Statistical Methods in Applied Computer Science Lecture Notes Course DD2447, FDD3342, Jan-April 2011.

Stefan Arnborg [*]
KTH CSC/Nada

February 7, 2011

Abstract

In the real world, lots of things are uncertain; including the proper way to handle uncertainty.
Old Jungle Saying

Contents

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

1 Introduction

Computers have since their invention been described as devices performing fast and precise computations. For this reason it might be easy to conclude that probability and statistics has little to do with computer science, as quite many of our students have indeed done. This attitude may have had some justification before computers became more than a prestigious and expensive piece of equipment. Today, computers are involved everywhere, and the ubiquitous uncertainty characterizing the real non-computer world can no longer be separated from computing. Automatically recorded data are collected in vast amounts, and the kind of interpretation of sounds, images and other recordings that was earlier made by humans, or under close supervision by humans, must now inevitably be performed to larger and larger extent by computers, unaided by humans. If we emphasize the human performance of interpretation as a gold standard, we may call the methodology used in such applications *artificial intelligence*. If we emphasize that we must make machines learn from experience we may call it *machine learning*. Other popular names of the emerging technologies are *data mining*, emphasizing the large amounts of data and the importance of finding business opportunities in them, or perhaps *uncertainty management*, emphasizing that computer applications must be made more and more aware of the basic uncertainties in the real world. The terms *pattern recognition* and *information fusion* have also been used. *Knowledge discovery* is another term that evokes emotions, particularly in natural sciences. These notes were made for a course in statistical methods in computer engineering, where the theme is to unify the methodology in all the above fields with probabilistic and statistical tools. By this you may miss some lore of a specific field, but probably it will not matter much.

1.1 Professional context of statistics in Computer Science

Data acquired for analysis can have many different formats. We will describe the analysis of measurements that can be thought of as samples drawn from a larger population, and the conclusions will be phrased in terms of this larger population. I will focus on very simple models. My experience is that it is too easy to start working with complex models without understanding the basics, which can lead to all sorts of erroneous conclusions. As the investigator's understanding of a problem area improves, the statistical models tend to become complex. Some examples of such areas are genetic linkage studies[24], ecosystem studies[68] and functional MR investigations[109], where the signals extracted from measurements are very weak but potentially extremely useful for the application area. As an example, analysis of fMRI experiments have complex models for the signal distortion, noise, movement artifacts, variation of anatomy and function among different subjects, variations within the individual and even drift in a single experiment. All these disturbances can be modeled using basic physics, knowledge of anatomy and statistical summaries on inter-brain variation and the total model is overwhelmingly complex[69]. Experiments are typically analyzed using a combination of visualization, Bayesian analysis and conventional test and confidence based statistics. When it becomes necessary to develop more sophisticated models, it is vital that the analyst communicates the developments to the non-statistical members of the research team. In engineering and commercial applications of data mining, the goal is not normally to arrive at eternal truths, but to support decisions in design and business. Nevertheless, because of the competitive nature of these activities, one can expect well founded analysis methods and understandable models to provide more useful answers than ad hoc ones. In any case, even with short deadlines it seems important that the methodological discussion in a project contains more than the question of which buttons to press in commercial data analysis packages.

If each sample point contains measurements of two variables, it is easy to produce a scatter plot in two dimensions where sometimes a conclusion is immediate. The human eye is very good at seeing structure in such plots. However, sometimes it is too good, and constructs structure that does simply not exist. If one makes a scatter plot with points drawn uniformly inside a polygon, the point cloud will typically be deemed dense in some areas and sparse in others, in other words one sees structure that does not exist in the underlying distribution, as exemplified in figure 1.

There are several other reasons to complement visualization methods with more sophisticated statistical methods: Most data sets will contain many more than two variables, and this leads to a dimensionality problem. Producing pairwise plots means that many plots are produced. It is difficult to examine all of them carefully, and some of them will inevitably accidentally contain structure that would be deemed real in a single plot examination but not as an extreme among many plots. It would also mean that co-variation of several variables that cannot be explained as a number of two-variable co-variations cannot be found. Nevertheless, scatter plots are among the most useful explanatory devices once a structure in the data has been verified to be 'probably real'.

This text emphasizes characterization of data and the population from which it is drawn with its statistical properties. Nonetheless, the application owners have typically very different concerns: they want to understand, they want to be able to predict and ultimately to control their objects of study. This means that the statistical investigation is a first phase, which must be accompanied by interpretation, activities extracting meaning from the data. There is relatively little theory on these later activities, and it is probably fair to say that their outcome depends mostly on the intellectual climate in the team of which the analyst is only one part.

1.2 Summary and related work

Our purpose is to explain some advantages of the Bayesian approach and to show how probability models can capture 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 present projects in medical informatics, defense science and customer segmentation. Several of the computations we describe have been analyzed 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 mostly confined to pure Bayesian analysis and puts several established data mining methods in a joint Bayesian framework. 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 - which are computationally demanding - and variational Bayes methods - which are approximate - are available. There are several recent books describing the Bayesian method from both a theoretical[16], an ideological[95] and an application oriented[19] perspective. Particularly Ed Jaynes unfinished lecture notes[61] have provided inspiration for me and numerous students using them all over the world. A current survey of MCMC methods, which can solve many complex evaluations required in advanced Bayesian modeling, can be found in the book[48]. Books explaining theory and use of graphical models are Lauritzen[64] and Cox and Wermuth[26]. A tutorial on Bayesian network approaches to data mining is found in (Heckermann[56]) and they are thoroughly covered in (Jensen[62]). We will describe data mining in a relational data structure with discrete data (discrete data matrix) and the simplest generalizations to numerical (floating point) data. A recent book with good coverage of visualization integrated with models and methods is the second edition of Gelman et al[45]. Good tables of probability distributions can be found in [79, 16, 45], and a fascinating book on probability theory is Feller's [38].

These lecture notes were written for a Data Mining PhD course given annually 1996-2006. A significant update was made when it was decided to give the course at the advanced (Masters) level. The contents of these notes has been influenced by the research undertaken by me and by colleagues and PhD students I worked with. A short version of these notes was published in Wang[2, Ch 1].

1.3 Usage of probability concepts

We will stick to the conventions of applied mathematics using probability density functions, which can also be generalized in the sense of having point masses like Dirac's δ function. This function from real to real has the property that $\int_I \delta(x)\mathrm{d}x$ is one if the interval I contains zero, otherwise the integral value is zero. This gives the useful formula $\int f(x)\delta(x-y)\mathrm{d}x = f(y)$ for a smooth function $f : R \mapsto R$. The Dirac function is thus not an ordinary function, but a *generalized function*, a measure-theoretic construction which can be thought of as a limit of a sequence of functions whose supports are intervals around 0 with decreasing length going to zero, and such that each member integrates to one over the real line. We will not consider measure-theoretic complications or detailed handling of non-uniform state spaces. Thus, when we call, e.g., $q(z|x)$ a probability distribution, we mean that we have a probability distribution over the space of z that depends on x. The space of z is Euclidean n-dimensional space or a discrete finite set. In the latter case the distribution is simply a set of non-negative (probability) numbers summing to one. In the first case we may think of an integrable continuous function, possibly with discontinuities along $n-1$ dimensional sub-manifolds, and a set of singular 'peaks' 'ridges', etc, that can be modeled using the Dirac δ function. The case of state spaces consisting of unions of dissimilar spaces poses no problem except the need to visualize and understand. Such applications with variable dimension state (parameter) space are rapidly emerging, prominent examples being target tracking systems for multiple targets, and intelligent cruise control systems keeping track of nearby vehicles and obstacles.

1.4 Course Guide

Chapters and exercises marked * are either more demanding or more abstract than the main text which is easy (but long). Exercises marked + will be given as homework and their solutions are not supplied here. It is important to have access to a computer with Matlab, Octave, or R. Many of the methods described here can be found in publicly available code libraries (instead of indexing this moving target, I suggest you Google up the codes you need, or go through the Matlab central file exchange where the codes are quality reviewed by other users).

The most popular supplementary readings have been Jaynes[61], (particularly chapters 1, 2,4 and 5) for the philosophically minded, and [45] for those wanting thorough competence in practical application of Bayesian statistical tools and methods.

2 Schools of statistics

Statistical inference has a long history, and one should not assume that all scientists and engineers analyzing data have the same expertise and would reach the same type of conclusion using the objectively 'right' method in the analysis of a given data set. On the contrary, analysis proceeds by a trial-and-error process influenced by developments in the particular science or application community as well as in basic statistical sciences. A lot of science could once be developed using very little statistical theory, but this is no longer true. Probability theory is the basis of statistics, and it links a probability model to an outcome. But this linking can be achieved by a number of different principles. A pure mathematician interested in *mathematical probability* would only consider abstract spaces equipped with a probability measure. Whatever is obtained by analyzing such mathematical structures has no immediate bearing on how we should interpret a given data set collected to give us knowledge about the world – it is given that bearing only through interpretation linking it to human experience.

When it comes to inference about real-world phenomena, there are at least two different and complementary views on probability that have competed for the position as 'the' statistical method. With both views, we consider *models* that tell how data are generated in terms of probability. The models used for analysis reflect our - or the application owners - understanding of the problem area. In a sense they are *hypotheses* and in inference a hypothesis is often more or less equated with a probability model. With probability theory we can find the probability of observing data with specified characteristics under a probability model. But inference is concerned with saying something about which probability model generated our data - for this reason inference was sometimes called *inverse probability*[29]. The name Bayesian, referring to inference of the 19th century style of inverse probability, is however more recent[39].

2.0.1 *Kolmogorov's axioms

The Komogorov axioms are often referred to in a ceremonial way in parts of the engineering literature. They are not based on probability density functions which are mostly used in applied work. Instead, the concept of *probability space* is defined as a triple (Ω, F, P), where Ω is the outcome space, F is a σ-algebra of subsets (events) of Ω, and P is a real-valued function on F satisfying the three Kolmogorov axioms:

- P has non-negative values on F;
- $P(\Omega) = 1$;
- For every finite or countable sequence of disjoint sets(events) $E_i \in F$ we must have $\sum_i P(E_i) = P(\bigcup_i E_i)$.

The σ-algebra mentioned above is a set F of subsets of Ω such that $\Omega \in F$, if $f \in F$ then $\Omega - f \in F$, and if E_i is a finite or countable sequence of elements in F, then $\bigcup_i E_i$ is also in F. It is not difficult to see that a smooth probability density function defined on, e.g., R^n, will give rise to a probability function on smooth subsets (obtained by integration) that satisfies the Kolmogorov axioms. This more sophisticated approach to probability is sounder from a pure mathematics point of view. Engineers have a tendency to work with generalized probability density functions, which also give right results as long as they are manipulated in a standard fashion. In these notes we use the abbreviation *pdf* meaning probability density function (for continuous domains) or probability distribution (for discrete domains).

Consider, e.g., a variable that with probability $1/2$ is uniformly distributed over the unit interval, and with probability $1/2$ has the exact value $1/4$. The appropriate probability function P for this situation assigns to a subset r of the unit interval the probability that is obtained by taking the interval length $|r|$ and, if $1/4$ is in r, adding 1, and then halving the result. An engineer might describe this situation with a density function $p(x) = 0.5(1 + \delta(x - 0.25))$ on $0 \leq x \leq 1$, where δ is the Dirac function. Since the integral of $\delta(x - 0.25)$ over a set is one if the set contains 0.25 and otherwise zero, these two definition methods give the same probability for all 'nice' sets. The reason that Kolmogorov's axioms define P only on the σ-algebra is that there are weird subsets of most dense sets, like the real unit line, whose probability cannot be determined – they are non-measurable. Such sets do not pop up in normal applications. One common example of an unmeasurable set is the Cantor set: Start out with the closed unit interval, remove the middle third open interval which results in two closed intervals. Recursively perform the same operation on all closed intervals appearing during this construction process. The resulting Cantor set contains no closed interval but is also non-denumerable and non-measurable.

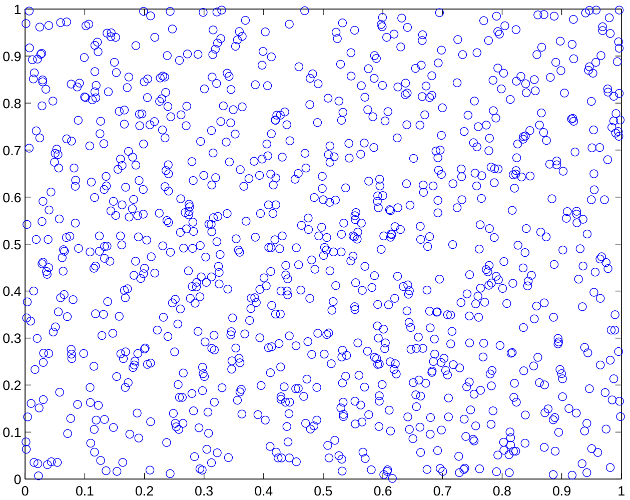

Figure 1: A sample of points distributed uniformly in the unit square. It is easy to find areas with apparently lower density of points. But there is no structure in the underlying distribution.

Figure 2: Florence Nightingale (1820-1910) is perhaps best known as the founder of the medical nursing profession. She was also an able administrator and gifted mathematician, and pioneered the use of graphical statistical summaries to identify areas for improvement in military and civilian health care, and public health. This is an early example of systematic data mining and visualization used to influence decision makers to improve operations. She is recognized as inventor of circular statistical diagrams (visualizing, e.g., seasonal variations) and pie charts.

Figure 3: Andrei Kolmogorov(1903-1987) is the mathematician best known for shaping probability theory into a modern axiomatized theory. His axioms of probability tells how probability measures are defined, also on infinite and infinite-dimensional event spaces. He did not work on statistical inference about the real world, but many of the analytical tools he developed are useful for such work.

2.1 Bayesian Inference

Figure 4: Thomas Bayes, (1701?-1762), was a Presbyterian clergyman and amateur mathematician, but he was also a member of the Royal Society. It is believed that he was admitted because he participated in a discussion on the rationality of the new, Newton type, mathematics when it was attacked by the bishop Berkeley[14]. Bayes was also the first to analyze a way to infer a probability model from observations, although his essay is technically weak by modern standards. It is not verified that this contemporary portrait of a priest actually is a likeness of Thomas Bayes, but it is often used as such.

2.1.1 Very short introduction to Bayesian inference

The first applications of inference used *Bayesian analysis*[29], where we can directly talk about the probability that a hypothesis in the form of a probability model generated our observed data. The basic mechanism is easy: There is an *observation space* of possible observations, and a *state space* of possible states (of the world, or of significant aspects of the world). Let our models be defined by their data generating probabilities, i.e, for every state there is a probability distribution over the possible observations. The *likelihood* is obtained by regarding probability of a given observation as a function of the state. The likelihood is not itself a probability density function, since it is not normalized. It is sometimes convenient to think of the likelihood (after normalization) as a probability distribution. The *prior* is a probability distribution over state, representing belief in different possible states. When an observation has been obtained, it will change our belief in the system's state from the prior to a new probability distribution, the *posterior.*

As a first example, let us look at the simplest case where both spaces have exactly two elements. This is quite common in many cases where an observation, e.g., on the reaction of a test kit (outcomes are often called positive and negative), is used to asses a persons state wrt a disease. These tests are made in such a way that the test outcome is normally correct, positive if you have the disease and negative when you do not have the disease. But there are possibilities of both false negatives (where you actually have the disease in spite of a negative test) and false positives, where the test gives a 'false alarm'. Assume the probabilities of having and not having the disease before taking the test are (p_d, p_{nd}), where $p_d + p_{nd} = 1$. Also, for test outcome o the probabilities of getting outcome o with the disease and without the disease are (p^o_d, p^o_{nd}), we form the vector $(p_d p^o_d, p_{nd} p^o_{nd})$. This is not a probability vector, but it becomes one, (p'_d, p'_{nd}) after normalization, i.e., by dividing each component by $p_d p^o_d + p_{nd} p^o_{nd}$. The Bayesian interpretation (which we shortly will motivate in more detail) is that the prior probability p_d of having the disease is transformed to the posterior probability of having the disease, p'_d, after obtaining the observation o.

In a slightly more complex case the state space is infinite, e.g. the set of real numbers. Here the prior is a probability density $f(x)$, a non-negative function that integrates to one over the real line. For a continuous observation space, like the real numbers, we have instead a density function $f(x, o)$ where for each fixed x, $f(x, o)$ is a density over the reals. In this case we come from the prior, $f(x)$, to the posterior, $f'(x)$, by multiplying functions pointwise in x, $f'(x) = cf(x)f(x, o)$ for observation o. Here c is a normalization constant that is chosen so that $f'(x)$ integrates to one over the real line. If now the state x is the true length of an object and o is the estimated length of the object, then the prior $f(x)$ can be a normal distribution with mean μ and variance σ^2, and the measurement error can be another normal distribution with mean 0 and variance σ^2_m, thus with the density $c_m \exp(-\frac{(x-o)^2}{2\sigma^2_m})$, where the normalization constant c_m is $\frac{1}{\sqrt{2\pi}\sigma_m}$. Multiplying this with the prior $c \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ and normalizing we get another normal distribution with mean $\frac{\sigma^2_m \mu + \sigma^2 o}{\sigma^2_m + \sigma^2}$ and variance $((\sigma^2_m)^{-1} + (\sigma^2)^{-1})^{-1}$. The mean of the posterior is thus a weighted average of the prior mean and the measurement. If the measurement has small variance, this mean is close to the measurement, if the variance of the prior is small then it is close to the prior mean. If a another and independent measurement is made, our new belief in the objects length is handled similarly, with the old posterior taking the place of the new prior. Since multiplication of functions is associative and commutative, it does not matter in which order a set of such independent measurements are obtained or grouped, as long as each measurement is included exactly once. Handling of discrete ad continuous state spaces are thus analogous in Bayesian inference, and we will sometimes use the same notation in both cases. In this example, if two measurements are made giving o_1 and o_2, the posterior is $f(x|o_1, o_2) = f(o_1|x)f(o_2|x)f(x)$ under the assumptions that the measurements are independent. One significant source of dependence is a systematic error affecting both o_1 and o_2.

The notation used below is quite compact and maybe gives a simple picture, but you should observe that when actual probability distributions are substituted into the formulas one usually has a quite demanding analysis to make, analytical, numeric, or stochastic. Examples are given in exercises and in Chapter 3.

2.1.2 Choosing between two alternatives

With two models H_1 and H_2, the probability of seeing data D is $P(D|H_1)$ and $P(D|H_2)$, respectively. Using the standard definition of conditional probability, $P(A|B) = P(AB)/P(B)$, we can invert the conditional of the model:

$$P(H_1|D) = P(D|H_1)P(H_1)/P(D).$$

The data probability $P(D)$ regardless of generating model is not unexpectedly difficult to assess and cannot be used. For this reason, Bayesian analysis does not normally consider a single hypothesis, but is used to compare different models in such a way that the data probability regardless of model is not needed. Using the same equation for H_2, we can eliminate the data probability:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)}\frac{P(H_1)}{P(H_2)} \tag{1}$$

This rule says that the odds we assign to the choice between H_1 and H_2, the *prior odds* $P(H_1)/P(H_2)$, is changed to the *posterior odds* $P(H_1|D)/P(H_2|D)$, the odds after also seeing the data, by multiplication with the *Bayes factor* $P(D|H_1)/P(D|H_2)$. In other words, the Bayes factor contains all information provided by the data relevant for choosing between the two hypotheses. The rule assumes that we have *subjective probability*, dependent on information the observer holds, *e.g.*, by having seen the outcome D of an experiment.

If we assume that exactly one of the hypotheses is true, we can normalize probabilities by $P(H_1|D) + P(H_2|D) = 1$ and find the unknown data probability $P(D) = P(D|H_1)P(H_1) + P(D|H_2)P(H_2)$. This probability has however no obvious meaning in an application, it is merely an inverse normalization constant.

2.1.3 Finite set of alternatives

If we have more than two hypotheses $\{H_i\}_{i\in I}$ for finite set I, a similar calculation leads to a formula defining a posterior probability distribution over the hypotheses that depends on the prior distribution:

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_{j\in I} P(D|H_j)P(H_j)} \tag{2}$$

The assumption behind the equation is that exactly one of the hypotheses is true. Then the inverse normalization constant and the data probability is $P(D) = \sum_{j\in I} P(D|H_j)P(H_j)$.

When the number of hypotheses $|I|$ goes to infinity, we can intuitively see the prior and the resulting posterior as parameterized probability density functions. Conversely, one way to solve the continuous inference problem is to approximate it by a finite problem using equation (2).

2.1.4 Infinite sets of alternatives and parameterized models

If we have a *parameterized model* H_λ, the data probability is also dependent on the value of the parameter λ which is taken from some *parameter space* or *possible world space* Λ. In Bayesian analysis the parameter space can be overwhelmingly complex, such as in the case of image reconstruction where Λ is the set of possible images represented, e.g., by voxel vectors of significant length in 3D images. But the parameter space can also be simple, like a postulated 'true value' of a measured real valued quantity. We can also let the parameter space be a finite set and so can regard equation (2) as a special case of equation (3) below.

There are also interesting cases where the parameter space has complex structure. Typically, a parameter value in such cases is a vector whose dimension is not fixed. Such models can describe a multiple target problem in defense applications, where one wants to make inference both about the number of incoming targets and the state (position and velocity) of each, or in genetic studies where one assumes that a disease is determined by combinations of an unknown number of gene variants, or simply in an effort to describe a probability distribution as a mixture of an unknown number of normal distributions. In the latter case, the parameter can be the number of such normal components followed by a sequence of weights, mean values and variances of the components, the sequence $(\lambda_1, \mu_1, \sigma_1^2, \lambda_2, \ldots)$ of equation (25).

Consideration of parameterized models leads to the parameter inference rule, which can be explained as a limit form of equation (2):

$$f(\lambda|D) \propto P(D|H_\lambda, \lambda) f(\lambda), \tag{3}$$

where $f(\lambda)$ is the prior density and $f(\lambda|D)$ is the posterior density, and $\propto$ is a sign that indicates that a normalization constant (independent of λ) has been omitted. For posteriors of parameter values the concept of *credible set* is important. A credible set is a set of parameter values in which the parameter has a high probability of lying, according to the posterior distribution. For a real-valued parameter, a credible set which is a (possibly open-ended) interval is called a *credible interval.* Likewise, a credible interval with posterior probability 0.95 is called a 95% credible interval. There are typically many 95% credible intervals for a given posterior. It may seem natural to choose that interval for which the posterior probability density is larger in the interval than outside it. This interval is not robust to rescaling, however. A robust credible interval is one that is defined by the quantiles at its end-points. It is usually centered in the sense that the probability is the same for the posterior to lie above as below the interval. In the case of a 95% credible interval, these two probabilities are thus 2.5% and 97.5%.

