where statistical fluctuation can be ignored. In order to asses the significance of an association rule, we need the machinery of Bayes factors, and then we can easily assess the generalization expectable from a proposed rule. In short, let the *significance* of a rule be the Bayes factor between a model that gives dependence between $A$ and $B$ and a model that does not. This gives us a simpler version of the decision tree rule, equation (31):

$$s(A \to B) = \frac{\Gamma(n_{..}+2)}{\Gamma(n_{A.}+c)\Gamma(n_{\overline{A}.}+c)} \frac{\Gamma(n_{AB}+1)\Gamma(n_{\overline{A}\overline{B}}+1)}{\Gamma(n_{.B}+1)} \frac{\Gamma(n_{A\overline{B}}+1)\Gamma(n_{\overline{A}\overline{B}}+1)}{\Gamma(n_{.\overline{B}}+1)} \tag{32}$$

Since the significance depends on four quantities which can have a very large span of values in practical applications, this concept seems necessary for throwing out rules that cannot be expected to generalize because either the data base, the support or the lift (or some combination) is too small. However, there are more dangers in applying data mining results, particularly the problem of biased sampling which no test on sampled data can reveal.

Mining of large files for association rules typically reveals very large quantities of significant rules. Many papers have been devoted to finding an interesting subset of such rules. Two basic approaches exist: in one, a measure of interestingness or surprisingness is defined for a particular rule, in another a rule is evaluated in the context of an already existing set.

# 11 Approximate analysis with Metropolis-Hastings simulation

Several of the cases mentioned previously, where analytical solutions become infeasible because of breakdown of the simple conjugacy principle (missing and erroneous values) or the large number of models to be considered (graphical models on many variables) or because of the analytical difficulties in computing data probability (classification) have been separately attacked with monte carlo methods. We will outline a method solving all these cases at once.

The basic problem solved by MCMC methods is sampling from a multivariate distribution over many variables. The distribution can be given generically as $p(x, y, z, \ldots, w)$. If some variable, e.g., $y$, represents measured signals, then the actual values measured, say $a$, can be substituted, and sampling will be from a conditional distribution $p(x, a, z, \ldots, w)$. If some other variable, say $x$, represents the measured quantity, then sampling and selecting the $x$ variable will give samples from the posterior of $x$ given the measurements. In other words, we will get a best possible estimation of the quantity given the measurements and the statistical model of the measurement process.

The two basic methods for MCMC computation are the Gibbs sampler and the Metropolis-Hastings algorithm. Both generate a Markov chain with states over the domain of a multivariate target distribution and with the target distribution as its unique limit distribution. Both exist in several more or less refined versions. The Metropolis algorithm has the advantages that it does not require sampling from the conditional distributions of the target distribution but only finding the quotient of the distribution at two arbitrary given points, and it can be chosen from a set with better convergence properties. A thorough introduction is given by Neal[25]. To sum it up, MCMC methods can be used to estimate distributions that are not tractable analytically or numerically. We get real estimates of posterior distributions and not just approximate maxima of functions. On the negative side, the Markov chains generated have high autocorrelation, so a sample over a sequence of steps can give a highly misleading empirical distribution, much narrower than the real posterior. Although signifi-

cant advances have been made in the area of convergence assessment and choice of samples, this problem is not yet completely solved.

The Metropolis-Hastings sampling method is organized as follows: given the pdf $p(q)$ of a state variable $q$ over a state space $Q$, and an essentially arbitrary symmetric move function $m(q, q')$, a sequence of states is created in a Markov chain. In state $q$, draw $q'$ according to the move function $m$. If $p(q')/p(q) > 1$, let $q'$ be the new state. Otherwise, let $q'$ be the new state with probability $p(q')/p(q)$, otherwise keep state $q$. It is easy to verify that $p(q)$ is a stable distribution of the chain, and from general Markov chain theory there are several conditions that ensure that there is only one limiting distribution and that it will always be reached asymptotically. It is much more difficult to say when we have a provably good sample, and in practice all the difficulties of hill-climbing optimization methods must be dealt with in order to assess convergence.

Nevertheless, there have been great successes with this method for those cases of Bayesian analysis where closed form solutions do not exist. With various adaptations of the method, it is possible to express multivariate data as a mixture of multivariate distributions and to find the posterior distribution of the number of classes and their parameters [14, 15, 16, 28].

The missing data problem can also be solved in the sense that parameters and dependence structures can be estimated with missing data without the simple expedient of wasting incomplete cases.