There are many situations where the 'data' part of the analysis is not precisely known, but only a probability distribution is available, perhaps obtained as a posterior in a previous inference. In this case the inference rule will be

$$f(\lambda|f(x)) = \int f(\lambda|x) f(x) \mathrm{d}x. \tag{4}$$

where $f(\lambda|x) \propto f(x|\lambda)f(\lambda)$, and $f(x)$ is a probability density function over the observation space, a *fuzzy observation.*

2.1.5 Simple, parameterized and composite models

It is common in statistical work to refer to probability distributions generating experimental results as *models*. We can distinguish three types of models throughout this course:

- A *simple model* is a model like $P(D|H)$. Its data generating distribution is fixed. Simple models can be used in inference work as stated in equations (1,2).
- A *parameterized model* is a probability distribution that depends on one or more parameters like $P(D|H_\lambda, \lambda)$ above. They are used to make inference on one or more of the parameters as in equation (3). This equation gives a posterior for all parameters. If the parameter is composite (consisting of several values like mean and variance for a normal distribution), we can get a posterior for one of them by integrating out the others. So if inference about a normal distribution parameter gives us the posterior as a joint pdf $f(\mu, \sigma^2)$, and we are only interested in the mean, the the posterior of the mean is the so-called *marginal distribution* $f(\mu) = \int f(\mu, \sigma^2) \mathrm{d}\sigma^2$. This process of eliminating uninteresting parameters in Bayesian inference is called integrating out *nuisance parameter*.
- A *composite model* is obtained from a parameterized model $P(D|H_\lambda, \lambda)$ by specifying a prior distribution $f(\lambda)$ for the parameter and integrating out the parameter. The resulting composite hypothesis H_c has the data probability distribution $P(D|H_c) = \int P(D|H_\lambda, \lambda) f(\lambda) \mathrm{d}\lambda$

An example of a composite model is given in section 2.1.10 to describe an unbalanced coin. Since we do not know exactly how the coin is unbalanced, we average over all possible unbalanced coins, assuming a uniform 'probability of heads' distribution.

2.1.6 Recursive inference

Bayesian analysis for many observed quantities can be formulated recursively, because the posterior obtained after one observation can be used as a prior for the next observation, so called *recursive inference*:

$$\begin{aligned} f(\lambda|D_t) &\propto f(d_t|\lambda) f(\lambda|D_{t-1}) \qquad (5) \\ f(\lambda|D_0) &= f(\lambda) \end{aligned}$$

Here $D_t = (d_1, \ldots, d_t)$ is the sequence of observations obtained at different times, and we have assumed they are independently generated, $f(D_t|\lambda) = \Pi_{i=1}^t f(d_i|\lambda)$.

The analysis of a sequence of data items, each obtained by an identical procedure is fundamental in statistics. If one wants to be very careful, one uses the concept *exchangeability*: see below.

Exercise 1 *Show that (5) follows from (3) if the d_i are independently drawn from a common distribution*

2.1.7 *Exchangeability

A term occurring frequently in theoretically oriented statistical papers is *exchangeability.* It originates in work by de Finetti, but has been reinvented many times. Assume the data obtained from an experiment is a sequence of items, all of the same type. If the only thing we know about its distribution is that all permutations of the sequence have the same probability (or probability density), then the distribution is exchangeable. This seems to be the weakest condition resulting in what is colloquially referred to as a sequence of independent and identically distributed (*iid*) observations. Exchangeability is often considered a weaker assumption than iid, but the difference is subtle. The representation theorem, due originally to de Finetti([30]), states that an exchangeable sequence is always describable as a sequence of independent variables generated according to some, typically unknown, distribution. The 0-1 representation theorem of de Finetti says that an exchangeable distribution over 0-1 (i.e., binary) sequences has a probability distribution obtainable by first specifying a 'success probability' p from some distribution over probability (i.e., reals from 0 to 1) and then considering the sequence as generated by coin tossing with success probability p.

For a fuller account of representation theorems and their proofs, see, e.g., [16].

2.1.8 Dynamic inference

The recursive formulation (5) is the basis for the Chapman-Kolmogoroff approach where the task is to track a system state that changes with time, having state λ_t at discrete time t (t is thus an integer). This case is called *dynamic inference*, since it tries to estimate the state of a moving target at times indexed from 1 upwards. In this case the sequence of observations is not exchangeable. The Chapman-Kolmogoroff equation is:

$$\begin{aligned} f(\lambda_t|D_t) &\propto f(d_t|\lambda_t)\int f(\lambda_t|\lambda_{t-1})f(\lambda_{t-1}|D_{t-1})\mathrm{d}\lambda_{t-1} \qquad (6)\\ f(\lambda_0|D_0) &= f(\lambda_0), \end{aligned}$$

where $f(\lambda_t|\lambda_{t-1})$ is the maneuvering (process innovation) noise assumed, often called the *transition kernel.* The latter is a pdf over state λ_t dependent on state at the previous time-step, λ_{t-1}. This distribution may depend on t, but often a stationary process is assumed where the distribution of λ_t depends on λ_{t-1} but not explicitly on t. As an example, if we happen to know that the state does not change, we would use Dirac's delta $\delta(\lambda_t - \lambda_{t-1})$ for the transition kernel and equation (6) will simplify to (5). If we constrain the model to be linear with Gaussian process and measurement noise with known covariance, equation (6) simplifies to a linear algebra equation, whose solution is the classical Kalman filter (this is pedagogically explained in [13]).

In many applications, additional information can be obtained by recomputing the past trajectory of the system. This is because later observations can give information about the earlier behavior. This is known as *retrodiction.* When the system has been observed over T steps, the posterior of the trajectory $\overline{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_T)$ is given by:

$$f(\overline{\lambda}|D_T) \propto f(\lambda_0) \prod_{t=1}^{T} f(d_t|\lambda_t) f(\lambda_t|\lambda_{t-1}) \tag{7}$$

Dynamic inference is a quite common case of inference, examples being target tracking in civil and military vehicle surveyance, spread of epidemics, climate, and more.

A continuous-time version of dynamic inference leads to study of the Focker-Planck (or Focker-Planck-Kolmogorov) equation, of which the Chapman-Kolmogorov equation is a discretization.

2.1.9 Does Bayes give us the right answer?

Above we have only explained how the Bayesian crank works. It would of course be nice to know also that we will get the right answers.

The Bayes factor (1) estimates the support given by the data to the hypotheses. Inevitably, random variation can give support to the 'wrong' hypothesis. A useful rule to apply when choosing between two hypotheses as in (1) is the following: If the Bayes factor is k in favor of H_1, then the probability of getting this factor or larger from an experiment where H_2 was the true hypothesis is less than $1/k$. For many specific hypothesis pairs, the bound is much better[87].

There is also a nice characterization of long run properties of the equation (5) that has an accessible proof in [45]: If the observations are generated by a common distribution $g(d)$ and we try to find the parameter λ by using equation (5), then as the number of observations tends to infinity, the posterior $f(\lambda|D_n)$ will concentrate around a value $\hat{\lambda}$ that minimizes the Kullback-Leibler distance between the true and the estimated distribution of observations, $\hat{\lambda} = \text{argmin}_{\lambda} KL(g(\cdot), f(\cdot|\lambda))$, where $KL(g, f) = \int f(x) \log(f(x)/g(x)) \mathrm{d}x$. It is not difficult to see that the KL distance is minimized and zero when the two distributions $f(d|\lambda)$ and $g(d)$ are equal. So if the real distribution $g(x)$ is equal to the considered distribution $f(x, \lambda)$, then the dynamic Bayesian inference will asymptotically give the right answer. Convergence rate can however be very slow, as discussed in [45].

2.1.10 A small Bayesian example.

We will see how Bayes' method works with a small example, in fact very similar to the example used by Thomas Bayes. Assume we have found a coin among the belongings of a notorious gambling shark. Is this coin fair or unfair? The data we can obtain is a sequence of outcomes in a tossing experiment, represented as a binary string D. Let one hypothesis be that the coin is fair, H_r. Then $P(D|H_r) = 2^{-n}$, where $n = |D|$ is the number of tosses made. Since the tosses are assumed independent, the number of ones, s, and the number of zeros, f, completely characterizes an experiment. Since every outcome has the same probability, we can not evaluate the experiment with respect to only H_r, but we introduce another hypothesis that can fit better or worse to an outcome. Bayes used a parameterized model where the parameter is the unknown probability, p, of getting a one in a toss. For this model H_p, we have $P(D|H_p) = p^s(1-p)^f$, and the probability of an outcome is clearly a function of p. We can consider H_p a whole family of models for $0 \leq p \leq 1$. If we assume, with Bayes, that the prior distribution of p is uniform in the interval from 0 to 1, we get a posterior distribution equal to the normalized likelihood, $f(p|D) = cp^s(1-p)^f$, a Beta distribution where the normalization constant is (as can be found from a table of probability distributions or proved by double induction) $c = (n+1)!/(s!f!)$. This function has a maximum at the observed frequency s/n. We cannot say that the coin is unfair just because we have $s \neq f$ since the normal variation makes inequality very much more likely than equality if we made a large number of tosses, even if the coin is fair.

If we want to decide between fairness and unfairness we can introduce a composite hypothesis/model by specifying a probability distribution for the parameter p in H_p. A conventional choice is again the uniform distribution. Let H_u be the hypothesis of unfairness, expressed as H_p with a uniform distribution on the parameter p. By integration we find $P(D|H_u) = \int_0^1 P(D|H_p)dp = \int_0^1 p^s(1-p)^f \mathrm{d}p = s!f!/(n+1)!$. Suppose now that we toss the coin twelve times and obtain the sequence `00011000001`, three successes and nine failures. The probability of this outcome D under the fairness hypothesis H_r is $P(D|H_r) = 2^{-12}$, and under the unfairness hypothesis H_u we have $P(D|H_u) = 3!9!/13!$. The Bayes factor in favor of unfairness will be

$$P(D|H_u)/P(D|H_r) = 2^n s!f!/(n+1)!$$

Inserting $n = 12$, $f = 9$ and $s = 3$ we obtain the Bayes factor $2^{12}3!9!/13! = 1.4$, slightly in favor of unfairness. But this is a too small value to be of interest. Values above 3 are worth mentioning, above 30 significant, and factors above 300 would give strong support to the first hypothesis, whereas values below 1/3, 1/30 and 1/300 give similar support to the second hypothesis. But we are only comparing two hypotheses – this scheme cannot tell us that none of the alternatives is plausible.

The above is an analytical way to solve Bayesian inference, and it relies on having priors and likelihoods such that their product can be integrated analytically. This is often not the case since several functions do not have a primitive function that can be defined in terms of functions we have named. We can obtain a simple approximate solution by approximating the continuos parameter space with a finite set, e.g., probability values spaced uniformly from 0 to 1. We then use the formula (refeq:db) as an approximation to (refeq:parm). This works well for a one-dimensional parameter space, but with many parameters the curse of dimensionality sets in and we will have to use MCMC methods.

The application of Bayes factor for the case of a composite and a simple model has been criticized for the case where the simple model is a special case of the parameterized model from which the composite is obtained. An alternative, proposed in [45], uses a high probability symmetric credible interval and chooses the simple model if the corresponding parameter value falls in the interval. Another alternative is to find the posterior probability that the parameter value is larger than the value of the simple model. If this probability is not close to one or zero, this indicates that the simple model is sufficient. This is not a pure Bayesian approach but has the flavor of hypothesis testing and p-values, as we will describe in Ch 2.2. The method is a good alternative to the Bayes factor in the case of real valued parameters like the success probability for a coin being 0.5 or a regression coefficient being zero (Ch 3.7). In one of our important applications of the Bayes' factor, the dependency test of Ch 3.2.2, two composite models are compared and despite the fact that the simpler (coarser) model (modeling independency) is a special case of the finer one, there is no easy way to define a real valued measure which can be tested for zero. In this case the Bayes factor approach is the only reasonable alternative with high credibility.

We have here compared two data generating mechanisms, out of a potentially unbounded set of possibilities. We have simply assumed that the tosses are independent and identically distributed in the sense that the success probability is constant over the experiment. It is perfectly possible to assume that there is autocorrelation in the experiment, or that the success probability drifts during the experiment. In order to investigate these possibilities we need different data generating models (and longer experiments, because there are more parameters to consider). Once credible models of those aspects of an experiment we want to consider are available, the analysis follows the same steps as those above.

Exercise 2 *A cheap test is available for a serious disease. Assume that this test is cheap in the sense of having* 5% *probability of giving negative result if you have the disease, and* 10% *probability of giving positive result if you do not have it. Moreover, in a population of individuals similar to you,* 0.1% *have the disease. How would you compute your probability of having the disease when the test gives*

a) positive result?

b) negative result?

c) Can the precision be improved by repeating the test? What assumptions are reasonable for answering this question?

d) Do there seem to be systematic errors in your analysis when applied in a practical setting?

Exercise 3 *+The assumption that an unfair coin has a uniform distribution for p is not very convincing if the coin is physical and we have had a chance to look at it without actually tossing it. Assume that we instead assign a prior distribution for p by saying that the coin is well expected to be balanced in a series of $2k$ trials. This can be formalized as stating the prior to be be proportional to $p^k(1-p)^k$ in the unbalanced case. How would such an assumption change the posterior and interpretation of the example (with three successes and nine failures)? Find the posterior probability of $p > 0.5$ in the parameterized model for the two cases uniform and Beta(k+1,k+1) distributed prior, for some suitable values of k, like 5 and 50!*

Exercise 4 *Assume that we use equation (1) to choose between two models stating respectively probability a and b for heads in a coin tossing series. Assume also that the true probability of heads is c. For which values of c will the Bayes factor in favor of a against b go to zero when the number of tosses increases?*

2.1.11 Bayesian decision theory and parameter estimation

The posterior odds in equation (1) gives a numerical measure of belief in the two hypotheses compared. Suppose our task is to decide by choosing one of them. If the Bayes factor is greater than one, H_1 is more likely than H_2 assuming no prior preference of either. But this does not necessarily mean that H_1 is true, since the data can be misleading by natural random fluctuation. Two types of

Figure 5: Pierre Simone de Laplace was a court (later Napoleon's minister of the interior) mathematician contributing to quite many problems in mathematics, fluid mechanics and astronomy. He rediscovered Bayes' method and developed it in a more mathematically mature way. He used the same uniform prior for an unknown probability as Bayes did, and used it to compute the probability that the sun will rise tomorrow, given that it has risen N times without exception. The answer, $(N+1)/(N+2)$ is obtained by using the posterior mean value estimator which in a discrete setting is known as the Laplace estimator: for a discrete distribution over d outcomes, where n_i observations of outcome i were observed, the Laplace estimate of the probability distribution $(p_1, \ldots, p_d)$ is $p_i = (n_i+1)/(n+d)$, the relative frequencies after one more observation of each outcome has been added. In this course we can also use *Laplace's parallel combination*: this term is sometimes used to describe the computation (5) for a finite Λ: Take two vectors indexed by state, multiply them component-wise and normalize. In Matlab: `tmp=likelihood.*prior; posterior=tmp/sum(tmp);`.

error are possible: Choosing H_1 when H_2 is true, and choosing H_2 when H_1 is true. Our choice must take the consequences of the two types of error into account. If it is much worse to act on H_1 when H_2 is true than vice versa, the choice of H_1 must be penalized in some way, whereas if the consequences of both error types are similar we should indeed chose the alternative of largest posterior probability. Statistical Bayesian decision theory resolves the problem by introducing a cost for each pair of choice and true hypothesis. For finite hypothesis sets, the table of these costs forms a matrix, the *cost matrix*, which is column indexed by true hypothesis and row indexed by our choice. The recipe for choosing is to make the choice with smallest expected cost[11]. This rule is applicable also when simultaneously making many model comparisons. The term *utility* is equally common in statistical decision making. It differs from cost only by a sign, and we thus strive to maximize it. The general Bayesian decision making paradigm can be captured in the prescription:

$$a* = \mathrm{argmax}_{a \in A} \int u(a, \lambda) f(\lambda | D) \mathrm{d}\lambda,$$

where $u(a, \lambda)$ is the utility of executing action a from a set of actions A in world state λ.

When making inference for the parameter value of a parameterized model, equation (3) gives only a distribution over the parameter value. If we want a point *estimate* $\hat{\lambda}$ of the parameter value, we should also use Bayesian decision theory. We want to minimize the loss incurred by stating the estimate $\hat{\lambda}$ when the true value is λ. Let this loss be given by a *loss function*[1] $L(\hat{\lambda}, \lambda)$. But we do not know λ, we only know its posterior distribution. As with a discrete set of decision alternatives we minimize the expected loss over the posterior for λ, $\int L(\hat{\lambda}, \lambda) f(\lambda|D) \mathrm{d}\lambda$. If λ is real valued and the loss function is the squared error, the optimal estimator is the mean of $f(\lambda|D)$; if the loss is the absolute value of the error, the optimal estimator is the median; with a discrete parameter space, minimizing the probability of an error gives the *Maximum A Posteriori*(*MAP*) estimate. When the parameter is continuous, MAP is the argument of highest probability density for the parameter. The estimate is often called the *mode* of the posterior distribution. Certainly, the probability of an error in this case is usually one, but the estimate gets closer and closer to the mode if we postulate a sequence of decreasing error bounds. As an example, when tossing a coin gives s heads and f tails, the posterior with a uniform prior is $f(p|s, f) = cp^s(1-p)^f$, the MAP estimate for p is the observed frequency $s/(s+f)$, the mean estimate is the *Laplace estimator* $(s+1)/(s+f+2)$ and the median is a fairly complicated quantity expressible, when s and f are known, as the solution to an algebraic equation of high degree. The Laplace estimator is often preferable to the simple observed frequency estimator, see figure 5.

To see that minimizing expected squared error leads to the mean value estimate, consider the squared error loss function $L(\lambda, \hat{\lambda}) = (\lambda - \hat{\lambda})^2$. Minimizing the expected loss wrt $\hat{\lambda}$ leads to finding a zero of $\frac{\mathrm{d}}{\mathrm{d}\hat{\lambda}} \int_\Lambda L(\lambda, \hat{\lambda}) f(\lambda|x) \mathrm{d}\lambda = -\int_\Lambda 2(\lambda - \hat{\lambda}) f(\lambda|x) \mathrm{d}\lambda = -2\bar{\lambda} + 2\hat{\lambda}$. The solution is $\hat{\lambda} = \bar{\lambda}$, and it is easily verified to give a minimum, since the second derivative with respect to $\hat{\lambda}$ is 2.

[1] Of course, it is not necessary to introduce both utility and loss since they differ only in sign. But both are used in different communities

Exercise 5 *(+) (Sivia's lighthouse example[95], modified) Light pulses are emitted horizontally from a lighthouse at angles uniformly (and iid) distributed. The lighthouse is placed at distance d from a straight coastline, and the normal from the lighthouse hits the coastline at coordinate* x_0. *Consider only light rays that actually hit the coastline (and forget those sent away from the coastline). So the emitting angles* (α_i) *are uniformly distributed on* $[-\pi/2, \pi/2]$ *and the points where they hit the coastline have coordinates* $x_i = d\tan(\alpha_i)$. *(i) Derive the probability distribution of the points* (x_i) *where the coastline is hit, and verify that it is in the family of Cauchy distributions.*
(ii) Consider estimating x_0 *from the* x_i *when d is known to be 1, using sample mean, sample median and maximum likelihood estimators. Implement the estimators and check them on samples of size 10, 100, 1000 and 10000. Some of them seem to behave unexpectly bad. Explain why, and assess the adequacy of your estimators.*
(iii) Similar problem, but design and analyze a maximum likelihood estimator for d when x_0 *is known to be 0. Optional: what would mean and median estimators look like? How good are they? (Hint: You will probably have to do this numerically)*

Exercise 6 *Verify the optimal estimator of* λ *to be*

+a) the median when the loss function is $L(\lambda, \hat{\lambda}) = |\lambda - \hat{\lambda}|$;

+b) the mode when the loss function is $L(\lambda, \hat{\lambda}) = -\delta(\lambda - \hat{\lambda})$; *where* δ *is Dirac's delta function.*

c) Which of the three estimates (mean, median, MAP) are scale-invariant when applied to a real valued variable (Scale invariance: if the variable is transformed by $u = g(x)$, *and thus the distribution to* $f(x)/g'(x)$, *then* $\hat{u} = g(\hat{x})$, *where the hat denotes one of the three estimators. Mode: estimate with highest probability or probability density).*

Exercise 7 *Find the optimal estimator from the posterior when the estimated quantity is a probability distribution and we know that this distribution will only be used for maximum expected utility decision making.*

2.1.12 *Bayesian analysis of cross-validation

The current state-of-the-art in Bayesian (and also in test-based) inference is that non-parametric models are avoided because we do not fully understand them when applied to complex multi-variate data sets. So we are coerced into using models that are not universal, and then Bayesian inference gives us a posterior for the parameters that may differ substantially from some 'true' distribution whose existence, but not form, we may postulate or which may become obvious in the future. The Bayesian model selection framework we have earlier described may be called the *model-closed perspective.* In that perspective we assume that there is a true model somewhere in the set of models we consider. If we consider the possibility of a true model which is not part of the set we analyze, we call it the *model-open perspective.* A thorough analysis of this problem area is given in [16, Ch 6.1]. We shall look at a selected part of this analysis, which is concerned with analyzing an exchangeable sequence for the purpose of making a decision. In [16, Ch 6.1] the authors also consider an in-between case, a *model-completed perspective*, where a true model is known, but not in the set of models M_i. The reason for this perspective may be that although we know the true model, it has a too complicated form to be used in decision making.

Suppose that we wish to consider a set of simple or composite models $\{M_i\}_{i\in I}$, having made a sequence of independent observations $\overline{x} = x_1, \dots, x_n$ from a process. The notation suggest that I can be a discrete set, an interval of real numbers, a set of vectors in Euclidean space filling a part of it, or a more complex object such as the set of all distributions, even if we would be hard pressed to actually use the latter. We can handle both model selection (select a value for i) and model averaging (select some probability distribution over I) by, in the latter case, extending the model set to the set of weighted averages of the original models in I, thus possibly getting a larger set of actual models than the original $\{M_i\}_{i\in I}$. Such set is the *convex closure* of the original set. Suppose we must now choose one $i \in I$ and use it as a model, and then make a decision based on i, which is followed by some type of payoff. This payoff can either be based on the accuracy of a prediction of a previously unknown value, or on the accuracy of a parameter estimate or predicted distribution. The situation can be described with a utility function $u(m_i, a_i, \omega)$, where m_i is a model selected, a_i is an action performed based on i, and ω is the response of nature - a future observation or a distribution. The objective is to select m_i and a_i in a way that optimizes $\int u(m_i, a_i, \omega)\mathrm{d}\omega$. However, when a_i is selected from some set A_i after m_i has been chosen, we must assume that $\{M_i\}$ is the true model. The second selection is thus easy, $a_i = \mathrm{argmax}_{a\in A_i} \int u(m_i, a, \omega) f_i(\omega|\overline{x})\mathrm{d}\omega$, where $f_i(\omega|\overline{x}) = P(\omega|M_i, \overline{x})$.

In the case of simple models M_i, the function $f_i(\omega|\overline{x})$ is independent of $\overline{x}$. A similar procedure solves the combined model selection and decision making problem in the model-closed perspective:

$$(m_i, a_i) = \mathrm{argmax}_{m_i^* \in M, a_i^* \in A_i} \int u(m_i^*, a_i^*, \omega) f_i(\omega)\mathrm{d}\omega. \quad (8)$$

This is little more than the formulation of inference as decision making we saw in section 2.1.11. Another way to express the solution is that the expected utility of choosing model m_i is $\int u(m_i, a_i, \omega) f(\omega|\overline{x})\mathrm{d}\omega$, where $f(\omega|\overline{x})$ is the posterior of ω given the data observed $\overline{x}$ in the model-closed perspective. In the model-open perspective we have no reason to put great confidence in the posterior, and we must use some other method to estimate $f(\omega|\overline{x})$. However, we can remove one item x_j from $\overline{x}$ and use the rest of the sequence $x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots)$ as a possible observation sequence and then use x_j as an approximation for ω. For a long sequence we can thus get many examples of an initial sequence followed by a new observation, and this yields a possible method to select models outside the standard Bayesian framework.

As an example, with the decision problem of predicting the next observation with quadratic loss, the state ω will be the next observation y, the actions a_i will be predicted values for y, and the expected utility will be $\int u(m_i, a_i, y) f(y|\overline{x}) \mathrm{d}y = -\int (y - a_i)^2 f(y|\overline{x}) \mathrm{d}y$. By cross-validation we can estimate this quantity (remember, we do not know $f(y|\overline{x})$) by $-1/(k \sum_{j=1}^{k} (x_j - a_i)^2)$. This is a kind of Monte Carlo estimation of the utility of model i, and in Bayesian cross-validation one tries a number of models and selects one with best estimated utility. For other types of decisions, like the classical machine learning task of predicting a future variable given partial information about it, the procedure can be easily adapted. We have not here gone into detail concerning reliability estimates. Such estimates are rather infrequent in the literature, and one is usually content with experimental validations of the method selected for a particular application, the reason (i.e., the problem) being that the model-open perspective admits any model as possible.

2.1.13 Biases and analyzing protocols

We will see in the chapter on inference by testing (frequentism), exemplified by the coin-tossing example, that a frequentist analysis must consider the protocol of an experiment, not just the outcome. However, there are also important factors of experimental protocols that can influence both a Bayesian and a frequentist analysis.

In investigations based on questionnaires, the detailed formulation of questions and the administration of the questionnaire has significant effect on what subjects answer. If such investigations touch very sensitive parts of society or the individual, they have little reliability unless their protocols are very well documented, designed and validated. When asked about 'antisocial' and politically incorrect or even illegal habits, subjects' actual trust in the anonymity offered can be crucial, and work in unexpected ways. When asked about circumstances that can lead to changes in legislation or public surveillance, subjects can also give the answer that tends to turn legislation their way. The intents of the actors creep back into the protocol when they are regarded as subjects.

In medical 'meta-analyses', many different investigations are pooled from publications to give better statistical strength. Pitfalls in this might be that the investigations are not quite comparable, and that investigations not leading to significant conclusions might not even be published. Such investigations may also be published in less prestigouos ways and fail to be found by the meta-analyst. This phenomenon is known as *publication bias*.

In clinical trials, where a new treatment is compared with the currently standard one, biases can be avoided by randomizing the two groups (i.,e., patients are not given a choice) and if possible hiding the choice even for health care personnel involved. This can be shown to eliminate completely the commonly observed biases like those described below in section 3.3.2, but there are still possible problems in the procedure by which patients are included and excluded from the study, particularly in that all subjects(participants) in a trial must nowadays give informed consent. This can for example bias the population in a study with potential connection to psychiatric conditions.

Exercise 8 *In questionaires with sensitive questions is is difficult to assess to which extent the answers are true, maybe because subjects fear that their anonymity will not be respected. Suggest some reasons why subjects can give deliberately wrong answers in investigations of their attitudes. One way to tackle this is to ask subjects to give wrong answers with some probability (and governed, e.g., by throwing coins or dice while answering). Assume we want to assess the frequency of some legally or morally inapproppriate behaviour and expect that a 'yes' answer is not obtained when it should with probability* p_0*, but 'no' answers are correct. Assume also that we ask subjects to give the wrong answer with probability* p_1*. When does this procedure give more accurate estimates, assuming the answers will be truthful in the second case? (Hint: use a MATLAB simulation).*

Exercise 9 *In a famous TV quiz show, the contestant has to guess in which of three boxes a valuable prize can be found and obtained. After the contestant's first selection of a box, the host sometimes opens one unselected box and shows that it is empty. The contestant is given the opportunity to switch his guess to the unselected remaining box. It is a popular quiz in probability courses to prove that this switch is indeed advantageous and the offer should be accepted.*

Analyze and advice on the decision under the following assumptions:

i) The host is forced by the protocol to show one empty box and allow the contestant to change selection.

ii) (+) The opening is voluntary for the host, and he will try to make you fail.

iii) (+) The opening is voluntary for the host, and he will try to make you win

iv) (+) The opening is voluntary for the host, and you do not know the desire, if any, of the host.(Hint: You should assume that the host is as clever as you are. Is there a randomized optimal choice which performs well regardless of the intent of the host?).

2.1.14 How to perform Bayesian inversion

Development of Bayesian analysis was only enabled when desk-top computing power was made available. Before computers were standard tools, Bayesian analysis was completely dependent on formulating the inference problem using equation (1), or (2) with a small number of alternatives, and various analytical models that happened to be tractable because of conjugacy properties for the equations (3), (5) and (6). We have already mentioned the Kalman filter, where assuming all distributions Gaussian transforms the solution of (6) into a linear algebra problem. Several such special cases exist, and the corresponding families of functions with their properties are listed, e.g., in [16, App A] and [45]. In our analysis of graphical models (also known as influence diagrams or Bayesian Networks) we will thoroughly analyze one such example, the use of Dirichlet distributions to make inference about discrete probability distributions. For inference about low-dimensional parameter spaces, it is possible to discretize parameter space and reduce the problems (3, 5, 6) to the discrete problem (2), either using some sophistication with numerical analysis tools or just assuming that both likelihoods and priors are piecewise constant within rectilinear boxes. Then we work exclusively with probabilities that parameters fall into specific boxes. We can thus work with distributions over discrete (but large) sets.

Figure 6: Andrei Markov (1856-1922) was a Russian mathematician developing, among other things, the theory of Markov processes. Markov assumptions are common assumptions in statistical modeling saying that a variable's distribution, conditional on the other variable's values, depends effectively only on a subset of those variables. Examples of Markov assumptions are found in Markov chains, where only the current state and not the full history of the process decides the probability distribution of the next state, and in graphical models, where the immediate ancestors (neighbors for undirected models) in a graph give all available information (among ancestors in the case of directed models) about a variable.

In general, however, the parameter and observation spaces of interest are very high-dimensional (in genetic linkage studies and imaging, the parameter dimension can be a couple of thousand and many millions, respectively). In these cases Markov Chain Monte Carlo (MCMC) is the only presently feasible and readily implementable technique (in certain key applications, however, other very specialized numerical procedures have been implemented, as in military tracking systems). The MCMC method assumes that a probability density function on a continuous or discrete space is given in unnormalized form, $\pi(x) = cf(x)$, where we can evaluate f but not c or π. A Markov chain is constructed in such a way that it has a unique stationary distribution which is π. This means that a long chain $x_1, \ldots, x_N$ can be used as an approximation to the distribution, and the average of a function $g(x)$ over π can be approximated by the sum $\sum_{i=1}^{N} g(x_i)/N$. The details will be given in Ch. 4. The chain is designed with a proposal distribution $q(z|x)$ which is used to generate a next proposed state z when the last state of the chain is x. So the chain $x_1, \ldots, x_t$ is extended by drawing a proposed new state z from the distribution $q(z|x_t)$, and this proposal is evaluated by computing the ratio

$$r = \frac{f(z)q(x_t|z)}{f(x_t)q(z|x_t)}. \tag{9}$$

If $r > 1$, the proposal is accepted. Otherwise the proposal is accepted with probability r, and rejected otherwise. If the proposal is accepted, x_{t+1} is set to z, otherwise x_{t+1} is set to x_t. This can be implemented by comparing a standard uniformly distributed pseudo random u number with r and accepting the proposed state if $r > u$. Equation (9) shows that a move is preferred to the extent that it would increases the probability of the state but also if the proposal distribution makes it easy to return to the old state. A proposal distribution with $q(z|x) = q(x|z)$ is called a *symmetric proposal* distribution. Symmetric proposals are the most common, and they simplify (9) a lot.

We can illustrate the MCMC method with a small example. Assume a distribution consisting of two different normal distributions, so that with probability about 0.5 each is selected as the distribution of a new item (data point). Figure 7 shows a distribution with two high- probability regions (top frame). Then three traces with different proposal distributions are shown. In the first case with a small proposal step, the trace has difficulty switching between the two high-probability regions, and therefore the estimate of their relative weights will be of low accuracy. The second trace mixes well and gives an excellent estimate. In the last trace with a large step size, most of the proposals are rejected because they fall in low probability regions. Therefore the estimate becomes inaccurate - the trace has too many repeated values and looks like an estimate of a distribution over a finite set of values.

It is a significant problem in MCMC to determine the proposal distribution so that good mixing is guaranteed. Often, experiments are required before a good proposal distribution is found.

An alternative method, *variational Bayes method* is gaining increasing attention as an alternative to the computer intensive MCMC method. Instead of computing an intractable posterior over many parameters, one tries to find factored posteriors where each factor depends on only one or a few parameters[7].

Exercise 10 *The file* `x33.mat` *contains three traces with different proposal distributions similar to those in figure 7). Assume that the target distribution is a mixture of two normal distributions. Estimate the mixing coefficients, mean and variance for the two distributions using each of the three traces.*

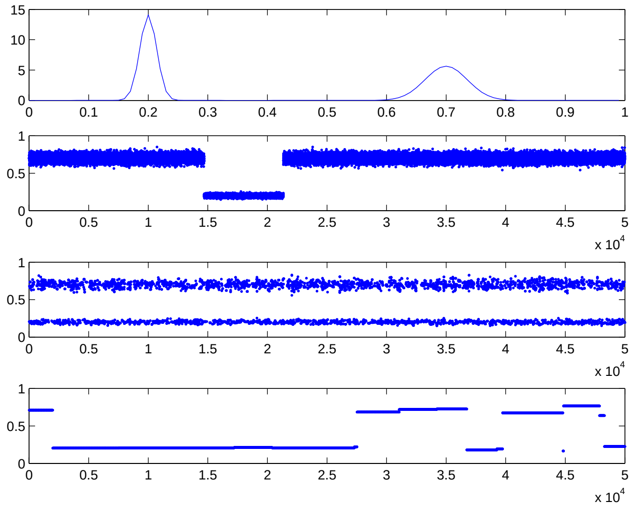

Figure 7: MCMC computation. From top: a) Density function is a mixture of two normals of equal weight. b) MCMC trace with small proposal size - too few jumps between peaks. c) Well chosen proposal. Mixes well. d) Too large proposal step: most proposals rejected.

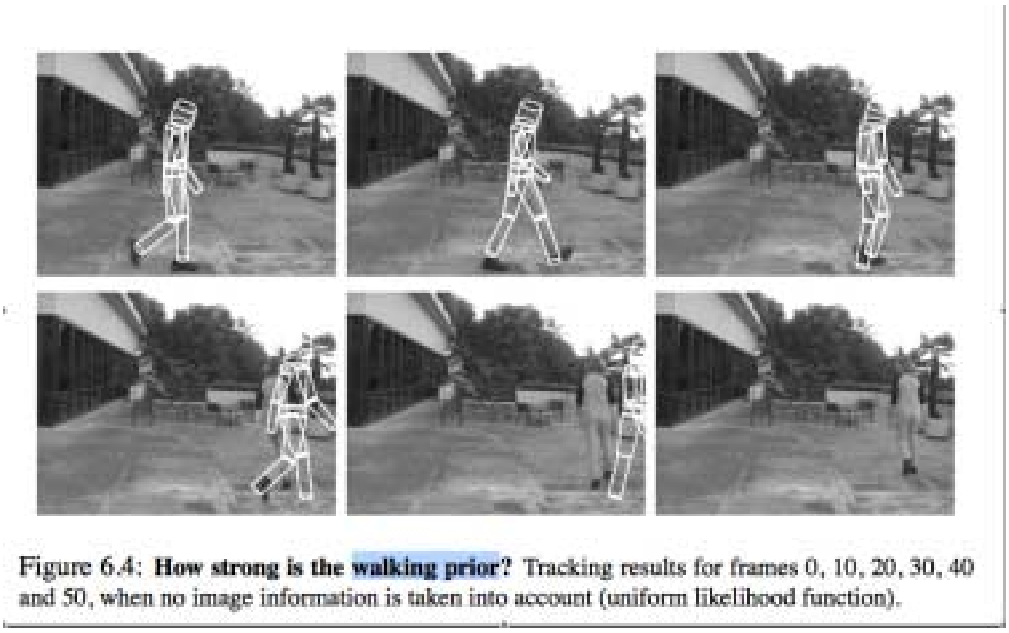

Figure 6.4: **How strong is the walking prior?** Tracking results for frames 0, 10, 20, 30, 40 and 50, when no image information is taken into account (uniform likelihood function).

Figure 8: The MCMC method has been developed to solve dynamic inference problems as defined by equation (6). This method is often called a particle filter. In this example, from the Thesis of Hedvig Kjellström[94], a person walking in a cluttered scene is tracked by a model built from cylinders. In each video frame, several copies of the current best matching models are forwarded one time step using the innovation process. Then all these forwarded models are scored by matching its contours with edges in the next video frame, forming a weighted sample. By resampling, a new unweighted sample is obtained, and the tracking process starts anew. In the example, the innovation process was not wide enough to prevent the subject from escaping the tracker. More on this method in section 4.1.2.

2.2 Test based inference

It is completely feasible to reject the idea of subjective probability, in fact, this is the more common approach in many areas of science. If a coin is balanced, the probability of heads is 0.5, and in the long run it will land heads about half the number of times it was tossed. However, it might not, but even then one can claim that the probability is still the same. In other words, we can claim that the updating of prior to posterior through the likelihood has nothing to do with the 'objective' probability of the coin landing heads.

Figure 9: Antoine Augustine Cournot (1801–1877). He formulated 'Cournot's bridge' or *Cournot's principle* as a link between mathematical probability and the real world: An event, picked in advance, with very small or zero probability, will not happen. If we observe such an event, we can conclude that one (or more) of the assumptions made to find its probability is incorrect. In hypothesis testing we apply Cournot's bridge to a case where the null hypothesis is the only questionable assumption we made in computing the p-value – thus we can reject the null hypothesis if we observe a low p-value. The principle was considered of fundamental importance by several pioneers of probability and inference, but it does not figure prominently in current historic surveys. It was recently brought to attention (for another purpose than ours) by Glenn Shafer[93].

The perceived irrelevance of long run properties of hypothesis probabilities made one school of statistics reject subjective probability altogether. This school works with what is usually known as *objective probability* or *frequentist statistics*. Data is generated in repeatable experiments with a fixed distribution of the outcome. Since we cannot talk about the objective probability of a hypothesis, we can not use such probabilities to express our belief in them. In

photo A Barrington-Brown (c) R A Fisher Memorial Trust.

Figure 10: R A Fisher (1890–1962) was a leading statistician as well as geneticist. In his paper Inverse probability[40], he rejected Bayesian analysis on grounds of its dependency on priors and scaling. He launched an alternative concept, 'fiducial analysis'. Although this concept was not developed after Fishers time, possibly because of erroneous or at least highly questionable statements in his second fiducial paper[41], the standard definition of confidence intervals (developed by J Neyman) has a similar flavor. The fiducial argument was apparently the starting point for Dempster in developing evidence theory[32].

the development of probability, many intriguing developments occurred that are now not visible in the standard engineering statistics and probability, and a very readable account can be seen in the recent book [93]. In particular, Cournot hypothesized that the connection between statistics and the real world is that an event (picked in advance) with probability zero will not occur. In the world of finite time and resources, this translates to the thesis that an event (also picked in advance of testing) with small probability will usually not happen. This is the basis for the modern hypothesis-testing framework that is standard in many applied disciplines, for example, in clinical testing.

One device used by a practitioner of objective probability is testing. For a single hypothesis H, a *test statistic* is designed as a mapping t of the possible outcomes to an ordered space, normally the real numbers. The data probability function $P(D|H)$ will now induce a distribution of the test statistic on the real line. I continue by defining a rejection region, an interval with low probability, typically 5% or 1%. Next the experiment is performed or the data D is obtained, and if the test statistic $t(D)$ falls in the rejection region, the hypothesis H is rejected. For a parameterized hypothesis, rejection depends on the value of the parameter. In objective probability inference about real valued parameters we use the concept of a *confidence interval*, whose definition is unfortunately rather awkward and is omitted. It is discussed in all elementary statistics texts, but often the students leave introductory statistics courses believing that the confidence interval is a credible interval. Fortunately, this does not matter a lot, since numerically they are practically the same. The central idea in testing is thus to reject models. If we have a problem of choosing between two hypotheses, one is singled out as the tested hypothesis, the *null hypothesis*, and this one is tested. The test statistic is chosen to make the probability of rejecting the null hypothesis maximal if data is in fact obtained through the alternative hypothesis. Unfortunately, there is no strong reason to accept the null hypothesis just because it could not be rejected, and there is no strong reason to accept the alternative just because the null was rejected. But this is how testing is usually applied.

The *p-value* is an important concept in testing. This value is used when the rejection region is an infinite interval on one side of the real line, say the left one. The p-value is the probability of obtaining a test statistic not larger than the one obtained, under the null hypothesis, so that a p-value less than 0.01 allows one to reject the null hypothesis on the 1% level – the observed value is less than it could plausibly be. We can define and compute p-values also for ordered sets of rejection regions that are not half-infinite intervals on the real line, but caution must be exercised. For the observation space D we must, before obtaining or at least before looking at the data to be tested, define a dense or discrete ordered set of subsets R_t of D such that $R_s \subset R_t$ whenever $t < s$. The p-value for an observation d with respect to the null hypothesis and the rejection sets is now $\sup\{P(d' \in R_t) : d \in R_t\}$. In other words, if we compute the probability under the null hypothesis for the observation to fall in R_t for all t and call this probability p_t, then the p-value is the upper bound of p_t for the sets R_t which contain the observation.

2.2.1 A small hypothesis testing example

Let us analyze coin tossing again. We have again the two hypotheses H_f and H_u. Choose H_f, the coin is fair, as the null hypothesis. Choose the number of successes as test statistic. Under the null hypothesis we can easily compute the p-value, the probability of obtaining nine or more failures with a fair coin tossed twelve times, $\sum_{i=9}^{12} \binom{12}{i} 2^{-12} = .075$. This is 7.5%, so the experiment does not allow us to reject fairness at the 5% level. On the other hand, if the testing plan was to toss the coin until three successes have been seen, the p value should be computed as the probability of seeing nine or more failures before the third success: $\sum_{i=9}^{\infty} \binom{i+2}{i} 2^{-(i+3)} = .0325$. Since this is 3.25%, we can now reject the fairness hypothesis at the 5% level. This is a feature of using test statistics: the result depends not only on the choice of null hypothesis and significance level, but also on the experimental design, *i.e.* on data we did not see but that could have been seen.

If we chose H_u as the null hypothesis we would put the rejection region in the center, around $f = s$, since in this area H_u has smaller probability than H_f. In many cases we will be able to reject either both of H_f and H_u, or neither of them. This shows the importance of the choice of null hypothesis, and also a certain amount of subjectivity in inference by testing.

2.2.2 Multiple testing considerations

Modern analyses of large data sets involve making many investigations, and many tests if a testing approach is chosen. This means that some null hypotheses will be rejected despite being 'true'. If we test 10000 null hypotheses on the 5% level, 500 will be rejected on the average even if they are all true. Correcting for multiple testing is necessary. This correction is an analog of the bias introduced in Bayesian model choice when the two types of error have different cost.

There are basically two types of error control proposed for this type of analysis: In *family-wise error control* (*FWE*[57]), one estimates the probability of finding at least one erroneous rejection if there is one, whereas in the recently analyzed *false discovery rate control* (*FDR*[9]) one is concerned with the fraction of the rejected hypotheses that is erroneously rejected.

A *Bonferroni correction* divides the desired significance, say 5%, with the number of tests made. So in the example of 10000 tests, the rejection level should be decreased from 5% to 0.0005%. Naturally, this means a very significant loss in power, and most of the 'false' null hypotheses will probably not be rejected. It was shown by Hochberg[57] that one could equally well truncate the rejection list at element k where $k = \max\{i : p_i \leq q/(m - i + 1)\}$, $(p_i)_1^m$ is the increasing list of p-values and q is the desired FWE rate.

A recent proposal is the control of false discovery rate(FDR). Here we are only concerned that the rate(fraction) of false rejections is below a given level. If this rate is set to 5%, it means that of the rejected null hypotheses, on the average no more than 5% are falsely rejected. It was shown in [9] that if the m tests are independent, one should truncate the rejection list of m sorted p-values at element k where $k = \max\{i : p_i \leq qi/m\}$. The correct interpretation of the k so found is: The expected number of the k tests giving p-values $p_1, \ldots, p_k$ that are falsely rejected is kq.

If we do not know how the tests are correlated, it was shown in [10] that the cut-off value is safe for FDR control regardless of dependencies if it is changed from qi/m to $qi/(mH_m)$, where $H_m = \sum_{i=1}^{m} 1/i$, the mth harmonic number. A more surprising (and more difficult to prove) result is that, for many types of positively correlated tests, the correction factor $1/H_m$ is not necessary [10]. People practicing FDR seem often to have come to the conclusion that the correction factor is too drastic for all but the most exotic and unlikely cases. This may yet turn out to be an example of wishful thinking, however, since it is difficult to test.

These new developments in multiple testing analyses will have a significant effect on the practice of testing. As an example, from the project described in section 5.2, in a 4000-test example, the number of p-values below 5% were 350, but only 9 would be accepted by the Bonferroni correction and 257 with FDR controlled at 5%. This means that the scientists involved can analyze 257 effects instead of 9, with a 5% risk that one of the 9 was wrong and only 13 of the 257 expected to be wrong.

FDR control is a rapidly developing technique, motivated by the need for acceptable alternatives to the Bonferroni correction which has very little power when many tests have been performed. Such applications are rapidly developing in medical research, where fMRI (functional Magnetic Resonance Imaging, a technique that shows brain activity by a contrast between oxygenated and deoxygenated blood which suggests energy consumption by neurons and their appendages, dendrites and axons) and micro-array investigations (a technique for genotyping and measuring gene activity in many points or genes on the genome simultaneously) produce very large numbers of very weak signals.

It is possible to produce an even stronger multiple testing procedure which only tells us that with high confidence there is at least one false null hypothesis. Such a test would build on either the Kolmogorov-Smirnov test that the p-values are uniformly distributed, or a graphical/visual test like that proposed to test for constant intensity in section 3.1.5. The drawback would be that we then know only that at least one of possibly quite many null hypotheses is false. The identification of such a case could however motivate that a larger study is performed.

Exercise 11 *(+) Five different gene variants were determined for a population of 200 normal subjects and 300 subjects diagnosed with schizophrenia. Each subject is classified with a genotype of five components, one for each gene, and the component is 11, 12 or 22, showing which variant the subject has in his/her two genes of this type. Using a standard test to decide if there is an association between gene and diagnosis, five p-values were computed related to the null hypothesis of no association (when the two subject classes have the same distribution over gene variants). These were 1.1%, 1.1% 2.3%, 2.4% and 20%. With the Bonferroni correction it is not possible to reject any null hypothesis on the 5% level. Make an FDR analysis of the situation, finding a set of hypotheses of which on the average 95% are true.*

2.2.3 Finding p-values using Monte Carlo

.

In Statistics there is an overwhelming battery of test statistics produced for various purposes, and each has its own method for determining p-values. We will shortly describe a general method to compute an approximate p-value for any real valued test statistic. It is applicable whenever we are able to generate samples from the distribution taken as the null hypothesis. We start by generating a large sample of N points, and compute the test statistic t_i for each point i, $i = 1, \ldots, N$, after renumbering the sequence so that it is increasing (this is simply done by sorting the sequence). Now the value t_i for the test statistic is given the p-value i/N for a left rejection interval, and $1 - i/N$ for right-sided rejection. When the test statistic of the data obtained in an experiment has been computed, its approximate p-value is found by interpolation in the sequence. The statistical error in this procedure can be estimated, but typically a value of N giving enough precision is easy to find. The use of this method is illustrated in the next section.