The structure of a graphical model can be obtained as a sample from the posterior distribution[4, 13].

## 11.1 Example: Univariate Gaussian Mixture modeling

Consider the problem of deciding, for a set of real numbers, the most plausible decompositions of the distribution as a weighted sum (mixture) of a number of distributions each being a univariate normal (Gaussian) distribution. This problem has significance when we try to find 'discrete' circumstances behind a measured variable which is also influenced by various chance fluctuations. Note that a single column of discrete data is not decomposable in this way because a mixture of discrete distributions is again a discrete distribution. But a mixture of normal distributions is not itself a normal distribution. In the frequently used Enzyme problem[28], the discovered components, if any, could correspond to genetic factors in a population. There are quite many approaches to solve this problem, and many carry over to the more general problem of modeling a matrix of reals as coming from a mixture of multivariate Gaussians[]. The approach presented here seems to have an advantage in that it is not necessary to include the parameters (mean and variance) of the participating Gaussians. It simulates only the assignment of variables to classes, and for each such assignment it computes the exact posterior of the data including the latent class variable. This posterior is, for class $I$:

$$p(D|I) = |I|! \int \prod_{i \in I} p(d_i|\Theta) p(\Theta) d\Theta, \tag{33}$$

If we assume a uniform prior for the distribution over classes, this factor will only depend on the total sample size and can be ignored (see equation (20)),

so we just multiply together the contributions from each class to get the model probability for the current class assignment.

Just as the discrete probability distribution has the Dirichlet conjugate family, the Gaussian univariate has a conjugate family, which is a distribution over the mean and variance parameters of the Gaussian. It is possible to define priors were the mean and variance parameters of the Gaussian are a priori independent, see [7], but with that prior the data probability is not expressible in closed form (it involves the exponential integral). For the natural conjugate of a Gaussian, in the notation of Bernardo and Smith[3], we use the inverse of the variance as a parameter (the precision) and have the Gaussian distribution:

$$f(x|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi}} e^{-\lambda/2(x-\mu)^2}, \tag{34}$$

and the natural prior for $\mu$ and $\lambda$ is a normal-gamma distribution where the precision $\lambda$ is drawn from a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. The mean $\mu$ is then normally distributed with mean $\mu_0$ and precision $n_0\lambda$, where $n_0 > 0$ and $\mu_0$ are new constants:

$$p(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \tag{35}$$

$$p(\mu|\mu_0,n_0,\lambda) = \sqrt{\frac{n_0\lambda}{2\pi}} e^{-n_0\lambda/2(\mu-\mu_0)^2}, \tag{36}$$

The joint probability distribution is thus:

$$p(D,\mu,\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} (\frac{\lambda}{2\pi})^{n/2} e^{-\lambda/2\Sigma_i(x_i-\mu)^2} \sqrt{\frac{n_0\lambda}{2\pi}} e^{-n_0\lambda/2(\mu-\mu_0)^2} \tag{37}$$

and we want to obtain the data probability by integrating over $\mu$ and $\lambda$, which after some calculation and the substitutions $n's^2 = \Sigma x_i^2 + n_0\mu_0^2$, $n'm = \Sigma x_i + n_0\mu_0$ and $n' = n + n_0$ yields:

$$p(D|n_0,\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} n!(2\pi)^{-n/2} \sqrt{\frac{n_0}{n'}} \int_0^\infty e^{-n'\lambda/2(s^2-m^2)} \lambda^{n/2+\alpha-1} e^{-\beta\lambda} d\lambda \tag{38}$$

The value of the integral is the inverse normalization constant for a gamma distribution with parameters $(\alpha + n/2, \beta + n'/2(s^2 - m^2))$ and can be obtained by substituting in (35). The probability of data $D$ and class assignment $\{I_c\}_{c\in C}$ is then

$$p(D,\{I_c\}_{c\in C}) = (2\pi)^{n/2} \sqrt{n_0} \Pi_c \frac{n_c!\Gamma(\alpha + n_c/2)}{\sqrt{n_c'}(\beta + n_c'/2(s_c^2 - m_c^2))} \tag{39}$$

The choice of the parameters can be made so that $\mu$ and $\lambda$ are fairly evenly distributed over a range covering the values found likely by inspection of the data. The coupling of the precisions of the distributions of the data points and of the prior class mean seems to constrain this model in a bad way, since it penalizes sharp peaks in the outskirts of the distribution. A theoretically

sound way to deal with this problem is to assign a so called *hyper-prior* for the parameter $\mu_0$. The result of such an exercise is a *hierarchical* model.