2.3 Example: Mendel's results on plant hybridization

Gregor Mendel[70] made important discoveries on inheritance of traits in plants. In one of his experiments, he examined the self-hybridization of plants (actually a kind of peas) that were heterozygotes (hybrids) with respect to a trait such as seed shape, seed color, stem length. He found that on average three out of four offspring would get the dominant trait, that can be explained by assuming a hybrid plant has one dominant and one recessive gene, denoted Aa, and assuming that the genes are randomly selected for the offspring, each with probability 1/2, giving equal probability to the four outcomes AA Aa aA and aa, only the last giving offspring with the recessive trait. Mendel's statistics where as summarized in the following Matlab table:

```
N1=[5474,1850];%seed shape
N2=[6022,2001];%seed color
N3=[705,224];%seed-coat color
N4=[882,299];%pod shape
N5=[428,152];%pod color
N6=[651,207];%flower position
N7=[787,277];%stem length
```

When Fisher got hold of these results, he felt that the proportions were closer to Mendel's stipulated 3:1 law than one would expect if the process had been a true Bernoulli process with success probability 3/4. Fisher used the χ^2 test to find that the square deviations were smaller than expected. Today we can directly simulate Mendel's stipulated process and check the variability of the results. The necessary Matlab code to compute the p-values for a test with rejection region around the 3:1 point can be written and tested out in no time (figure 13). As test statistics we take the sums of the relative squared deviations from the mean for each trait. In other words, consider the first row `N1=[5474,1850];` above, standing for an experiment where 7324 seeds obtained by self-fertilization from a hybrid parent gave 5474 peas with dominant trait and 1850 with the recessive. The mean, under Mendel's law, is the exact proportions 3:1, namely 5493 and 1831. The relative squared difference is $(1850-1831)^2/1831 + (5474-5493)^2/5493$, and this is the test statistic for the first row. For a test on the whole 7-trait experiment, the 7 test statistics are just added to get the test statistic for the whole experiment.

In figure 11 we get the distribution of the test statistic and the outcome for the seven traits. In figure 12 we see that the joint outcome has a p-value of 0.043, a bit on the low side. But if two experiments with the same outcome are added, or if the 'best' of two experiments were chosen, the hypothetical p-value would have been completely normal. Of course, it is somewhat questionable to do this analysis at all, since Cournot's principle only applies to events picked in advance. So Fisher probably made the error of using the observations to pick a set of rejection intervals containing zero, where he could equally well have chosen intervals containing infinity or a sequence of intervals containing the mode (value at largest value for the pdf under the null hypothesis of success probability 3/4). Doubling the counts as we did in the lower part of figure 12 was only an illustration, of course. Doing this in a test where we want to strengthen a real hypothesis would be completely forbidden. We are also not allowed to repeat an experiment until we get data allowing rejection of null hypothesis. If we repeat any experiment 20 times there is a good chance (probability more than 1/2) that some of them rejects any null hypothesis at level 5%. When our investigation forces us to perform many tests, we must use the theory of multiple testing(section 2.2.2).

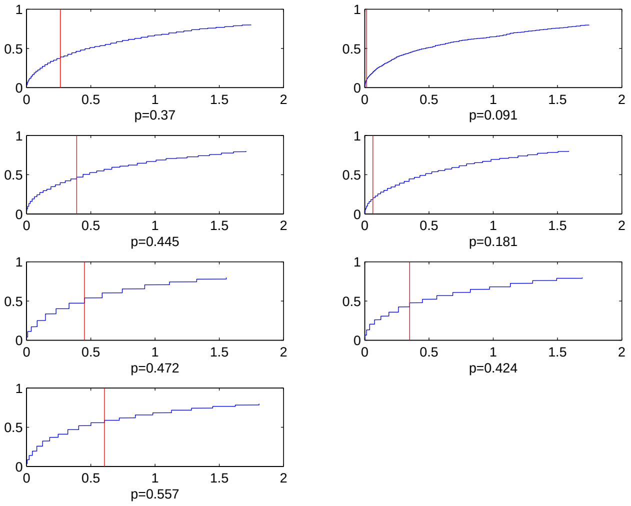

Figure 11: Stochastic simulation of Mendel's experiment, assuming that his 3:1 law holds. The p-values shown are for rejecting Mendel's law because of too little square sum deviation from the expectation values in each of the 7 trials. These values do not support rejection.

Exercise 12 *(+) A not so reputable person claims to have discovered an event with exact probability 0.5, and he wants to support his claim with an experiment where the event happened in exactly 40 of 80 occasions. Do you think he cheated? By cheating we mean reporting a score too balanced even under the assumption that the event has indeed probability 0.5. Quantify and motivate your judgment! What if it happened in 4000 of 8000 occasions?*

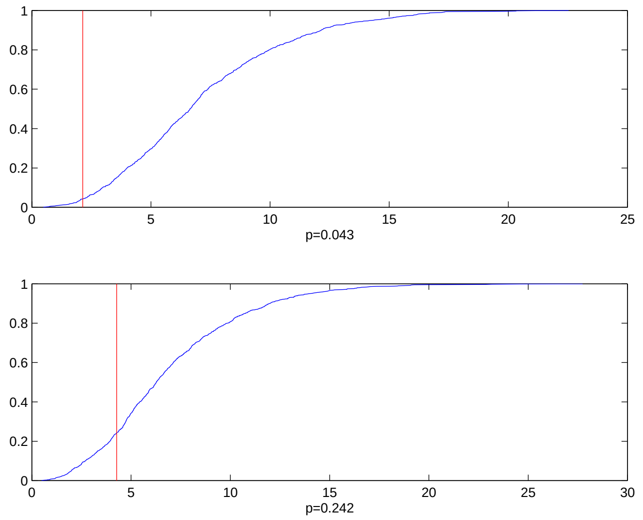

Figure 12: Putting together the 7 trials in one gives us a small p-value, 0.043, that allows rejection on the 5% level. To test the robustness of this conclusion, we can make the same computation with doubled counts (a hypothetical, twice as large, experiment). This gives a totally normal p-value.

2.4 Discussion: Bayes versus frequentism

Considering that both types of analysis is used heavily in practical and serious applications and by the most competent analysts, it would be somewhat optimistic if one thought that one of these approaches could be shown right and the other wrong. Philosophically, Bayesianism has a strong normative claim in the sense that every method that is not equivalent to Bayesianism can give results that are irrational in some circumstances, for example if one insists that inference should give a numerical measure of belief in hypotheses that can be translated to a *fair betting odds*[30, 89], or if one insists that this numerical measure be consistent under propositional logic operations[27]. Among stated problems with Bayesian analysis the most important is probably a non-robustness sometimes observed with respect to choice of prior. This has been countered by introduction of families of priors in robust Bayesian analysis[12]. Another is the need for priors in general, because there are situations where it is claimed that priors do not exist. However, this is a weaker criticism since the subjective probability philosophy does not recognize probabilities as existing: they are assessed, and can always be assessed[30, Preface]. Assessment of priors introduces an aspect of non-objectivity in inference according to some critics. However, objective probability should not be identified with objective science: good scientific practice means that all assumptions made, like model choice, significance levels, choice of experiment, as well as choice of priors, are openly described and discussed.

One can also ask if Bayesian and testing based methods give the same result in some sense. Under many different assumptions it can be proven that frequentist confidence intervals and Bayesian credible intervals will asymptotically, for

```
% Mendel's data to table NM1
NM1=[5474,1850;6022,2001;705,224;882,299;428,152;651,207;787,277];
NN=1000;%number of trials simulated
y=[1:NN]/NN;%scale y-axis to  $p$-value
%Simulation of experiment and doubled experiment
for k=1:2
  NM=NM1*k; % NM will hold the counts observed or twice these counts
  chi2=zeros(size(NM,1),NN); %Allocate space for chi2-values
  chi2s=zeros(size(NM,1),1);
  for iN=1:size(NM,1) % Consider each trait tabled in iN
    Ni=NM(iN,:); %Select the right column
    N=sum(Ni); % Total count for trait
    etr=0.75*N;% Expected number of plants/peas with dominant trait
    emiss=0.25*N; %Expected number of plants/peas with recessive trait
    R=rand(N,NN)<0.75;
    tr=sum(R); %Simulated result (dominant) for sample
    miss=N-tr; % and recessive
    %Compute chi2-values for the NN simulated trials
    chi2(iN,:)=((tr-etr).^2/etr+(miss-emiss).^2/emiss);
    % and for Mendel's table
    chi2s(iN)=(Ni(1)-etr).^2/etr+(Ni(2)-emiss).^2/emiss;
    if k==1
       p=sum(chi2s(iN,:)<chi2s(iN))/NN;
       %Plot the subtrials
       subplot(4,2,iN);xlabel(['p=',num2str(p)]);
       ss=sort(chi2(iN,:));
       cut=round(0.8*NN); % Cut tails
       plot(ss(1:cut)',y(1:cut),'b-', ...
          [chi2s(iN),chi2s(iN)]',[0,1]','r-');
       xlabel(['p=',num2str(p)]);
    end
  end;
  if k==1
    print  figur1
  end;
  subplot(2,1,k);
  schi2=sum(chi2);
  schi2s=sum(chi2s);
  schi2=sort(schi2');
  p=sum(schi2<schi2s)/NN;
  plot(schi2,y,'b-',[schi2s,schi2s]',[0,1]','r-');
  xlabel(['p=',num2str(p)]);
end
print  figure
```

Figure 13: Matlab-program for Monte-Carlo-simulation of p-values.

Figure 14: Gregor Mendel (1822-1884) was a monk and the first to discover important statistical properties of inheritance in plants. He did not get a lot of attention during his lifetime, but after three botanists independently rediscovered his laws around 1900, Mendel became one of the best known amateur scientists.

many observations, converge.

2.4.1 Model checking in Bayesian analysis

It is possible to combine the Bayesian and hypothesis-testing frameworks[45]. The parameter inference (3) gives a posterior over the parameter space, and a reasonable interpretation is that data is generated by the predictive distribution $P(D^*) = \int_\Lambda P(D^*|\lambda)f(\lambda|D)\mathrm{d}\lambda$. Now select a test statistic $t(D)$. By obtaining the test statistic of many predictive data sets D^* by this distribution we can obtain an approximate distribution of the test statistic, and if the test statistic of the real data D (from which the posterior was obtained) falls far out in the tails of the approximate distribution it is an indication that the parameterized model does not capture the data set obtained. It can be seen as a test of the hypothesis that the data was actually generated by the parameterized model used in the derivation of the posterior. The technique is suggested by Gelman also for 'informal', visual, testing. In figure 23 we can see an example of this method applied.

2.5 Evidence Theory and Dempster-Shafer structures

An important development in uncertainty management is the *Dempster- Shafer* or *evidence theory.* Originally formulated by Dempster[32], the idea was to handle the case where several sources of information having slightly different frames of reference were to be combined. Later, it was popularized by Shafer, who also decoupled the method from statistics. Philippe Smets[96] and many others have developed the method, which now exists in quite many versions.

The basic concept in DS theory is the *DS-structure*, which is simply a probability distribution over the power-set 2^Λ of the possible world set Λ, and with zero probability for the empty set. When the possible world set is conveniently modeled as a Euclidean space of some dimension, the power set is replaced by the set of boxes defined by upper and lower coordinates in each coordinate direction.

We will assume here that Λ is a finite set. The DS-structure is represented as a function, often called a *mass assignment*, $m : 2^\Lambda \rightarrow [0, 1]$, with $\sum_i m(i) = 1$ and $m(\emptyset) = 0$. Typically, many values of a mass assignment are zero, and the subsets of Λ with a non-zero mass are called *focal elements* of the assignment. Singleton subsets are sometimes called *atoms.* How should the DS-structure be regarded as evidence? Even if it is often denied to be essential, we can note that most tutorials in DS-theory start out with an example of assignment of probabilities to events, where the set $e \subset \Lambda$ represents the union of its members, and $m(e)$ is said to be the amount of belief/probability that can be attributed to the event e but not to any of its proper subsets $e' : e' \subset e, e' \neq e$. We can think of the evidence as a set of possible probability distributions over Λ. Such sets are called *imprecise probabilities*, defined by upper and lower envelopes. We will call a set of pdf:s obtainable from a DS-structure a *capacity.*

Example: In the very simple case of two states, $\Lambda = \{A, B\}$, the capacity of the mass assignment $m(A) = 0.8, m(B) = 0.1, m(\{A, B\}) = 0.1$ is a set of pdf:s giving A a probability in the interval between 0.8 and 0.9. For larger state sets, the capacities can be described as polytopes, convex sets spanned by a finite number of corner points.

Figure 15: Philippe Smets(1938-2005) developed the Transferable Belief Model, based on Dempster-Shafer's evidence theory. He makes a distinction between *credal belief* which is best represented by DS structures and two such beliefs are combined using Dempster's rule; and *pignistic belief* (used in the context of betting and decision making), represented by a probability distribution. From a credal belief, a DS-structure, the corresponding pignistic belief is obtained with the *pignistic transformation*.

In Dempster's view, a DS-structure gives an upper and a lower bound for the probability of each event $e \subset \Lambda$, called its belief (lower bound) and plausibility (upper bound), respectively:

$$\sum_{e' \subset e} m(e') \leq P(e) \leq \sum_{\{e' : e \cap e' \neq \emptyset\}} m(e') \tag{10}$$

In evidence theory, the specification of priors is usually ignored. The DS-structure is thought of as an evidence (prior, likelihood, or something else) about the system state. Two central ideas are *pignistic transformations* which take a DS- structure to a probability distribution over Λ, and *combination rules*, which say how two evidences can be combined to form a new one. The *pignistic transformation* estimates a precise probability distribution from a DS-structure by distributing evenly the probability masses of non-singleton focal elements to their singleton members. The *relative plausibility transformation*, on the other hand, consists of the normalized plausibility values for the singletons of 2^Λ. The pignistic and relative plausibility transformations are given by:
$P(w) = \sum_{w \in e} m(e)/|e|$, all $w \in \Lambda$, and
$P(w) \propto \sum_{\{w\} \cap e \neq \emptyset} m(e)$, $w \in \Lambda$, respectively.

The *Dempster's combination rule* combines two DS-structures into a new one using a random set operation: the random set intersection of the operands (regarded as random sets) conditioned on being non-empty. An alternative combination rule, the MDS rule, was recently proposed by Fixsen and Mahler [42].

Whereas Dempster's combination rule can be expressed as

$$m_{DS}(e) \propto \sum_{e = e_1 \cap e_2} m_1(e_1) m_2(e_2), e \neq \emptyset, \tag{11}$$

the MDS rule is

$$m_{MDS}(e) \propto \sum_{e = e_1 \cap e_2} m_1(e_1) m_2(e_2) \frac{|e|}{|e_1||e_2|}, e \neq \emptyset. \tag{12}$$

It is illuminating to see how the pignistic and relative plausibility transformations emerge from a precise Bayesian inference: The observation space can in this case be considered to be 2^Λ, since this represents the only distinction among observation sets surviving from the likelihoods. The likelihood will be a function $l : 2^\Lambda \times \Lambda \rightarrow [0, 1]$, the probability of seeing evidence $e \subset \Lambda$ given world state $\lambda \in \Lambda$. Given a precise $e \in 2^\Lambda$ as observation and a uniform prior, the inference over Λ would be $f(\lambda|e) \propto l(e, \lambda)$, but since we in this case have a probability distribution over the observation space, we should use equation (4), weighting the likelihoods by the masses of the DS-structures. Applying the indifference principle, $l(e, \lambda)$ should be constant for λ varying over the members of e, for each e. The other likelihood values ($\lambda \notin e$) will be zero. Two natural choices of likelihood are $l_1(e, \lambda) \propto 1$ and $l_2(e, \lambda) \propto 1/|e|$, for $\lambda \in e$. Amazingly, these two choices lead to the relative plausibility transformation and to the pignistic transformation, respectively:

$$\begin{aligned} f_i(\lambda|m) &\propto \sum_{\{e:\lambda\in e\}} m(e)l_i(e,\lambda) \qquad (13)\\ &= \begin{cases} \sum_{\{e:\lambda\in e\}} m(e)/\sum_e |e|m(e) & ,i=1\\ \sum_{\{e:\lambda\in e\}} m(e)/|e| & ,i=2 \end{cases} \end{aligned}$$

It is also possible to combine two pieces of fuzzy (DS) evidence in the form of two DS-structures m_1 and m_2. We find the task of combining the two likelihoods $\sum_e m_1(e)l(e,\lambda)$ and $\sum_e m_2(e)l(e,\lambda)$ using Laplace's parallel composition as in equation (4) over Λ, giving

$$f(\lambda) \propto \sum_{e_1,e_2} m_1(e_1)m_2(e_2)l_i(e_1,\lambda)l_i(e_2,\lambda).$$

For the choice $i=1$, this gives the relative plausibility of the result of fusing the evidences with Dempster's rule; for the likelihood l_2 associated with the pignistic transformation, we get $\sum_{e_1,e_2:\lambda\in e_1\cap e_2} m_1(e_1)m_2(e_2)/(|e_1||e_2|)$. This is the pignistic transformation of the result of combining m_1 and m_2 using the MDS rule. There is some current debate on which estimation (pignistic or relative plausibility) and combination (Dempster's DS or Fixen/Mahler's MDS) operations that should be used. I hope the above derivations show convincingly that the choice of such operators depend on the statistical model chosen for the application. In other words, none of the possible choices are intrinsically correct.

For a discussion of the relationships between standard and robust Bayesian analysis and evidence theory, see, e.g., [6]. An alternative interpretation of DS-structures exists, and was first argued for in [91].

Exercise 13 *The US Air Force has several target classification systems that give their output as a DS-structure. One such system outputs at some time its belief that an incoming target is either an Attack(Fighter), Bomber or Civilian aircraft. The parameter set is thus A, B, C. The output on one occasion is: $m(A) = 0.2, m(B) = 0.050, m(C) = 0.083, m(\{A,C\}) = 0.022, m(\{B,C\}) = 0.534, m(\{A,B,C\}) = 0.111$.*

Represent the capacity corresponding to m as a polygon in the plane containing possible combinations of $P(A)$ and $P(B)$.

2.6 Estimating a distribution, Decision and Maximum Entropy

In robust Bayesian analysis one considers convex sets of probability distributions like the capacities of DS-structures. For decision making one uses either expected utility maximax or maximin criteria, or estimates a precise probability distribution to decide from. Examples of the latter are the pignistic and relative plausibility transformations. An example of a decision-theoretically motivated estimate is the maximum entropy estimate, often used in robust probability applications [61]. This choice can be given a decision-theoretic motivation since it minimizes a game-theoretic loss function, and can also be generalized to a range of loss functions [53]. Specifically, a Decision maker must select a distribution q while Nature selects a distribution p from a convex set Γ. Nature

Figure 16: Ed Jaynes (1922-1998) championed the Maximum Entropy method, originally in physics and then in other application areas.

selects an outcome x according to its chosen distribution p, and the decision makers loss is $-\log q(x)$. This makes the decision makers expected loss equal to $E_p\{-\log q(X)\}$. The minimum (over q) of the maximum (over p) expected loss is then obtained when q is chosen to be the maximum entropy distribution in Γ. It is thus, if this loss function is accepted, optimal to use the maximum entropy transformation for decision making. An intuitive explanation of this lies in the structure of the payoff, $-\log(q(x))$. It is thus bad if we choose a q giving low probability (because then the neg logarithm is large) to x chosen by Nature. But x can only be chosen if Nature selects a p giving it a large probability. Consequently, we should avoid giving x a small probability if Nature can give it a large one. Putting the result in context, the MaxEnt estimate is appropriate when we can assume that Nature knows which distribution q Decision Maker chose when it selects its distribution p. Of course, this is a pessimistic case in general, and most likely there are many cases where a better strategy can be found, maybe the pignistic estimate, or the center of smallest enclosing sphere of the polygon.

The maximum entropy principle differs significantly from the relative plausibility and pignistic transformations, since it tends to select a point on the boundary of a set of distributions (if the set does not contain the uniform distribution), whereas the pignistic transformation selects an interior point.

The pignistic and relative plausibility transformations are linear estimators, by which we mean that they are obtained by normalization of a linear function of the masses in the DS-structure. If we buy the concept of a DS-structure as a set of possible probability distributions, it would be natural to require that as estimate we choose a possible distribution, and then the pignistic transformation of Smets gets the edge – it is not difficult to prove the following:

Proposition 1 *The pignistic transformation is the only linear estimator of a probability distribution from a DS-structure that is symmetric over Λ and always returns a distribution in the capacity represented by the DS-structure.*

Although we have no theorem to this effect, it seems as if the pignistic transformation is also a reasonable decision-oriented estimator approximately minimizing the maximum Euclidean norm of difference between the chosen distribution and the possible distributions, and better than the relative plausibility transformation as well as the maximum entropy estimate for this objective function. The estimator minimizing this maximum norm is the center of the smallest enclosing sphere. It will not be linear in m, but can be computed with some effort using methods presented, e.g., in [44]. The centroid is sometimes proposed as an estimator, but it does not correspond exactly to any known robust loss function – it is rather based on the assumption that the probability vector is uniformly distributed over the imprecision polytope.

The standard expected utility decision rule in precise probability translates in imprecise probability to producing an expected utility interval for each decision alternative, the utility of an action a being given by the interval $I_a = \cup_{f\in F} \int u(a,\lambda)f(\lambda|x)\mathrm{d}\lambda$. In a refinement proposed by Voorbraak [104], decision alternatives are compared for each pdf in the set of possible pdfs: $I_{af} = \int u(a,\lambda)f(\lambda|x)\mathrm{d}\lambda$, for $f \in F$. Decision a is now better than decision b if $I_{af} > I_{bf}$ for all $f \in F$.

Some decision alternatives will drop out as unfavorable because they are dominated in utility by others, but in general several possible decisions with overlapping utility intervals will remain. In principle, if no more information exists, any of these decisions can be considered right. But they are characterized by larger or smaller risk and opportunity.

Exercise 14 *(+) Consider again the USAF classifier of the previous exercise and the mass assignment given there.*

i)Find the relative plausibility, pignistic and Maximum Entropy estimates for the target class from m*, as three probability distributions over the parameter set.*

ii) Draw the points from i) in the polygon from the previous exercise . Conclusion or comment?

2.7 Uncertainty Management

Interpretation of observations is fundamental for many engineering applications, and is studied under the heading of *uncertainty management.* Designers have often found statistical methods unsatisfactory for such applications, and invented a considerable battery of alternative methods claimed to be better in some or all applications. This has caused significant problems in applications like tracking in command and control, where different tracking systems with different types of uncertainty management cannot easily be integrated to make optimal use of the available plots and bearings. Among alternative uncertainty management methods are Dempster-Shafer theory[91] and many types of non-monotonic reasoning. These methods can be - to some extent - interpreted as a *robust Bayesian analysis*, where the analyst need not give precise priors and likelihoods but can specify convex sets of priors and likelihoods, [12, 3] and Bayesian analysis with infinitesimal probabilities, respectively[108, 8]. We have also generalized the analyses by de Finetti, Savage and Cox, showing that under slightly weaker assumptions than theirs, uncertainty management where belief is expressed with families of probability distributions that can contain infinitesimal probabilities is the most general method satisfying compelling criteria on rationality[5, 4]. Other alternative methods, like *fuzzy logic* and *rough set theory*, neural networks and case-based reasoning should rather be seen as a search for more useful model families than those used traditionally in statistics. These methods can in principle be described in the Bayesian framework[110, 76, 83, 59], although frequently one makes heuristic validation tests rather than statistical ones. Moreover, since these methods were also developed with the intention of avoiding statistics, there is a certain reluctance to start viewing them as statistical models, besides the feeling that such analyses might be infeasible. A few statistical methods can be applied even to rather unorthodox statistical models. If my favorite method tells me that there is a certain structure in the data, how can I verify this? The standard way is to obtain a new set of observations and see if the new set shows the same structure. But new observations are in many cases unobtainable for cost or time reasons. By resampling techniques new observation sets can be created, by taking a part of the original data set. If the same structure can be seen in all or most parts of the original data set, this gives confidence that the structure seen is not only random noise. Likewise, a methods tendency to see things in data that has no real meaning can be checked by testing it on random data. By taking a real data set consisting of a number of cases each with a set of variables, and randomly shuffling each variable among the cases, one gets a data set where no serious method should find significant structure. Likewise, complex prediction methods can be tested by randomly partitioning the available data set into one training set from which the predictor is obtained and another one from which its performance is estimated. These methods should also be used when analyzing data using conventional statistical methods, since they provide good tests against programming and handling errors.

2.8 Beyond Bayes: PAC-learning and SVM

Whereas Bayesian analysis is based on the assumption that priors and likelihoods are precisely known, there is an interesting approach that is probabilistic although it makes no assumptions on the probability distributions involved. This paradigm was originated by Chervonenkis and Vapnik[103], and was recently developed in computational learning theory [101] and even more recently in the Support Vector Machine (SVM) method[28].

How can a method be probabilistic and at the same time be independent of the actual probability distributions involved? First of all, we consider only the problem of predicting an unknown quantity y from a known quantity x, and using a set of training samples $(x_i, y_i)_{i\in I}$. Here the x_i are vectors with real valued components drawn from a *feature space* R^d, and the y_i are either binary class indicators (the *classification problem*) or real numbers (for a *regression problem*). In the first case we want to use the training set to predict the class x (-1 or 1) from the new feature vector x; in the second case we want a real valued prediction y from x.

If more than two classes are to be distinguished, or if more than one real number is to be predicted, there are a number of ways to organize this using the basic binary classification or single variable prediction method we will describe.

In order to obtain distribution independence we must inevitably assume that there is some type of link between the training sample and the quantities that we try to predict. This assumption is that the training sample has been generated by the same probability distribution as those examples that we want to predict. Such a distribution can be described as a joint probability distribution $p(x, y)$ or as a family of conditional probability distributions $p(y|x)$ and a distribution over the $p(x)$. If we ignore the latter, we run the risk of having a training sample that is unrepresentative for future use, if we ignore the former we run the risk that the relation between x and y is different during training from what it is during use of the predictor. Unfortunately, even when the assumption of common distribution during training and testing is fulfilled, we run the risk of getting a bad predictor because the training sample was accidentally unrepresentative. An assumption in PAC and other machine learning is exchangeability. Since applications usually take examples in a time sequence, this assumption is seldom strictly fulfilled(see Hand[55]).

With the concept of PAC-learning, we consider a particular classifier (restricting ourselves temporarily to the binary classification case) $C(x) : x \mapsto \{-1, 1\}$. The *classifier error* is said to be the probability that $C(x)$ is different from y when (x, y) is drawn according to the distribution $p(x, y)$. The *empirical error* is the fraction of sample points where $C(x_i) \neq y_i$. If the classifier is constructed from a random sample $(x_i, y_i)_{i\in I}$ drawn according to $p(x, y)$ and the probability is less than δ to obtain a classifier error larger than ϵ, then we say that we can obtain a (δ, ϵ) classifier for the distribution p. If there is $\epsilon = \epsilon(n, \delta)$ so that this is true regardless of the distribution $p(x, y)$, then we have a PAC bound for the classification problem. This formulation has the flavor of statistical testing in the sense that it asserts that the probability of obtaining misleading data is small, regardless of the underlying distribution. On the negative side, it is obvious that bounds obtained over any distribution will usually be pessimistic compared to cases where some knowledge about the plausibility of possible distributions is encoded into a Bayesian analysis.

The basic form of the classifier in SVM stems from Rosenblatt's *perceptron* [86]. This was a hardware device taking a number of real valued inputs with correct binary classification as training data, used to set weights used to linearly separate further real-valued inputs. Let the input be a real valued vector, $x = (x_1, \ldots, x_n)$ and the class y be -1 or $+1$. The classifier decides the class of x by the rule $C(x) = \text{sign}(x_i \cdot w + b)$, where the dot $\cdot$ is scalar product and vector w and scalar b are the *weights* of the perceptron. A good deal of research has gone into finding good ways to set the weights using repeated scanning of the training set. However, if the classes of negative and positive examples can be linearly separated, finding a separating hyperplane is equivalent to finding a feasible solution of a linear program and is in principle easy[23]. A key empirical finding has been that the classifier gets better if the hyperplane is placed at maximum distance from all training points, giving a *wide margin classifier.* Such a classifier can be found using a further development of Lagrange's method with multipliers, developed by Karush, Kuhn and Tucker.

A similar technique can be used for the regression problem. Here we want to predict y by $C(x) = x_i \cdot w + b$, and the idea is to define the weights such that the maximum error $|y - C(x)|$ is minimized. This method is similar to the *Adaline*, an early neural network due to Widrow and Hoff[107].

Dynamic prediction problems can be handled with the SVM regression approach. If we have a time series (x_i, y_i), trying to predict y_i from the lagged vector $(x_{i-k}, \ldots, x_{i-1})$ for a suitable choice of k is a standard way to solve the dynamic prediction problem. It is here important to choose the time step (by sub-sampling the original time series) and lag vector dimension (the value of k) to obtain good performance. In practice the most critical aspect is the stationarity (time-independent dynamics) of the analyzed system.

The perceptron and the Adaline were designed as learning machines with a built-in assumption of linearity. This was heavily criticized by Minsky and Papert[71], whose book on the perceptron more or less blocked further research on neural networks for a long time.

The SVM builds on the concept of the Perceptron and Adaline. However, the specification of a wide-margin separator or smallest error hyperplane means that the PAC learning analysis can be performed, resulting in distribution-independent error bounds. Specifically, Cristianini and Shawe-Taylor[28, Ch 4] prove a large number of PAC bounds. The derivation of these bounds is somewhat lengthy, we will just give an example to show how they usually appear:

Proposition 2 *Given a set of n examples with binary classification, such that the support of the distribution of x lies in a ball of radius R. Fix γ. With probability $1-\delta$, any two parallel hyperplanes 2γ apart and separating the positive and negative examples (thus with margin γ), has error no more than*

$$\epsilon(n,\delta,\gamma) = \frac{2}{n}\left(\frac{64R^2}{\gamma^2}\log\frac{en\gamma}{9R^2}\log\frac{32n}{\gamma^2} + \log\frac{4}{\delta}\right), \tag{14}$$

if $n < 2/\epsilon$ *and* $64R^2/\gamma^2 < n^2$.

Similarly for regression (where y is a real number), if all training points lie within two parallel hyperplanes $2(\Theta - \gamma)$ apart, the corresponding predictor has residual greater than Θ with probability at most $\epsilon(n,\delta,\gamma)$ as defined in (14), again if $n < 2/\epsilon$ and $64R^2/\gamma^2 < n^2$.

As is obvious, the common distribution $p(x,y)$ may well prevent accurate prediction, particularly if it factors into a distribution for x and another for y, $p(x,y) = p(x)p(y)$. It is thus important in the formulation above that in such cases we are unlikely to obtain a single classifier or regressor with the required performance on the training set. We are also not allowed to create samples repeatedly until we find one on which a good predictor or classifier exists. It is also important to note that the bound γ is defined before we have seen the examples, although there are ways around this. Classification is often useful even on distributions where we are unlikely to obtain a good training set, but the real distribution of examples gives some outliers. With few outliers we can still find (ϵ,δ) bounds, and a number of related bounds are given in [28, Ch 4].

An interesting property of the bound (14) is that it is independent of the dimension (number of components) of x. In the perceptron discussion [71], it was soon noted that many important examples exist where a nonlinear mapping to a higher dimensional space makes classes linearly separable which are not linearly separable in the original space. As an example, if x is two-dimensional, the map $\Phi : (x_1, x_2) \mapsto (x_1, x_2, x_1^2, x_1x_2, x_2^2)$ has the interesting property that linear separators in the target space R^5 correspond to conic section separators in the original space R^2. This can be seen by considering the R^5 hyperplane defined by $w = (w_1, w_2, w_3, w_4, w_5)$ and b. In R^2 the R^5 plane $w \cdot x + b = 0$ is inverse mapped to:

$$w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_1x_2 + w_5x_2^2 + b = 0,$$

which is the general equation for a (possibly degenerate) conic section. Conic sections are e.g., ellipses, hyperbola and parabola, but also two parallel lines (separating the set between the lines from those outside the lines on both sides) and two crossing lines (separating the sections opposite each other from the other two opposite sections). The insensitivity of the bound (14) to the dimension of the feature space means that we can expect good generalization performance even after mapping the original examples to a high- or even infinite-dimensional space. This is accomplished in a general fashion with the Kernel trick (next section).

We can illustrate the SVM with a few examples visualizable in 2D. For the classification problem we can see in Figure 17 how positive and negative examples were uniformly generated with the red line demarcating the class. The wide margin classifier decides using the slightly different line between the two blue lines. In the generic 2D case, one blue line will usually be fixed by two support points, the other by one (case (a)). But it is also possible to have one support point on each side, in which case the blue lines are perpendicular to their connecting line (case (b)). In higher dimensions the picture is similar, but there are more cases on how the two sides of the margins are determined by support points. A midpoint and the normal vector is obtained from the multipliers of the support points.

In the regression case, x_i is 1D and extended with the y_i to get 2D z_i, Figure 18 . Here the examples were generated around the red dotted line and the SVM finds the smallest enclosing parallel lines. Here we also have two generic cases on how support points interact with the narrowest possible 'corridor', and more on higher dimension.

It is of course somewhat risky to demand full separation of classes and full compliance with the 'corridor', given the unavoidable prevalence of measurement and classification errors in real-world data sets. In many applications one uses 'soft margins' and 'soft corridors'. The detailed usage of this facility is explained in most SVM software packages.

We will not go into the detailed algorithm for finding wide margin separators and narrow margin approximators, for details see [28, Ch 5]. It is however quite important to understand one feature of the SVM method, namely the support vectors and corresponding Lagrange multipliers:

Consider the *primal optimization problem* involving functions f, g_i, with $i = 1, \ldots, k$ defined on n-dimensional space:

$$\begin{array}{rcc} \text{minimise} & f(w), & w \in R^n, \\ \text{subject to} & g_i(w) \leq 0, & i = 1, \ldots, k, \end{array}$$

For both the classification and the regression version we ask for two parallel hyperplanes with extremal (maximum or minimum) distance that fit to data samples. In the classification version we work with a space having the dimension d of the feature space given by the x_i, whereas in regression we work with dimension $d+1$ given by the x_i extended by corresponding y_i. We call these extended vectors z_i. A hyperplane is defined by $\{x : w \cdot x + b = 0\}$, and the pair of hyperplanes we are looking for will be

$$\{x^+ : w \cdot x^+ + b = +1\}, \{x^- : w \cdot x^- + b = -1\},$$

For the classification problem where $|y_i| = 1$ we ask for hyperplanes where positive examples are above the first one, and negative below the second one. This is obtained with the constraints $y_i(w \cdot x_i + b) \geq 1$. Since we want to maximize the distance between the two parallell hyperplanes we will maximize $||w||_2 = w \cdot w$.

For the regression problem, y_i is the last dimension of $z = (x_i, y_i) \in R^n$, we want all points to lie between the two hyperplanes that themselves are placed as close to each other as possible. The constraints are

$$\begin{aligned} w \cdot z + b &\leq +1, \\ w \cdot z + b &\geq -1. \end{aligned}$$

We want the borders as close as close to each other as possible, so we maximize $||w||_2 = w \cdot w$.

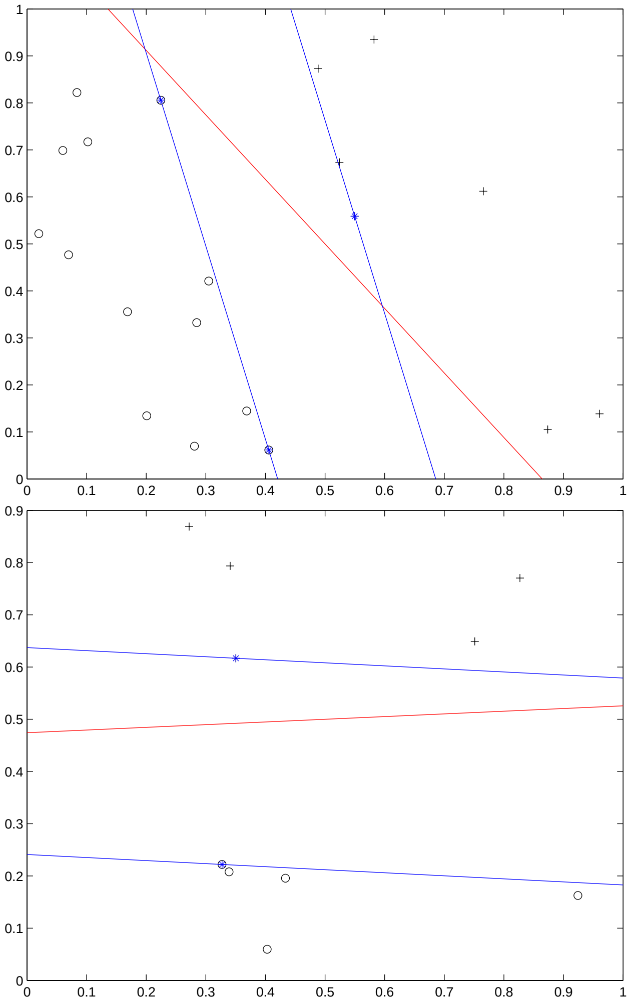

Figure 17: A sample of points distributed uniformly in the shown square. Those above the red line are positive. Blue lines indicate the margin of the wide margin classifier on these examples. With growing training set, the blue lines will approach the red one. (a): three generic support points; (b): Two generic support points

A development of Lagrange's fundamental method due to Kuhn and Tucker([63] is based on:

Proposition 3 *A necessary condition for a point w^* to be an extremal point of $f(w)$ subject to constraints $g_i(w) \leq 0$, where $i = 1, \ldots, m$, where f is convex and g_i are linear (affine), is*

$$\begin{aligned} \frac{\partial L(w^*, \alpha^*)}{\partial w} &= 0, \\ \alpha_i^* g_i(w^*) &= 0, i = 1, \ldots m \\ \alpha_i &\geq 0, i = 1 \ldots m \end{aligned}$$

where

$$L(w, \alpha) = f(w) + \sum_{i=1}^{m} \alpha_i g_i(w).$$

In the SVM (classification) problem, the Lagrangian is

$$L(w, b, \alpha) = \frac{w \cdot w}{2} - \sum_i \alpha_i (y_i (w \cdot x_i + b) - 1).$$

We look for optimum, which is solution to

$$\begin{aligned} 0 = \frac{\partial L(w^*, b, \alpha^*)}{\partial w} &= w - \sum_i \alpha_i y_i x_i, \\ 0 = \frac{\partial L(w^*, b, \alpha^*)}{\partial b} &= \sum_i \alpha_i y_i, \end{aligned}$$

We can now eliminate w and b from the Lagrangian and express it in terms of the α_i and inputs. The optimal solution is obtained from the following quadratic programming problem:
Maximize

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j)$$

under the constraints

$$\begin{aligned} \sum_i y_i \alpha_i &= 0 \\ \alpha_i &\geq 0 \end{aligned}$$

The quadratic programming problem has been extensively studied, and efficient codes form the computational engine of most SVM packages.

Three fundamental observations: (i) The formulation involves the examples x_i only in the form of scalar products, and this is essential for the 'kernel trick' to go through; (ii) the resulting multipliers α_i are nonzero only for support vectors, *i.e.,* points lying on one of the two hyperplanes; (iii) At the stationary point, $L(w^*, \alpha^*) = f(w^*)$, since for each i either α_i or $g_i(w^*)$ is zero. Once the optimal w and b are obtained from the α_i, a new example x will be classified as $C(x) = \text{sign}(w \cdot x + b)$ and a new regression example will be $C(x) = y : w \cdot (x, y) + b = 0$, or, if $w = (w_{-(d+1)}, w_{d+1})$, where $w_{-(d+1)} = (w_1, \dots, w_d)$,

$$y = -\frac{w_{-(d+1)} \cdot x + b}{w_{d+1}}.$$

Lagrange's method was developed for analytical mechanics, and there the multipliers represent forces between bodies that are in contact (the force is zero if the constraint is not tightly satisfied), and thus it is natural to regard the support vectors as outliers and the α_i as a non-conformance measure, more non-conformance the higher the value of α_i. This is an essential property for the Vovk/Gammerman hedged prediction scheme (section 2.9).

It is possible to modify the method using *soft constraints*, by also stating maximum values for the α_i. Then the solution will admit a limited number of outliers in the training sample that may get incorrect classifications. This modification is available as an option in most SVM program packages.

The stringency and elegance of the distribution-independent framework has maybe promoted its use more than justified. It is not difficult to guess that the bounds obtained will in many cases be quite large, even larger than one (which is rather damaging for error bounds on probabilities). The popularity of the SVM approach stems of course mainly from the experience that the results it gives are 'useful in practice' despite the depressingly large strict error bounds. Alternative analyses of similar methods are given in a new book by Shafer, Gammerman and Vovk [92], and a very short introduction will be given next (section 2.9). The Relevance Vector Machine of Tipping [100] uses a Bayesian approach and realizes a prediction and classification structure that is in some respects more attractive than the SVM.

2.8.1 The Kernel Trick

As stated above, the SVM algorithms work using the example vectors only through scalar products in the feature space. Thus, if we map the features x_i to a higher dimensional space using the mapping Φ, we will have to compute scalar products $\Phi(x) \cdot \Phi(z)$. For the quadratic map $\Phi : (x_1, x_2) \mapsto (x_1, x_2, x_1^2, x_1 x_2, x_2^2)$ where subscripts denote vector components, we get $\Phi(x) \cdot \Phi(z) = x_1 z_1 + x_2 z_2 + x_1^2 z_1^2 + x_1 x_2 z_1 z_2 + x_2^2 z_2^2$. This can almost be expressed as $(x \cdot z + 1)^2$, the difference being that there is a constant term and that the different terms are re-weighted by constant multipliers, and corresponds to a slightly different map with the same 'power' to model surfaces as inverse images in R^2 of hyperplanes in R^5:

$$\Phi'(x) = (\sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1 x_2, x_2^2).$$

The kernel technique is built on Mercer's theorem, which says that certain types of functions $K(x, y)$ of two feature vectors will certainly correspond to some mapping Φ in the sense that $\Phi(x) \cdot \Phi(z) = K(x, z)$. In this case we do not need to explicitly map the example vectors to high-dimensional space: The SVM algorithm works by only using vectors through scalar products, and these can be directly evaluated from the kernel $K(x, z)$.

Mercer's theorem is part of the toolbox in mathematical physics [25] and is quite interesting: it says under what conditions the eigenfunctions of an integral equation have the property that the scalar product of the eigenfunction vectors at two points is equal to the kernel at the two points. This allows one to use mappings into infinite-dimensional space, since the eigenfunction sets are often infinite. The background is Mercer's condition on a symmetric kernel $K(x, y)$ of an integral equation:

$$\int_C K(x, y) f(x) \mathrm{d}x = \lambda f(y)$$

This equation can be though of as a simple eigenvector problem, but in an infinite-dimensional Hilbert space. There is a sequence of eigenvalues λ_i decreasing in magnitude, and a corresponding sequence of real valued (normalized) eigenfunctions $\Psi_i : C \rightarrow R$. We are interested in cases where all eigenvalues are positive and where the scaled eigenfunctions $\Phi_i = \sqrt{\lambda_i}\Psi_i$ satisfy our requirement

$$K(x, y) = \sum_i \Phi_i(x) \cdot \Phi_i(y).$$

If this is the case, the kernel trick works for the map $\Phi(x) = (\Phi_1(x), \Phi_2(x), \ldots)$ and we can run the margin algorithm without actually mapping the feature vectors to the possibly infinite-dimensional space. Mercer's theorem says that this will be the case for a symmetric kernel $K(x, y)$ whenever Mercer's condition is satisfied:

$$\int_C K(x, y) g(x) g(y) \mathrm{d}x \mathrm{d}y \geq 0,$$

for every square integrable function $g(x)$, and if C is a compact subset of some space R^d (the original feature space).

We can now generalize our conic section separator $(x \cdot z + 1)^2$ to $K(x, z) = (x \cdot z + 1)^d$. This kernel correspond to mapping a feature vector x to a high-dimensional vector whose components are all monomials of degree not greater than d.

There are large numbers of kernels suitable for different types of modeling problems, see, e.g., [28, 88]. The kernel trick is not only applicable in the SVM context: Whenever the computation required can be described in terms of scalar products of feature vectors the method is applicable. Several examples are given, e.g., in [18]. The polynomial kernels are easy to comprehend, but sometimes Gaussian kernels are preferable. The Gaussian kernel is defined by:

$$K(x, y) = \exp(-||x - z||^2 / \sigma^2).$$

This kernel contains the sigmoid functions popular in neural network modeling and is popular because of its flexibility.

In Figure 19 we show an example of using Gaussian kernels in a classifier for a checker-board training set. The positive and negative examples lie in alternate squares of a checker-board (say, black squares are negative, white squares positive). Positive and negative examples are marked + and × . The examples are mapped to a high-dimensional space using the Gaussian kernel, and the figure shows the inverse projection of the wide margin separator in the 10D space (levels -1, 0 and 1). The support vectors touch the 1 and -1 contour lines (marked by black dots).

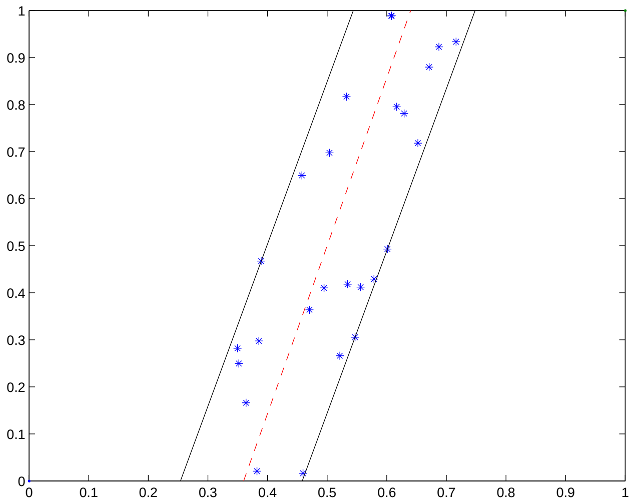

Figure 18: Support vector regression. Points lie around the red dotted line. Blue lines indicate the walls of the section in which points lie. Three generic support points, two to the right and one to the left.