Equation (39) gives an exact probability of data and class assignment, given parameters $\mu_0$, $\alpha$, $\beta$ and $n - 0$. The Metropolis-Hastings proposal will be a reassignment of the class of one data point. This changes the $n_c$, $m_c$ and $s_c^2$ values of two classes with amounts easily computed. The resulting density ratio of the distribution (39) which controls the probability of taking the proposed move, is also easy to compute. In the hierarchical model we would also have a class-specific prior mean $\mu_c$, which is also recomputed in a move. It is practical to forbid two class means to switch by these moves, so that the classes can always be recognized by their relative prior means.

# References


[1] J. O. Berger. *Statistical decision Theory and Bayesian Analysis.* Springer-Verlag, 1985.

[2] N. Berkman and T. Sandholm. What should be optimized in a decision tree? Technical report, University of Massachusetts at Amherst, 1995.

[3] Jose M. Bernardo and Adrian F. Smith. *Bayesian Theory.* Wiley, 1994.

[4] C. Berzuini, N. G. Best, W. R. Gilks, and C. Larizza. Dynamic graphical models and markov chain monte carlo methods. Technical report, MRC Biostatistics Unit, Cambridge, 1994.

[5] A. Birnbaum. On the foundations of statistical inference (with discussion). *J. American Statistical Ass.*, 57:269–326, 1962.

[6] B. P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57:473–484, 1995.

[7] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, 1997.

[8] P. Cheeseman and J. Stutz. Bayesian classification (AUTOCLASS): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining.* 1995.

[9] H. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart. Technical report, University of Chicago, 1995.

[10] D. R. Cox and Nanny Wermuth. *Multivariate Dependencies.* Chapman and Hall, 1996.

[11] R.T. Cox. Probability, frequency, and reasonable expectation. *Am. Jour. Phys.*, 14:1–13, 1946.

[12] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21:1272–1317, 1993.

[13] P. Dellaportas and J. Forster. Markov chain monte carlo model determination for hierarchical and graphical log-linear models. Technical report, Athens University of Economics, Greece., 1996.

[14] D. K. Dey, L. Kuo, and S. K. Sahu. A bayesian predictive approach to determining the number of components in a mixture distribution. Technical report, University of Connecticut, 1993.

[15] J. Diebolt and C. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56:589–590, 1994.

[16] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[17] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[18] David Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1:79–119, 1997.

[19] E. T. Jaynes. *Probability Theory: The Logic of Science*. Preprint: Washington University, 1996. `http://bayes.wustl.edu/etj/prob.html`.

[20] Harold Jeffreys. *Scientific Inference*. Cambridge University Press, 1931.

[21] Harold Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1939.

[22] Steffen L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

[23] D. Madigan and A. E. Raftery. Model selection and accounting fro model uncertainty in graphical models using occam's window. Technical report, University of Washington, 1993.

[24] David Madigan and Adrian E. Raftery. Model selection and accounting for model uncertainty in graphical models using occams window. *J. American Statistical Ass.*, 428:1535–1546, 1994.

[25] R M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical report, Department of Computer Science, University of Toronto, 1993. CRG-TR-93-1.

[26] Gerhard Paass and Jörg Kindermann. Bayesian classification trees with overlapping leaves applied to credit-scoring. *Lecture Notes in Computer Science*, 1394:234–245, 1998.

[27] M. Ramoni and P. Sebastiani. Parameter estimation in Bayesian networks from incomplete databases. *Intelligent Data Analysis*, 2, 1998.

[28] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, 59:731–792, 1997.

[29] Donald J. Rose. Triangulated graphs and the elimination process. *J. Math. Anal. Appl.*, 32:597–609, 1970.

[30] Donald J. Rose, Robert Endre Tarjan, and George S. Lueker. Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.*, 5:266–283, 1976.

[31] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

[32] D. S. Sivia. *Bayesian Data Analysis, A Bayesian Tutorial.* Clarendon Press: Oxford, 1996.

[33] D. J. Spiegelhalter, A. Thomas, N. G. Best, and W. R Gilks. *BUGS: Bayesian Inference using Gibbs Sampling. Version 0.50.* MRC Biostatistics Unit, Cambridge, 1996.

[34] G.I. Webb. Further experimental evidence against the utility of occams razor. *Journal of AI research*, 4:397–417, 1996.

[35] J. Whittaker. *Graphical Models in Multivariate Statistics.* Wiley, 1990.