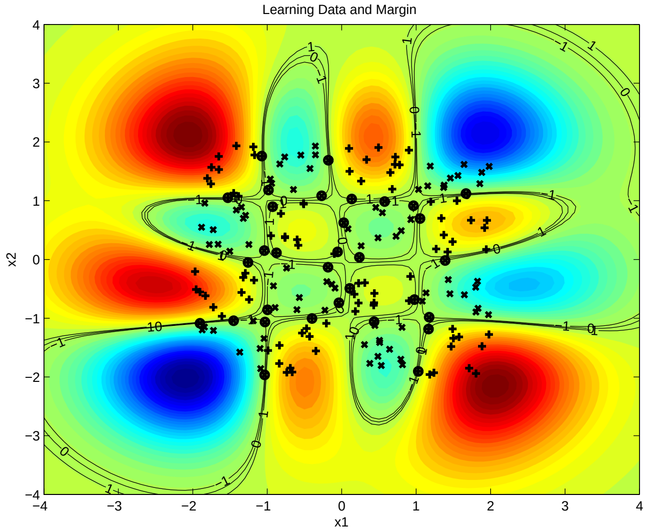

Figure 19: Gaussian kernel applied to the checker-board problem. Output of `SVM-KM` classification example[17].

2.9 Algorithmic conformal prediction and anomaly detection

A new (with some predecessors, see [43, Discussion]) approach to online learning and prediction was developed by Vovk and Gammerman and is pedagogically explained in [92]. The main advantage is that it gives well-founded confidence and credibility measures of individual predictions. The main practical idea is to base the analysis on a non-conformance measure. We can think of two situations where the examples $(z_1, \ldots, z_i, \ldots)$ are fed into our method, a prediction or anomaly measure is made for each new z_i based on its relation to the batch of previously seen examples. A prediction is as in the SVM scheme: each z_i is a pair (x_i, y_i) where the y_i are either from a finite set (classification) or a vector of reals (regression case). This method centers around the concept of non-conformance. A *non-conformance measure* can, in analogy with a test statistic, be chosen quite freely, but a proper choice is also a prerequisite for non-disappointing behavior. The non-conformance measure measures the difference of an example from a set of previous examples, an example based on linear regression thinking being:

$$c(\{z_i\}_{i \in I}, z) = (y - \text{argmin}_\beta(\sum_i (y_i - x_i\beta)^2)x)^2,$$

where the non-conformance of the new item $z = (x, y)$ thus is measured by the prediction error arising from a least-squares regression predictor found from the set of old items. The formula is equally applicable when the x, x_i and β are vectors instead of scalars.

It is required that the value of $c(Z, z)$ is invariant to permutations of Z (actually, we defined Z as a set, so this is implicit in our notation). It is now possible to define p-values for the members of the sequence $(z_i)_{i=1}^n$: The *p-value* for z_i in $(z_i)_{i=1}^n$ is

$$(1 - (r_i - 1)/(n - 1)), \tag{15}$$

where r_i is the rank of $c(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}, z_i)$ among the $c(\{z_1, \ldots, z_{j-1}, z_{j+1}, \ldots\}, z_j)$, $1 \leq j \leq n$ and the p-value of z in $\{z_i\}$ is $(1-(r_i - 1)/n)$ where r_i is the rank of $c(\{z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n\}, z)$ in the values $c(\{z_1, \ldots, z_{j-1}, z, z_{j+1}, \ldots, z_n\}, z_j)$, $1 \leq j \leq n$ with the former added (thus the rank among $n + 1$ values). The p-value is thus low if most z_i are less different from $\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$ than z is from $\{z_1, \ldots, z_n\}$.

Given a non-conformance measure, a set $\{z_1, \ldots, z_l\}$ and a new item x, a predictor of confidence level α for the corresponding y is the set of y-values giving p-value larger than $1 - \alpha$ to $z = (x, y)$.

Vovk and Gammerman go on to define *confidence predictors*, mapping a non-conformance measure, a training set, a new observation x, and a confidence level $1 - \epsilon$ to a set Γ^ϵ of predicted values, namely the set of values Y such that the p-value of the last element in $((x_1, y_1), \ldots, (x_N, y_N), (x, Y))$ is larger than ϵ.

For the classification problem, we will find a finite set of p-values for the possible classifications of the new item. We predict the one with highest p-value, and call this p-value our *credibility* - if it is high (particularly if it is one, its maximum possible value) it is completely plausible that we gave the right class. A low credibility tells us that the new instance is not well covered by the training set - every class is somewhat not in conformance with it. However, even with high credibility there is a possibility that another class is almost equally plausible (typically in cases with an insufficient training set or non-informative features, or when the new instance is somewhat a border-line case). The second highest p-value, subtracted from one, is the *confidence* of the prediction. This is 1 minus the largest ϵ such that the predictor set Γ^ϵ is a singleton, just the predicted value. A high confidence value shows that all alternatives are implausible, whereas a low value shows that one or more alternatives are completely plausible. with both confidence and credibility high, we can be relatively certain that the class predicted is correct, as always if the exchangeability assumption is satisfied.

The interesting result giving strong support to this methodology is the following: It is valid for *smoothed* p-values, obtained by probabilistically (with probability $1/2$) counting the non-conformance values that are ties with that of the last one (i.e. $l+1$), instead of counting them to zero as in (15):

The smoothed p-value is obtained as follows: Suppose α_i is the non-conformance measure of (x_i, y_i) in the $(l+1)$-sequence for $l = 1, \ldots, l$ and α_{l+1} is that for (x_{l+1}, Y). Then the smoothed p-value for Y is

$$p_Y = \frac{|i : \alpha_i > \alpha_{l+1}| + \eta |i : \alpha_i = \alpha_{l+1}|}{l+1}, \tag{16}$$

where η is a standard random variable uniformly distributed in $[0, 1]$. (This means that every time a p_Y is computed, η is obtained as a new standard random number.)

For an exchangeable sequence, the smoothed p-value and an iterated prediction sequence, where a predictor set Y with confidence ϵ is obtained for $y = y_{l+1}$ from $((x_1, y_1), \ldots, (x_l, y_l))$ and x_{l+1}, for $l = 1, 2, \ldots$, the frequency of $y \in Y$ will be ϵ. Likewise, for a classification sequence, the erroneous classifications will be independent with probability one minus the maximum confidence corresponding to a singleton predictor set.

Since the smoothed p-values are larger than the non-smoothed ones, the predictor based on non-smoothed p-values will perform better (on the average) than indicated by the confidence.

For the case of prediction, one must typically be content with ϵ-confident sets for the predicted value, because there are typically infinitely many of them. It is also not obvious how, in the general case, these sets shall be computed. Indeed, each conformance measure needs analysis of how to compute or approximate these sets efficiently.

A generally good way to apply the hedged prediction methodology is to use the Lagrange multipliers (α_i) in SVM to compute (smoothed) p-values.

A statistician is usually annoyed by application owners' frequently expressed desires to find anomalies in an ongoing process. In the application, anomalies may be signs of some previously unknown phenomenon that, when investigated and understood, will drive development of the field, or be a warning sign for an impending disaster. In the framework of conformal prediction, the concept of *anomaly* is easy to define: it is a new item with a low p-value, and anomaly is with respect to the chosen non-conformance measure.

It is also possible to select anomalies based on p-values using the FDR method. This aloows us to state that a high proportion of the detected anomalies will be 'real'.

3 Data models

The formulas in Chapter 2.1 always contain a data probability distribution term, in parameterized models as a function of (conditioned on) parameters. We will now go through a number of such distributions often used in Bayesian analyses and uncertainty management. You have probably seen most of it in standard statistics courses, but the Bayesian angle of these notes makes the emphasis different. We start out from simple data types and work towards complex ones.

3.1 Univariate data

Univariate data consist of real numbers, integers, ordinal or categorical values. The latter two are usually coded as integers, but the magnitudes of these have no significance – the values can be thought of as members of a set. This set is ordered for an ordinal variable but not for a categorical one. Ordinal variables are common in investigations of attitudes of persons. In a hotel room you may find a form where you are asked among other things if the service provided is very bad, bad, good or very good - your answer is sometimes considered an ordinal variable (but more often a numeric score which is then averaged over customers).

3.1.1 Discrete distributions and the Dirichlet prior

We have already seen the coin-tossing example (section 2.1.10). There the data are binary categorical (heads or tails), and the parameter is the probability of heads (or, by symmetry, tails). The inference is about this parameter, and the priors and posteriors in section 2.1.10 are distributions for the parameter. The distributions chosen are called Beta distributions, and many of them (specifically, those with integer parameters, since the number of successes and failures in our experiment must be integers) are generated from the uniform distribution by multiplication with likelihoods and normalization.

This can be generalized to general discrete distributions over values for a categorical variable taking d different values. The distribution in this case is a probability vector (with the constraints that the component values range from 0 to 1 as probabilities do, and that the components sum to one).

For a *discrete distribution* over d values, the outcome is a number from 1 to d. The parameter set is a sequence of probabilities $\overline{x} = (x_1, \ldots x_d)$, (in some textbooks the last parameter x_d is omitted - it is determined by the first $d-1$ ones), constrained to lie in the polytope L_d:

$$L_d = \{(x_1 \ldots, x_d) | 0 \leq x_i, \text{ for } i = 1, \ldots, d \text{ , and } \sum_{i=1}^{d} x_i = 1\}$$

The likelihood after observing a set of outcomes where outcome i occurred n_i times is

$$\Pi_i x_i^{n_i}.$$

Normalizing this quantity so that its integral over L_d becomes one, we have an example of the *Dirichlet distribution.*

The Dirichlet distribution with parameter set $\overline{\alpha}$ is

$$\mathrm{Di}(\overline{x}|\overline{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i x_i^{(\alpha_i-1)} = c_\alpha \prod_i x_i^{(\alpha_i-1)}, \tag{17}$$

where $\Gamma(n+1) = n!$ for natural number n. The normalizing constant $c_{\overline{\alpha}} = \Gamma(\sum_i \alpha_i)/\prod_i \Gamma(\alpha_i)$, although by no means obvious, can be verified using multiple induction (Exercise 15). It gives a useful mnemonic for integrating any monomial, like $\prod_i x_i^{(\alpha_i-1)}$ over the $d-1$-dimensional hyperplane segment L_d. It also gives the normalization constant $\Gamma(d)$ for the uniform distribution over L_d (which is often used as a prior). It is very convenient to use Dirichlet priors, for the posterior is also a Dirichlet distribution: After having obtained data with frequency count $\overline{n} = (n_1, \ldots, n_d)$ we just add this vector to the prior parameter vector $\overline{\alpha}$ to get the posterior parameter vector $\overline{\alpha} + \overline{n}$.

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 with $\alpha_i = 1$. This is, e.g, the prior used by Laplace to derive the rule of succession, see Ch 18 of [60] or [31]. Other priors have been used, but there are no strong reasons for them except where they correspond to prior knowledge. Such knowledge can in principle be fed into the analysis in the form of an equivalent prior sample.

3.1.2 Estimators and Data probability of Dirichlet distribution

The normalization constant:

$$c_{\overline{\alpha}} = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}$$

of the Dirichlet distribution makes it usually unnecessary to perform integrals over L_d. As an example, the Laplace estimator for the probability is the mean of the distribution, and its ith component is:

$$\int_{L_d} x_i \mathrm{Di}(\overline{x}|\overline{\alpha}) \mathrm{d}\overline{x} = \frac{c_{\overline{\alpha}}}{c_{\overline{\alpha}+1_i}} = \alpha_i / \sum_i \alpha_i,$$

where 1_i is the vector of zeros except for a one in the ith component. For inference of a probability from occurrence counts with uniform prior we have a posterior $\mathrm{Di}(\overline{x}|\overline{n}+\overline{1})$ and thus the mean estimator is the relative occurrence counts after 1 has been added to each outcome. Using Lagrange multipliers it is straightforward to show that the MAP estimator is just the (unmodified) observed relative frequencies. In many practical applications it has been found important to use the Laplace (mean) estimator instead of the MAP estimator. Particularly, the MAP probability estimate for an event that has not happened is exactly zero, and this is a much too drastic estimate if the total number of occurrences is small.

We will now find the data probability for a general discrete distribution $(x_1, \ldots, x_d)$, given that n_i occurrences of the ith outcome were observed, $i = 1, \ldots, d$. Let $n = \sum_i n_i$. The probability of observing a particular sequence with counts n_i is $\Pi_i x_i^{n_i}$, and to obtain the probability of the counts we should multiply with the multinomial coefficient $(\binom{n}{n_1, \ldots n_d})$. 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 over the $x_i$and a prior $\mathrm{Di}(\overline{x}|\overline{\alpha})$ on its parameters):

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

As is 'easily' seen, the uniform prior ($\alpha_i = 1$, all i) gives a probability for each sample size that is independent of the actual data:

$$p(\overline{n}|M) = \frac{\Gamma(n+1)\Gamma(d)}{\Gamma(n+d)}. \qquad (19)$$

The probability $p'(\overline{n}|M)$ of a sequence of outcomes with counts $\overline{n}$ is of course obtained by dividing with the corresponding multinomial coefficient:

$$p'(\overline{n}|M) = \frac{\Gamma(n+1)\Gamma(d)}{\Gamma(n+d)\binom{n}{n_1,\ldots,n_d}}. \qquad (20)$$

Exercise 15 *Derive the normalization constant for the Dirichlet distribution*
(i) with $\alpha_i = 1$, all i;
(ii) with integer parameters α_i.

3.1.3 The normal and t distributions

The most common assumption about a real valued variable is that it has the normal distribution. This assumption has some motivation as there are several theorems saying that under certain assumptions a quantity will be normally distributed. However, in practice most data sets analyzed do not have the normal distribution. Usually, the tails of empirical distributions are longer than the very short tails of the normal distribution. So whether or not the normal assumption is appropriate depends a lot on what kinds of data you have, and on the purpose of the analysis.

The normal distribution has parameter μ and σ^2, the mean and variance. The distribution is

$$f(x|\mu,\sigma^2) = N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

There is an alternative parameterization using the *precision* λ instead of variance, with $\lambda = 1/\sigma^2$. It is particularly popular in Bayesian statistics[16].

So let us analyze a sequence of real valued measurements $(x_i)_{i=1}^n$, assumed to be independently drawn from some normal distribution whose parameters we want to make inference about. The likelihood function is then

$$f((x_i)|\mu,\sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left(-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right)$$

which, applying the abbreviations $\overline{x} = \sum_i x_i/n$ and $s^2 = \sum_i(x_i - \overline{x})^2/n$, simplifies to

$$f((x_i)|\mu,\sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left(-\frac{1}{2\sigma^2}(ns^2 - 2n\mu\overline{x} + n\mu^2)\right). \qquad (21)$$

Since the likelihood is only dependent on data through the summaries $\overline{x}$ and s^2, these are so called *sufficient statistics* - we do not need to retain the individual values for inference under the normality assumption. However, the individual values can well be needed for other purposes, e.g., for testing whether or not the assumption of normality is plausible.

In order to apply (3), we must have priors on μ and σ^2. But we could also assume that one of these is known, and that the purpose of analysis is to make inference on the other. This allows for a more pedagogical progression, but is also prototypical for several real applications. We could for example assume that σ^2 is known, like in the case where we make repeated measurements of a fixed quantity in order to improve precision. We may have access to a statement of measuring accuracy that can be translated to a known value for σ^2 when making inference about μ. Or, we may want to make inference on measurement accuracy by repeatedly measuring a fixed quantity where the 'true' value is known, like measuring the length of the archive meter, when making inference about σ^2.

Assume we know the value of σ^2. A natural choice of uninformative prior for μ is the uniform distribution. However, since μ can vary over the whole real line, the uniform distribution does not exist as a (normalized) probability distribution (since the integral $\int_{-\infty}^{\infty} \mathrm{d}\mu$ for the normalization constant diverges). It is however possible to use priors that cannot be normalized to probability distributions, and it is standard for the normal distribution. Such non-normalizable priors are called *improper priors.*

The case of fixed σ^2 and improper uniform prior allows us to rewrite the likelihood (21) as

$$f((x_i)|\mu,\sigma^2) = f(\mu|(x_i),\sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(n(\mu-\overline{x})^2+C)\right),$$

where the quantity $C = \overline{x}^2 + ns^2$ does not contain the parameter μ. This quantity will disappear after normalization. This is, regarded as an unnormalized posterior for μ, a normal distribution with mean $\overline{x}$ and variance σ^2/n. So despite the fact that the prior is improper, the posterior, the product of the improper prior 1 and the likelihood, is a proper distribution for all non-empty data sets ($n > 0$), essentially because the likelihood has a convergent integral with respect to μ over the real line.

One standard prior assumption for the normal distribution is that $(\mu, \log\sigma)$ are uniformly distributed over the whole plane, and thus that σ has a prior marginal distribution proportional to $1/\sigma$. This is also an improper prior since the integral $\int 1/\sigma \mathrm{d}\sigma$ diverges, both at 0 and at $+\infty$. Again, the posterior is a proper distribution, since the normalization constant is the inverse of the convergent integral $\int_0^\infty \sigma^{-n-1}\exp(-\frac{1}{2\sigma^2}n(\mu-\overline{x})^2+C)\mathrm{d}\sigma$. You may not have seen this distribution, but looking in a table such as [79] or [16, App.A], it is readily identified as a Gamma distribution (see exercise 17).

Finally, we may want to make inference on both μ and σ. Their joint posterior under the standard improper prior assumption is given by equation (21) multiplied by the improper prior σ^{-1}, which is maybe somewhat unintuitive. It would be nice to find the marginal probabilities of the two parameters, like in the case where we are only interested in the mean μ. This is obtained by integrating out the σ, an instance of a common procedure known as integrating out *nuisance parameters.* Somewhat surprisingly, the marginal distribution of μ is no longer a normal distribution but the somewhat more long-tailed t-distribution. The reason is that averaging over a long tail of the distribution over σ makes the posterior a continuous mixture of normal distributions with varying variances. Indeed,

$$f(\mu|(x_i)) \propto \int_0^\infty \sigma^{-n-1}\exp\left(-\tfrac{1}{2\sigma^2}(n((\overline{x}-\mu)^2+\overline{x}^2+ns^2))\right)\mathrm{d}\sigma \propto \tag{22}$$

$$\left(1+\tfrac{n(\mu-\overline{x})^2}{ns^2}\right)^{-n/2}, \tag{23}$$

where the last line was obtained after a change of variable $z = (ns^2 + n(\mu - \overline{x})^2/(2\sigma^2)$. This is an instance of the t distribution, which is usually given the parametric form with parameters called mean (μ), variance (σ) and degrees of freedom (ν):

$$f(x;\mu,\sigma^2,\nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma}\left(1+\frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\right)^{-(\nu+1)/2} \tag{24}$$

Exercise 16 *A sample $y_1,\ldots,y_n$ of real numbers has been obtained. It is known to consist of independent variables with a common normal distribution. This distribution has known variance σ^2 and the mean is known to be either* 1 *or* 2 *(hypotheses H_1 and H_2, both cases considered equally plausible).*

(i) Which is the data probability function $P(D|H_i)$?
(ii) Describe a reasonable Bayesian method for deciding the mean value.
(iii) Characterize the power of the suggested procedure as a function of σ^2, assuming that the sample consists of a single point.

Exercise 17 *Show that the posterior for inference of normal distribution variance under the standard assumption is a gamma distribution, and find its parameters.*

3.1.4 Nonparametrics and mixtures

A common desire in analysis of univariate (and multivariate) data is to make inference on the distribution of a sample without assuming anything about the analytical form of the distribution. In other words we do not want to make inference about parameters of a schoolbook distribution, but about the distribution itself. This field of investigation is called *non-parametric* inference. In principle, the equation (3) is applicable also in non-parametric analysis. In this case the parameter set is the set of all distributions we want to consider, and we must naturally, because this set is overwhelmingly large, have a suitable prior for this parameter. We can also expect that posterior computation will be completely different from the simple estimation of a few parameters we saw in the previous examples. The set of all distributions includes also functions with strange sets of singularities.

In applying non-parametric Bayesian analysis we feed an observation set into (3). Here we have the first clue to non-parametrics: if the sample is small, the points will be far apart and we can never get a good handle on the small scale behavior of the distribution, so we must content ourselves with inference among a set of fairly regular (smooth) functions. We will consider a couple of ways to accomplish this. The first method consists in discretization: divide the range of the variable into bins and assume a general discrete distribution for the probabilities of the variable falling into each bin. From such a general distribution, a pdf that is piece-wise constant over each bin can be used as an estimate for the actual distribution.

The second example shows how a distribution can be modeled as a *mixture* of simpler distributions. If the components are normal distributions, the mixture can be expressed as

$$f(x) = \sum_{i=1}^{n} \lambda_i N(x, \mu_i, \sigma_i). \tag{25}$$

Here, the *mixing coefficients* λ_i form a probability distribution, i.e., they are non-negative and sum to one. We can think of drawing a value of the mixture as first drawing the index i according to the distribution defined by the mixing coefficients and then drawing the variable according to $N(x|\mu_i, \sigma_i^2)$. In this case we cannot know for certain which component generated a particular data item, since the support of the distributions overlap.

3.1.5 Piecewise constant distribution

In this section we use the method of section 3.1.1 above. We design an MCMC trace that approaches the posterior distribution of the breakpoints delineating the bins of constant intensity. Since different 'binnings' of the variable gives different resolutions of the estimate, we can approach the scale selection problem by trying different bin sizes and obtaining a posterior estimate for the appropriate number of bins, using (1) to compare the description of an interval either as one or as two bins in each case assuming that the probability density of values is constant over each bin.

Consider now a sample of real valued variables (x_i) and a division of the real line into bins, so that bin i consists of the interval from b_i to b_{i+1}. It does not matter much to which bin we assign the dividing points b_i – these points form a set of probability zero in this model. Each division into bins can be considered a composite model, and we compare the posterior probabilities of these models to find a probability distribution over models. As a preview of the MCMC method, we will describe an iterative procedure to sample from this posterior of bin boundaries and bin probabilities with respect to the observed data. We will also describe an advanced technique for keeping down the dimension of the state-space, Rao-Blackwellization. This sampling is organized as a sequence of model visits, where we visit the models in such a way that the chain, taken as a sample, is a draw from the posterior. When we are positioned at one model, we consider visiting other models produced either by splitting a bin into two, or by merging two adjacent bins into one. We are thus in each step comparing two models, the finer having d bins. We always consider a uniform Dirichlet prior for the probabilities $(x_1, \ldots, x_d)$ of a variable falling into each of the d bins. Bin i has width w_i. The data probability is then $\Pi_i x_i^{n_i}$. The normalization constant in (17) tells us that over uniformly distributed x_i, the data probability will be

$$\Pi_i \Gamma(n_i + 1)/\Gamma(\Sigma_i n_i + 1). \tag{26}$$

Consider the adjacent model where bins j and $j + 1$ have been merged. We now assume in the model that the probability density is constant over bins j and $j + 1$, so the probability x_j in the coarser model should be split into αx_j and $(1 - \alpha)x_j$, which probabilities generated the counts n_j and n_{j+1} in the finer model. The proportion α is the fraction of the combined bin length belonging to the first constituent, $\alpha = w_j/(w_j + w_{j+1})$. The data probability for the coarser model will now be the product $x_1^{n_1} \cdots (\alpha x_j)^{n_j} ((1-\alpha)x_j)^{n_{j+1}} x_{j+2}^{n_{j+2}} \cdots x_d^{n_d}$ Integrating out the x_i (now only $d-1$ variables) leads to the posterior probability

$$\alpha^{n_j} (1 - \alpha)^{n_{j+1}} \frac{\Gamma(n_j + n_{j+1} + 1)\Pi_{i \notin \{j,j+1\}} \Gamma(n_i + 1)}{\Gamma((\sum n_i + 1) - 1)}. \tag{27}$$

We must have some idea of how many bins there should be. A conventional assumption is that change-points are events with constant intensity, so the number of change points should follow some Poisson distribution and the probability of d change points is $\exp(-\lambda)\lambda^d/d!$, with a hyper-parameter λ saying how many change-points we expect. The Bayes factor in favor of the finer model is thus:

$$\alpha^{-n_j} (1 - \alpha)^{-n_{j+1}} \frac{\Gamma(n_j + 1)\Gamma(n_{j+1} + 1)}{(n + d - 1)\Gamma(n_j + n_{j+1} + 1)} \frac{\lambda}{(d + 1)},$$

where the last factor comes from the Poisson prior and the first is obtained by dividing (26) by (27). The technique used above to integrate out the x_i is known as Rao-Blackwellization[20]. It can improve MCMC-computations tremendously.

While this method of non-parametric inference is very easy to apply for a univariate data inference problem, it does not really generalize to high-dimensional problems. The assumption of constant probability density over each bin is sometimes appropriate, as in making inference on the intensity of an event over time, in which case it tries to identify a set of 'change points'. A classical example of change-point analysis is a data set over times at which coal-mining disasters occurred in England[80]. After having identified probable change points of the disaster intensity, one can for example try to identify changes in coal-mining practice associated with the change-points in intensity. A result from the popular coal-mining disaster data set is shown in figure 20. This data set has been analyzed with similar objectives in several papers, e.g., in [51]. Compared to the analysis there, which uses a slightly different prior on the function space, we get quite similar change-points and similar density estimates. It is typical in MCMC-approaches that subtle differences between models that seem equally plausible give different results. Since programming is involved, it is also useful to check the model by running it on a similar set of points generated with a uniform distribution, and check that it proposes a more even density of change points than the actual coal-mining disaster data does. One random example is shown in figure 21. The graph can be judged with experience to 'more random' for the synthetic data than for the real data, since it has a much more diffuse distribution of change-points, and the probability of one change-point is quite low (ca 7%). In order to convince ourselves more thoroughly that the coal-mining disaster data cannot reasonably be a sample from a uniform distribution, we can generate a large number of random data sets with matching number of points (simulated accidents uniformly distributed) and compare their cumulative plots with those of the coal mining data. The result is shown in figure 23. This is a popular and time-efficient way to do tests with visual inspection of a suitable plot. If you want a 'real' p-value you can define a test statistic that is the number of accidents in the first half of the interval. The p-value is thus approximated by the proportion of simulated plots that lie below the real data plot in the midpoint (ca 1906). This number is 0, and even with considerably more simulated data it would be 0, and we can state that the true value is clearly below any significance level used in practice (0.1% is the lowest in common use).

It may be interesting to know what might have caused the change around 1888: according to the analysis in [80], this time period was characterized by a fairly steep decline in productivity (yield per hour worked). This was apparently not caused by state regulation of safety measures in mining, but in the build-up of trade unions in the mining industry.

In other applications, the assumption of piece-wise constant densities may be less appropriate. An example is when one estimates a distribution in order to find clusters in the underlying generation mechanism. For example, we may be interested in finding possible subspecies in a population where each population member has a quantity (weight, height etc) that is explained as a normal distribution with sub-species specific parameters μ and σ^2. In this case we can model the distribution as a *mixture of normals* (next section 3.1.6).

Exercise 18 *(+) We tested above (figure 23) the hypothesis that the coal-mining disaster density is in fact constant. Is it possible to also test the hypothesis that the intensity is piece-wise constant? Can we test the hypothesis that there is exactly one breakpoint, and at 1888? How?*

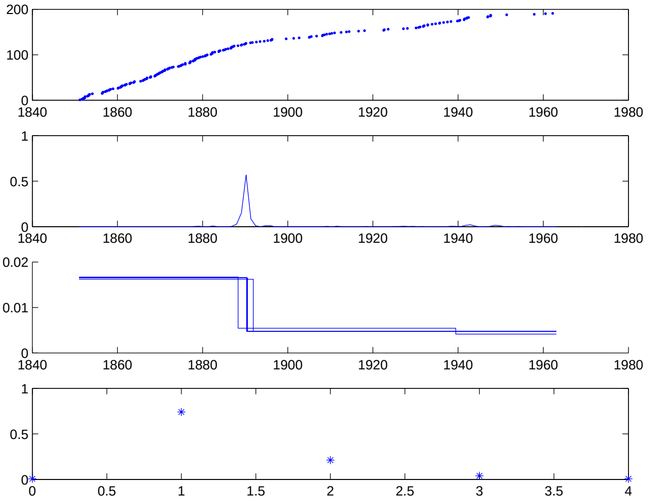

Figure 20: Coal mining disaster data. plots of occurrence times of disasters, MC estimate of density of change points, five overlaid density estimates and the observed number of change points distribution. The second plot shows what is known as the probability hypothesis density (PHD) in the more complex case of multi-target tracking[1], expected to be used, e.g., in future intelligent cruise controls to keep track of other vehicles and obstacles in road traffic.

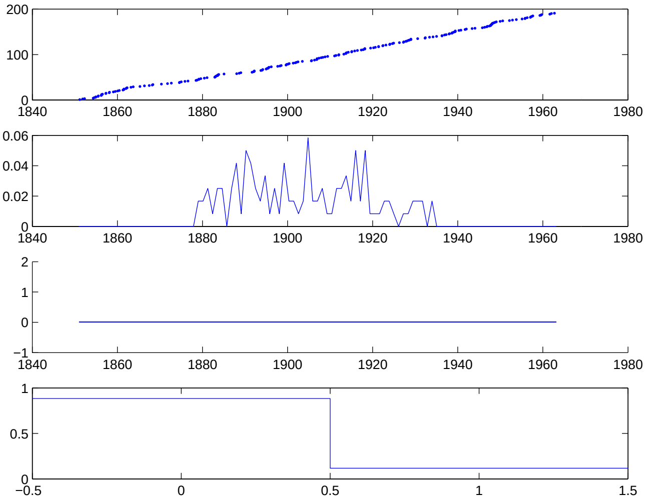

Figure 21: Checking the model: Similar data set with uniformly distributed points.

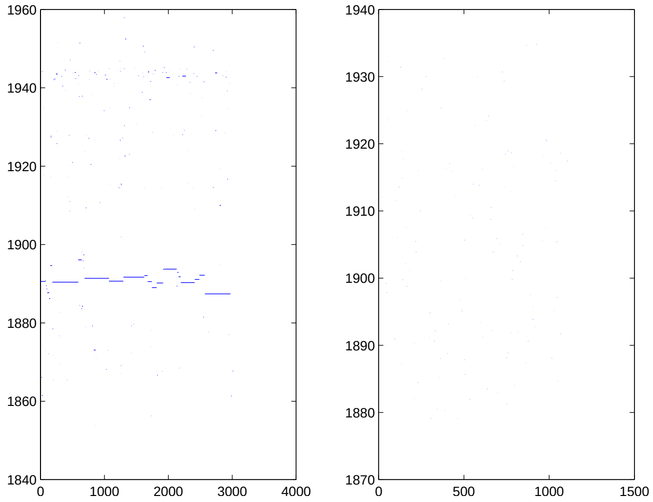

Figure 22: The trace of change-points for the real (left) and random test data. The real data trace shows an obvious concentration around 1888, and a less obvious one around 1944. The latter might be ascribed to the end of a world war, but it is not at all obviously significant. The creation of mine workers' unions seems to have had a more concrete effect than WW II. For the random test data, we can not, as expected, see any obvious structure (it is possible to compare it with the graph of figure 1, which also has no significant structure).

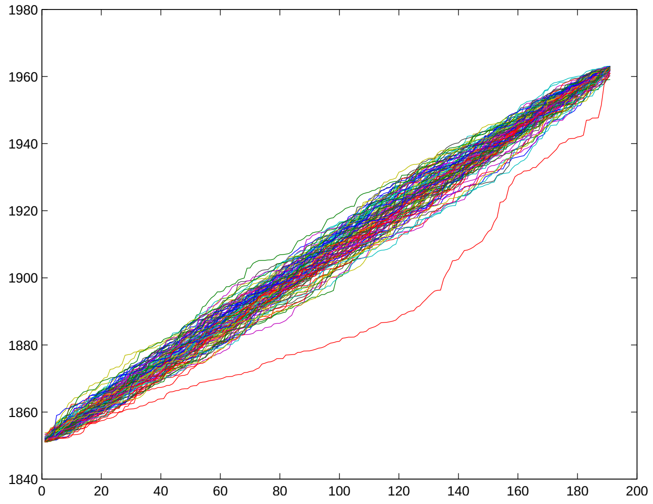

Figure 23: Coal mining data against 100 sets of uniformly distributed random data.The coordinates are cumulative count and year. The lowest curve comes from the real data and is considerably below those from the 100 random data sets. The coal-mining disaster data has thus significant non-uniform structure

3.1.6 Univariate Gaussian Mixture modeling - The EM and MCMC ways

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[82], the discovered components, if any, could correspond to discrete 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[58]. The MCMC method is quite straightforward but somewhat tedious to implement. For a number of components in the mixture, the algorithm keeps a trace that contains the current number of components, the weight, mean and variance of each component, and also the assignment of each data point to one of the components. A proposal will be the reassignment of a point to another component, a change in the mean or variance of a component, a change in weights of components, the deletion of a component (merging two components) or addition of a new component (splitting a component into two). In order to keep the identity of components in the trace, one should not allow the mean of one to change outside the interval between its neighbors. In order to speed up the computation, one should keep the sum and square sum of points in each component updated. The MCMC method can be rather slow for large data sets, but it gives a full picture of the posterior: a pdf of the number of components and for each component an empirical distribution of its weight and parameters. A detailed analysis of this method can be found in [74], and practical tips in [45]. A more sophisticated way to account for the 'label-switching' problem is found in [90].

It is much more common in practice to use Expectation Maximization for this problem. For a fixed number of components, the data points are assigned to components and then two steps alternate until convergence:

1: compute the values of weights, means and variances that maximizes the likelihood of the current assignment of points to components.

2: compute the assignment of points to components that maximize the likelihood, i.e., each data point is assigned to the component having highest density at that point.

The method will converge rapidly in most cases, but it does not give a good picture of the uncertainty of the MAP estimate. It is possible to get a picture of the sensitivity using resampling, but this is not equivalent to posterior estimation. A useful package in Matlab was published by Patrick Tsui (Mathworks file exchange). It covers also multivariate Gaussian mixtures.

For a mixture modelling using MCMC and univariate normal distributions it can be advantageous to make a Rao-Blackwellization by integrating out the variances (but not the means) of the component normal distributions. The data probability for a given component will then be (where summation is over the data points in the component) obtained by integration over the variance (from

0 to ∞) of the data probability (a product of Gaussians) times the standard prior ($1/\sigma$) (see Exercise 19)

Exercise 19 *Show that the data probability for a component in a mixture is*

$$\frac{\Gamma(n/2)}{(2\pi s)^n},$$

where n is the number of points and $s = (\sum x_i^2 - 2\mu \sum x_i + n\mu)$, sums being taken over the x_i of the component. Hint: with the substitution $t = 1/\sigma^2$, the integral can be pattern matched with a constant times an integral of the gamma distribution. The integral can be solved using the normalization constant for the gamma distribution obtained from a table of distributions.

3.2 Multivariate and sequence data models

In principle, a univariate data item or random variable can be generalized by two recursive operations, forming sets and sequences, respectively. A (multivariate) set is formed as a finite collection, or as an indexed sequence. Since the operations are recursive, we can in principle create sets of sequences, sequences of sets, and many more complicated structures, which we will avoid here. Sequences can be thought of as time series, but they are used also to describe, e.g., genes and proteins. For time series we may want to predict their future behavior, which is typically done using various filter operations that can be collectively described with equation (6). But for other sequences it may be more important to make inference about a hidden 'latent structure', using the retrodiction equation (7).

Sets of variables are typically analyzed using either a graphical dependency model (common for categorical data), or with various types of normality assumptions, like factor analysis and regression. Also here it may be useful to make inferences about hidden structure in the set, typically by decomposing the data set into classes that each have a simpler statistical structure than the total set.

3.2.1 Multivariate models

Consider a data matrix where rows are cases and columns are variables. In a medical research application, the row is associated with a person or an investigation (patient and date). In an internet use application the case could be an interaction session. 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. The former should be coded in a separate column of categorical (binary) data. 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.

In many applications the definition of the data matrix is not obvious. For example, in text mining applications, the character sequences of information items are not directly of interest, but a complex coding of their meaning must be done, taking into account the (natural) language used and the purpose of the application.

3.2.2 Dependency tests for categorical variables: Dirichlet modeling

Assume that we observe pairs (a, b) of discrete variables. How can we find out whether or not they are dependent? There is a large number of procedures proposed to answer this question, but one that is particularly elegant and intuitive in Bayesian analysis. We want to choose between two parameterized models, one that captures independence and one that captures dependency. The first one M_I is characterized by two discrete distributions, one for A assumed to have d_A outcomes and one for B assumed to have d_B outcomes. Their probability vectors are $(x_1^A, \ldots, x_{d_A}^A)$ and $(x_1^B, \ldots, x_{d_B}^B)$, respectively. This model gives the probability $x_i^A x_j^B$ for outcome (i, j). The second model M_D says that the two variables are generated jointly by one discrete distribution having $d_A d_B$ outcomes, each outcome defining both an outcome for A and one for B. This distribution is characterized by the probability vector $(x_1^{AB}, \ldots, x_{d_A d_B}^{AB})$.

In order to get composite hypotheses we must equip our models with priors over the parameter spaces. Let us choose uniform priors for all three participating distributions.

A set of outcomes for a sequence of discrete variables can be conveniently summarized in a *contingency table*. In our case this is a matrix $\overline{n}$ with element n_{ij} giving the number of occurrences of (i, j) in the sample. We can now find the probabilities of an outcome $\overline{n}$ for the two models by integrating out the parameters from the product of likelihood and prior.

As we saw in the derivation of equation (19), the uniform prior ($\alpha_i = 1$, all i) gives a probability for each sample size that is independent of the actual data:

$$p(\overline{n}|M) = \frac{\Gamma(n+1)\Gamma(d)}{\Gamma(n+d)}. \tag{28}$$

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 $n_{.j}$ and $n_{i.}$ be the marginal counts where we have summed over the 'dotted' index, and $n = n_{..} = \sum_{ij} n_{ij}$. The probability of the data given M_D is obtained by replacing the products and replacing d by $d_A d_B$ in equation (19):

$$p(\overline{n}|M_D) = \frac{\Gamma(n+1)\Gamma(d_A d_B)}{\Gamma(n+d_A d_B)}. \tag{29}$$

We now consider the model of independence, M_I. 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:

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

Integration over the uniform priors for A and B, $\Gamma(d_a)$ and $\Gamma(d_b)$, gives the data probability given the composite model M_I:

$$
\begin{aligned}
&p(\overline{n}|M_I) = \\
&\quad\int_{L_{d_a}L_{d_B}} p(\overline{n}|\overline{x}^A x^B)p(\overline{x}^A)p(\overline{x}^B)\mathrm{d}\overline{x}^A\mathrm{d}\overline{x}^B \\
&\quad= \int_{L_{d_a}L_{d_B}}\binom{n}{n_{11},\ldots,n_{d_A d_B}}\prod_{i=1}^{d_A}(x_i^A)^{n_{i.}}\prod_{j=1}^{d_B}(x_j^B)^{n_{.j}}\times \\
&\quad\Gamma(d_A)\Gamma(d_B)\mathrm{d}\overline{x}^A\mathrm{d}\overline{x}^B \\
&\quad= \frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)\Gamma(n+d_B)}\frac{\prod_i\Gamma(n_{i.}+1)\prod_j\Gamma(n_{.j}+1)}{\prod_{ij}\Gamma(n_{ij}+1)}.
\end{aligned}
$$

From the above and equation (29) we obtain the Bayes factor in favor of independence:

$$
\begin{aligned}
&\frac{p(\overline{n}|M_I)}{p(\overline{n}|M_D)} = \\
&\quad\frac{\Gamma(n+d_A d_B)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)\Gamma(n+d_B)\Gamma(d_A d_B)}\frac{\prod_j\Gamma(n_{.j}+1)\prod_i\Gamma(n_{i.}+1)}{\prod_{ij}\Gamma(n_{ij}+1)}. \qquad (30)
\end{aligned}
$$

The analysis presented above will be used in, and continued in, section 3.3 on graphical models. The gamma functions have very large values, and expressions involving them should normally be evaluated as logarithms to prevent numeric overflow. For example, Matlab has a useful function `gammaln` that evaluates the logarithm of the gamma function directly, as a sum of logarithms for integer arguments.

Exercise 20 *For deriving the data probability in the independence model it was crucial to observe that*

$$
\prod_{i,j=1}^{d_A,d_B}(x_i^A x_j^B)^{n_{ij}} = \prod_{i=1}^{d_A}(x_i^A)^{n_{i.}}\prod_{j=1}^{d_B}(x_j^B)^{n_{.j}}.
$$

Convince yourself that this is true.

Exercise 21 *Design a method to decide whether or not a set of data triplets is generated as three independent variables. For domain size d_A, d_B, d_C, and contingency table n_{ijk} introduce notation similar to above equation (30).*

Exercise 22 *(+) In Pediatrics 2005;116;1506-1512, a study on hospital organization is reported where in one case the mortality of admitted patients was 39 out of 1394, and in another case 36 out of 548. Can this result be explained by random variation, the underlying mortality rate being the same in both cases? Quantify your conclusion!*

3.2.3 The multivariate normal distribution

The multivariate normal distribution is ubiquitous in statistical modeling and has an abundance of interesting properties. We will describe it very shortly and indicate some central analysis methods based on it. The distribution is a distribution over points in a d-dimensional space, and one way to get a multivariate normal distribution is to multiply together a number of univariate normal distributions, one over each dimension x_i. Some of these may have variance zero, which effectively is a Dirac δ-function and constrains the density to a subspace (such distributions are called degenerate). This gives a distribution whose equidensity surfaces form an ellipsoid in d-space, possibly but not necessarily with different lengths on the principal axes which are parallel to the coordinate axes. In case several principal axes have the same length, it is not possible to tell the orientation of them, we only know that they are mutually orthogonal and span a subspace. So, e.g., the ball used in American football has only one well-defined principal direction while the ball used in soccer has none. We have now the most general form of the multivariate normal distribution, with one important exception, namely that any distribution obtained by rotating the frame of reference is also a multivariate normal distribution. The analytic form of the density function is

$$p(x) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \tag{31}$$

where x varies over R^d, μ is the mean and Σ is the variance (often called covariance) matrix, a symmetric positive definite d by d matrix. The inverse of the variance matrix is sometimes called the *precision matrix.* The eigenvectors of Σ are indeed the principal directions and its eigenvector with the largest eigenvalue shows the direction of largest variation of the distribution.

Models based on the multivariate normal distribution may be oriented towards making inferences about the orientation of the axes particularly in principal component and factor analysis, and regression, or in expressing an empirical data set as a mixture of multivariate distributions. The latter can be used both to explain structure of the data and as a non-parametric inference method for multivariate data, a generalization of the method described in section 3.1.6.

Exercise 23 *The multivariate normal distribution has the following nice recursive characterization: A distribution over d-dimensional space is normal if and only if either, $d = 1$ and it is the univariate normal distribution, or, $d > 1$ and both all marginals and all conditionals of it are multivariate normal distributions over $d-1$-dimensional space. Show that this is true! Hint: A conditional of a multivariate distribution is obtained by conditioning on one of its variables. A marginal is obtained by integrating out one of its variables.*

3.2.4 Dynamical systems and Bayesian time series analysis

The analysis of time series is one of the central tasks in statistics. We will here only describe a number of useful but maybe less known methods.

For the case of a series with known likelihoods connecting observations to system state and known system state dynamics, the task will normally be to predict the next or a future state of the system. The standard solution method is conveniently formulated in the Chapman Kolmogoroff equation (6), which can be solved with a Kalman filter for known linear dynamics and known Gaussian process noise. For non-linear, non-Gaussian systems, the sequential MCMC or particle filter method described in section 4.1.2 can be used. There are tricky cases if the state and observation spaces have non-uniform structure, such as in the multiple target tracking problem, where the task is to find the probable number of targets (and their states) from cluttered radar readings. This problem has an established solution in the FISST[50] method.

Bayesian methods can also be used to make inference, not only about measurement and process noise, but also on the system state space dimension. In this case we consider a system with a not directly visible state $s(t)$ at time t, and where some information about the state is obtained by observations $x(t)$ at time t. For such systems, a basic theorem is Takens[97], considering a system governed by a first-order differential equation and an observation process, contaminated by process and measurement noise $\chi(t)$ and $\psi(t)$:

$$\begin{aligned} s'(t) = &\quad G(s(t)) + \chi(t) \\ x(t) = &\quad F(s(t)) + \psi(t). \end{aligned} \tag{32}$$

Here the state and observation spaces are assumed to be of finite dimension (but not finite!), although a major interest in the physical interpretation of (32) is that the finite-dimensional state space of physical systems is often embedded in an infinite-dimensional state space, as in some laminar flow problems. The theorem of Takens says that under suitable but general assumptions, the space of lag-vectors $(x(t), x(t-T), \ldots, x(t-(d-1)T))$ is diffeomorphic to the state space for d large enough, and its interest stems from the possibility of constructing an approximate state space from this principle, i.e., by estimation the lowest value of $d+1$ that gives a degenerate set of lag-vectors obtained from experimental data. The exact formulation of the theorem has a number of regularity assumptions and it is formulated in asymptotic form assuming no noise. However, it is in many cases possible to construct non-linear predictors for such systems with extremely good performance. The idea is to find suitable values for the dimension of a state space and time step, collect a large number of trajectories, and to find approximations of the functions F and G using numerical approximation methods. The main problem with finding such predictors are that the process and measurement noise may be too large for reliable identification and that the system may be drifting so that the functions F and G change with time. An analysis of such methods used to solve challenge problems in time series prediction in the 1991 Santa Fe competition can be found in [46]. A more rigorous analysis based on Bayesian analysis and where the functions F and G of (32) are represented as special types of neural networks can be found in [102]. The functions F and G are represented parametrically as:

$$\begin{aligned} F(s) = & \quad B\tanh(As+a)+b \\ G(s) = & \quad s+D\tanh(Cs+c)+d, \end{aligned}$$

where the number of columns of B and rows of A is the number of hidden cells in the observation function neural network and the number of columns of D and rows of C is the number of hidden cells in the transition kernel ANN. The components of matrices and vectors are the weights of the networks. The method puts independent Gaussian priors on all weights. In their Matlab implementation, Valpola and Karhunen do not use MCMC but a version of the variational Bayes method they claim to be significantly faster than MCMC.

3.2.5 Discrete states and observations: The hidden Markov model

This special case occurs naturally in some applications like speech analysis after some signal processing (the number of phonemes in speech is finite, around 27), and is intrinsic in bio-informatics analysis of genes and proteins.

Consider a finite sequence of observations $(x_1, x_2, \ldots, x_n)$. Assume that this sequence was obtained from an unknown Markov chain on a finite state space: $(s_1, s_2, \ldots, s_n)$, with state transitions and observations defined by discrete probability processes $P(s_t|s_{t-1})$ and $P(x_t|s_t)$. This is the same setup as in equation (6). The sizes of the state space and sometimes the observation space is however often not known, but can be guessed, or subject to inference based on the retrodiction equation (7). Assume they are fixed to d_o and d_s, respectively. Now the unknowns are the state transition probabilities, a $d_s \times d_s$ matrix of d_s discrete probability distributions over d_s values, and one $d_o \times d_s$ observation distribution matrix of one discrete observation distribution for each system state. If no particular application knowledge about the parameters is available, it is natural to assume priors giving uniform Dirichlet priors to the probability distributions, and the missing observations can be either assumed uniformly distributed, or if they are few, distributed as the non-missing observations.

An MCMC procedure for making inferences about the two ($d_s \times d_s$ and $d_o \times d_s$) matrices goes as follows: The variables are the state vector $(s_1, s_2, \ldots, s_n)$ and the two matrices. We can also accommodate a number of missing observations as variables. The discrete distributions, the state vector and the missing values are initialized to essentially arbitrary values. In each step of the chain we propose a change in one of the $2d_s$ distributions, a state s_i or one or more of the missing observations. By considering the change in probability (7) and symmetry of the proposal distribution we compute the acceptance probability for the move, draw a standard uniform random number and accept or reject the proposal depending on whether or not the random number is less than the acceptance probability.

The probability is the product of all transition and observation probabilities connecting a state with its successor state in S, or a state s_t with the corresponding observation. Only a few of these probabilities actually change by the proposal and it is important to take advantage of this in the computation. Evaluation of the run must be made by examination of summaries of a trace of the computation. Typically, peaked posteriors are a sign that some real structure of the system has been identified whereas diffuse posteriors leave the possibility that there is no structure in the series (high noise level or even only noise in relation to the length of the series) still plausible.

3.3 Graphical 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[67]. As an example, in figure 24 M_1 indicates a model where A and B are dependent, whereas they are independent in model M_2. These are thus graphical representations of the models called M_D and M_I, respectively, in section 3.2.2. In figure 25, 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. Directed graphical models described here are also known as *Markov Random Fields(MRF)* . Our analysis in this section is oriented towards graphs with relatively simple structure (tree-like). For such models it is possible to make inference about edges in the graph. Another family of mathematically identical statistical models but with a much denser, grid-like, neighborhood structure is used for example in image processing and spatial statistics. For such models it is infeasible to make inference about the neighborhood structure, and methods originating in statistical mechanics can be applied. In this case we call the models Markov Random Fields (see section 3.6). Technically, our undirected model family is also a Markov Random Field model family, but it is usually not described as such.

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[66, 67], although one must - as always in statistics - constantly remember that dependencies are not necessarily causalities.

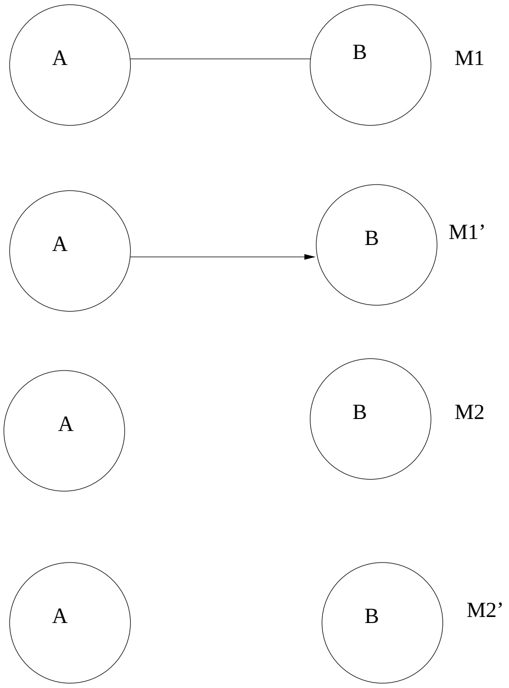

Figure 24: Graphical models, dependence or independence?

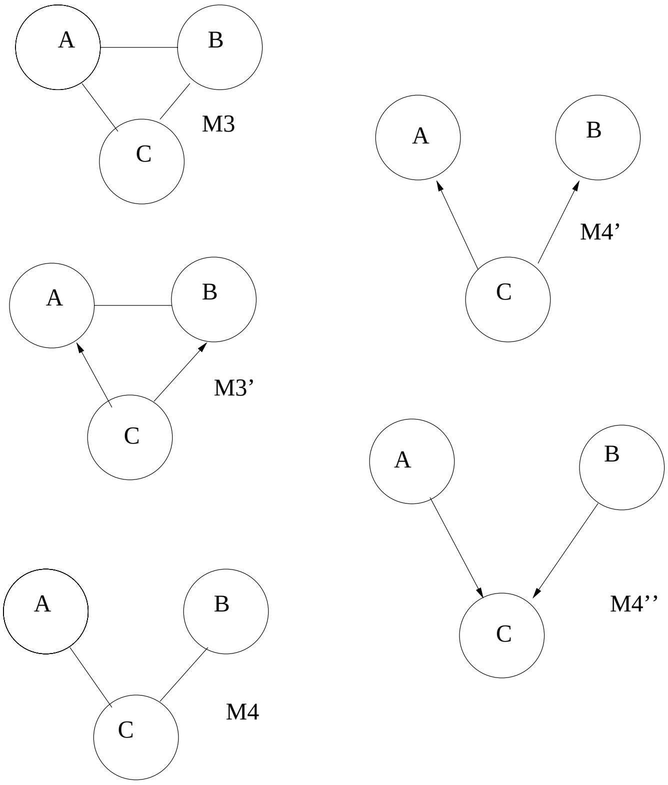

Figure 25: Graphical models, conditional independence?

3.3.1 Causality and direction in graphical models

Normally, the identification of cause and effect must depend on ones understanding of the mechanisms that generated the data. There are several claims or semi-claims that purely computational statistical methods can identify causal relations among a set of variables. What is worth remembering is that these methods create suggestions, and that even the concept of cause is not unambiguously defined but a result of the way the external world is viewed. The claim that causes can be found is based on the observation that directionality can in some instances be identified in graphical models: Consider the models M_4'' and M_4' of figure 25. In M_4', variables A and B could be expected to be marginally dependent, whereas in M_4'' they would be independent. On the other hand, conditional on the value of C, the opposite would hold: dependence between A and B in M_4'' and independence in M_4'! This means that it is possible to identify the direction of arrows in some cases in directed graphical models. It is difficult to believe that the causal influence should not follow the direction of arrows in those cases, and this opens the possibility, e.g., of scavenging through all kinds of data bases recording important variables in society, looking for causes of unwanted things. Since causality immediately suggest the possibility of control by manipulation of the cause, this is both a promising and a dangerous idea. Certainly, this is a potentially useful idea, but it cannot be applied in isolation from the application expertise, as the following example illustrates. It is known as Simpson's paradox, although it is not paradoxical at all.

3.3.2 Simpson's paradox

Consider the application of drug testing. We have a new wonder drug that we hope cures an important disease. We find a population of 800 subjects who have got the disease; they are asked to participate in the trial and given a choice between the new drug and the alternative treatment currently assumed to be best. Fortunately, half the subjects, 400, chose the new drug. Of these, 200 recover. Of those 400 who chose the traditional treatment, only 160 recovered. Since the test population seems large enough, we can conclude that the new drug causes recovery of 50% of patients, whereas the traditional treatment only cures 40%. Since this is known to be a nasty disease difficult to cure, particularly for women, this seems to be a cause for celebration until someone suddenly remembers that the recovery rate for men using traditional treatment is supposed to be better than the 50% shown by the trial. Maybe the drug is not advantageous for men? Fortunately, in this case it was easy to find the sex of each subject and to make separate judgments for men and women.

So when men and women are separated, we find the following table:

	recovery	no recovery	Total	rec. rate
men				
treated	180	120	300	60%
not treated	70	30	100	70%
women				
treated	20	80	100	20%
not treated	90	210	300	30%
total				
treated	200	200	400	50%
not treated	160	240	400	40%

Obviously, the recovery rate is lower for the new treatment, both for women and for men! Examining the table reveals the reason, which is not paradoxical at all: the disease is more severe for women, and the explanation for the apparent benefits of the new treatment is simply that it was tried by more men, for whom the disease is less severe. The sex influences both the severity of the disease and the willingness to test the new treatment, in other words sex is a *confounder*. This situation can always occur in studies of complex systems like living humans and most biological, engineering or economic systems that are not entirely understood, and the confounder can be much more subtle than sex. A first remedy is to balance the design with respect to known confounders. But in these types of systems there are always unknown confounders, and the only safe way to test new treatments is to make a 'double blind' test, where the treatment is assigned randomly, and neither the subject nor the physician knows which alternative was chosen for a subject. Needless to say, this design is ethically questionable or at least debatable, and also one reason why drug testing is extremely expensive. For complex engineered systems, however, systematically disturbing the system is one of the most effective ways of understanding its dynamics, hindered not by ethical but sometimes by economic concerns.

When we want to find the direction of causal links, the same effect can occur. In complex systems of nature, and even in commercial warehouse data bases, it is not unlikely that we have not even measured the variable that will ultimately become the explanation of a causal effect. Such an unknown and unmeasured causal variable can easily turn the direction of causal influence indicated by the comparison between models M_4'' and M_4', at least unless the data is abundant. A pedagogical example of this effect is when a causal influence from a barometer reading to tomorrows weather is found. The cause of tomorrows weather is however not the meter reading, but the air pressure it measures, and one can even claim that the air pressure is caused by weather in the vicinity that will be tomorrows weather here. This is not the total nonsense it may appear to be, because it is probably more effective to control tomorrows weather by controlling the weather in the vicinity than by controlling the local air pressure or by manipulating the barometer. In other cases this confounding can be much more subtle.

Nevertheless, the new theories of causality have attracted a lot of interest, and if applied with caution they should be quite useful[77, 49]. Their philosophical content is that a mechanism, causality, that could earlier not or only with difficulty be formalized, has become available for analysis in observational data, whereas it could earlier only be accessed in controlled experiments.

3.3.3 Segmentation and clustering

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[21]. In figure 26 we illustrate the use of segmentation: The segment number can be regarded as a hidden variable, and if the segments are chosen to minimize within-segment dependencies, and succeeds, then the data is given a much more explanatory graphical model.

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. Aggregation 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 aggregation is clearly related to the problems of 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.

3.3.4 Graphical model analysis - overview

Graphical models are quite popular, but the literature is unfortunately somewhat incoherent. It is probably best to approach the problem using our repertoire of Bayesian analysis. Both directed and undirected models (Bayesian nets or Markov Random Fields) describe a multivariate distribution. They are characterized by (i) a graph describing 'dependencies' in a visually attractive way, and (ii) probability tables, describing the precise numerical nature of these dependencies. In the choice between directed and undirected models there is really no general rule to apply; in many areas one of these have become 'standard' for no obvious reason. The directed models are the most common today, particularly when it is felt that qualitative causality reasoning can go a long way to identify 'the right' graph. The undirected (MRF) graphical models are however

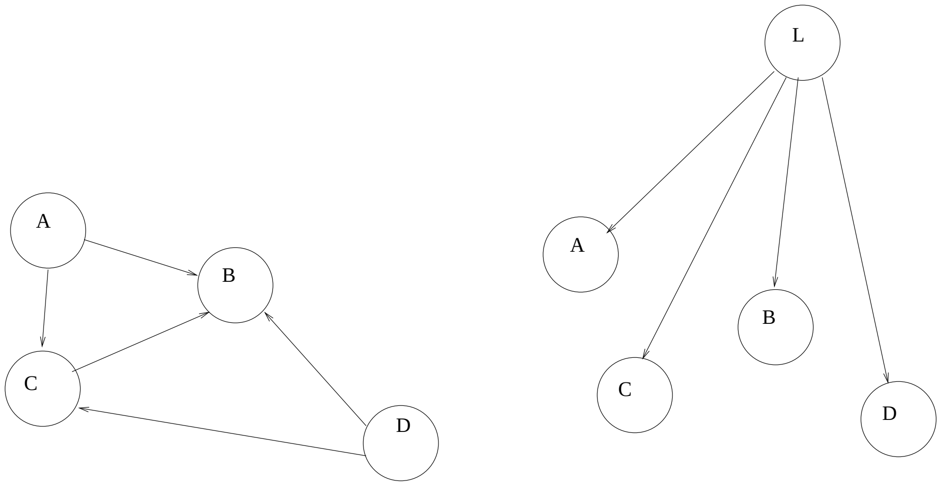

Figure 26: Segmentation explains data: The variables themselves are highly dependent (left). If a segmentation into components with independent variables can be found, then the variable indicating segment number can be added to get simpler dependencies. It is called a latent or hidden variable.

simpler and easier to work with. In particular, they do not suffer from the problem found for Bayesian nets that there are many different graphs defining the same family of probability distributions, which makes inference about the graph somewhat error-prone, if the the inference is over a set of graphs where one edge can have both directions.

After having chosen between a few main families of models, one has to use either prior information or a past set of outcomes - a training data set to fill in the items (i) and (ii) above.

The prior information is normally not sufficient for deciding which graph to use and which probability distributions to enter into the graph. Using Bayesian inference one can find a posterior distribution over graphs and probability tables, and hopefully some standard estimation can select one exact model that is good enough for the application. Typically, the graph structure is found with a MAP estimate on the marginal distribution (integrating out probability tables as nuisance parameters), followed by a Laplace (mean) estimate for the probability tables. This is another example of the Rao-Blackwellization method used already in Ch 3.1.5. We will now go through some details in this procedure.

3.3.5 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. The analysis in section 3.2.2 shows how we can choose between models M_1 and M_2:

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.

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. This corresponds to the directed model M_1': With uniform priors we get:

$$p(\overline{n}|M_1') = \frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)^{d_A}}{\Gamma(n+d_A)} \prod_i \frac{\Gamma(n_{i.}+1)}{\Gamma(n_{i.}+d_B)} \tag{33}$$

Observe that we are not allowed to decide between the undirected M_1 and the directed model M_1' based on equations (29) and (33). 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.

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. The first approach corresponds to the analysis in section 3.2.2, equation (30). From equations (29) and (30) we obtain the Bayes factor for the undirected data model:

$$\begin{aligned}&\frac{p(\overline{n}|M_2)}{p(\overline{n}|M_1)} = \\ &\qquad \frac{\Gamma(n+d_Ad_B)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)\Gamma(n+d_B)\Gamma(d_Ad_B)} \frac{\prod_j \Gamma(n_{.j}+1)\prod_i \Gamma(n_{i.}+1)}{\prod_{ij}\Gamma(n_{ij}+1)}.\end{aligned} \tag{34}$$

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

$$\begin{aligned}&p(\overline{n}|M_2') = \\ &\quad \frac{\Gamma(n+1)\Gamma(d_A)}{\Gamma(n+d_A)} \int_{L_{d_B}} (\prod_i \binom{n_{i.}}{n_{i1}\dots n_{id_B}} \prod_j x_j^{n_{ij}})\Gamma(d_B)\mathrm{d}\overline{x}^B = \\ &\quad \frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)} (\prod_i \binom{n_{i.}}{n_{i1}\dots n_{id_B}}) \int_{L_{d_B}} \prod_j x_j^{n_{.j}} \mathrm{d}\overline{x}^B = \\ &\quad \frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)} (\prod_i \binom{n_{i.}}{n_{i1}\dots n_{id_B}}))/c_{(n_{.j}+1)} = \\ &\quad \frac{\Gamma(n+1)\Gamma(d_A)\Gamma(d_B)}{\Gamma(n+d_A)\Gamma(n+d_B)} \frac{\prod_i \Gamma(n_{i.}+1)\prod_j \Gamma(n_{.j}+1)}{\prod_{ij}\Gamma(n_{ij}+1)}.\end{aligned} \tag{35}$$

We can now find the Bayes factor relating models M'_1 (equation 33) and M'_2 (equation 35), with no prior preference of either:

$$\frac{p(M'_2|D)}{p(M'_1|D)} = \frac{p(\overline{n}|M'_2)}{p(\overline{n}|M'_1)} = \\ \frac{\prod_j \Gamma(n_{.j}+1)\prod_i \Gamma(n_{i.}+d_B)}{\Gamma(d_B)^{d_A-1}\Gamma(n+d_B)\prod_{ij}\Gamma(n_{ij}+1)} \tag{36}$$

Consider now a data matrix with three variables, A, B and C (figure 25). The analysis of the model M'_3 where full dependencies are accepted is very similar to M_1 above (equation 29). 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 30) respectively M'_2 and M'_1 (equation 36) for each block of the partition to get the Bayes factors sought:

$$\frac{p(M_4|D)}{p(M_3|D)} = \frac{p(\overline{n}|M_4)}{p(\overline{n}|M_3)} = \\ \frac{\Gamma(d_A)^{d_C}\Gamma(d_B)^{d_C}}{\Gamma(d_A d_B)^{d_C}}\prod_c \frac{\Gamma(n_{..c}+d_A d_B)\prod_j \Gamma(n_{.jc}+1)\prod_i \Gamma(n_{i.c}+1)}{\Gamma(n_{..c}+d_A)\Gamma(n_{..c}+d_B)\prod_{ij}\Gamma(n_{ijc}+1)} \tag{37}$$

and the directed case is similar[56]. The values of the gamma function are rather large even for moderate values of its argument. For this reason the formulas in this section are always evaluated in logarithm form, where products like formula (37) translate to sums of logarithms.

The above analysis takes into account the probability of the data to be generated by the two models. It may give unintuitive results when there are some data values for C that 'obviously' makes A and B dependent, and for these values there are indeed strong evidence in favor of a link between A and B. But the method is quite objective and if other values of C are frequent, then this data-dependent dependency will be considered a 'random accident' by equation (37). There are several recipes to overcome this problem:(i) When substantive qualitative knowledge is available it is motivated to force a link between A and B; (ii) There are more detailed ways to describe dependencies in a multivariate data set, although at present these methods have no really systematic and at the same time practical treatment; (iii) The ultimate method is to go into the use of the model and produce it using a formal utility optimization. Once quality of model output can be quantified, cross-validation is a possible practical approach (Ch. 2.1.12).

Exercise 24

Derive a formula similar to (37) for the directed case, $p(M'_4|D)/p(M'_3|D)$

3.3.6 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 27 is from

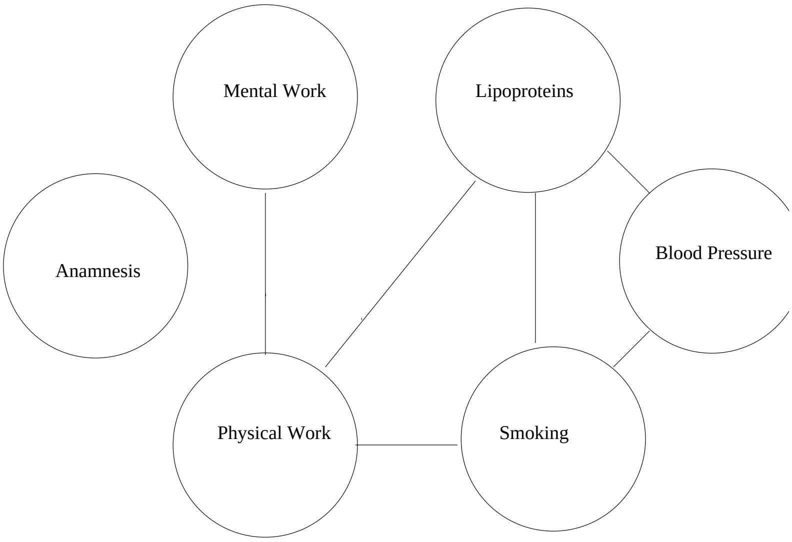

Figure 27: Symptoms and causes relevant to heart problems

[66], 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 $\{c_1, \dots c_d\}$. If two models G and G' differ only by the 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.

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 [84], Rose, Lueker and Tarjan[85]). 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 by edges. 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:

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

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).

The above finds a high-probability graph that may be globally optimal and similar to all high-probability graphs, but this is not always the case. It is possible to get a posterior distribution over the graph by using an MCMC style approach: Each time the possibility of adding or deleting an edge is contemplated, the acceptance criterion of MCMC is used (i.e., accept a proposal that improves the posterior probability, and accept other proposals by randomization depending on the Bayes factor). We may still have convergence problems, however.

For each graph found 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[67]), 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 know that we will find all high probability graphs. For directed graphical models, a similar method of obtaining high probability graphs is known as the K2 algorithm[9].

3.3.7 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. A quite useful and robust method is to discretize the variables. This is equivalent to assuming that their probability distribution functions are piecewise constant. Discretized variables can be treated as categorical variables, by the methods described above. The method wastes some information, but is quite simple and robust. Typically, the granularity of the discretization is chosen so that a reasonably large number of observations fall in each level. It is also possible to assign an observation to several levels, depending on how close it is to the intervals of adjacent levels. This is reminiscent of linguistic coding in fuzzy logic[110]. Discretization does however waste information in the data matrix. It is possible to formulate the theory of section 3.3.5 using inverse Wishart distributions as conjugate priors for multivariate normal distributions, but this is leads to fairly complex formulas and is seldom implemented. Since most practically occurring distributions deviate a lot from normality, it is in practice necessary to model the distributions as mixtures of normal distributions which leads to even higher conceptual complexity. A compromise between discretization and use of continuous distributions is analyses of the *rankings* of the variables occurring in data tables. When considering the association between a categorical and a continuous variable one would thus investigate the ranks of the continuous variable, which are independently and uniformly distributed over their range for every category if there is no association. Using a model where the ranks are non-uniformly distributed (*e.g.* with a linearly varying density), we can build the system of model comparisons of section 3.3.5. The difficulty is that the nuisance parameters cannot be analytically integrated out, so a numerical quadrature procedure must be used.

3.4 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[81], 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). A third 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 $p_0 = n_{00}/n_{0.}$. 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.

Assuming that data is missing at random, it is relatively easy to get an adequate analysis. It is not necessary to waste entire cases just because they have a missing item. Most of the analyses made refer only to a small number of columns, and these columns can be compared for all cases that have no missing data in these particular columns. In this way it is, *e.g.*, possible to make a graphical model for a data set even if every case has some missing item, since all computations of section 3.3.5 refer to a small number of columns. In this situation it is even possible to impute the values missing, because the graphical model obtained shows which variables influence the missing one most. So every missing value for a variable can be guessed by predicting it from values of the case for neighbors to the variable in the graph of the model. When this is done, one must always remember that the value is guessed. It can thus never be used to create a formal significance measure - that would be equivalent to using the same data twice, which is not permitted in formal inference.

The method of imputing missing values has a nice formalization in the Expectation Maximization (EM) method: This method is used to create values for missing data items by using a parameterized statistical model of the data. In the first step, the non-missing data is used to create an approximation of the parameters. Then the missing data values are defined (given imputed values) to give highest probability to the imputed data matrix. We can then refine the parameter estimates by maximization of probability over parameters with the now imputed data, then over the missing data, *etc.* until convergence obtains. This method is recommended for use in many situations despite the fact that it is not strictly Bayesian and it violates the principle of not creating significance from guessed (imputed) values. It is important to verify that data are really missing at random, otherwise the distribution of missing data values cannot be inferred from the distribution of non-missing data. A spot check that is easy to perform is to code a variable with many missing items as a 0/1 indicator column of missingness, and check its association to other columns using equation (30). The most spectacular use of the EM algorithm is for automatic (unsupervised) classification in the AUTOCLASS model (see section 3.5).

Imputation of missing values can also be performed with the MCMC method, by having one variable for each missing value, The trace will give a picture of their joint posterior distribution, and this has the advantage of not creating more significance than justified by the model and the data. The method is strongly recommended over simple imputation by Gelman.

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.

3.4.1 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[72]. 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 practice[105, 15]. The explanation is that the data one wants to classify has usually not been generated in such a way that the smallness of the tree gives a good indication of generalizing power. An example is shown in Figure 28. The distribution of the data is well approximated by two multivariate normal distributions, but since these lie diagonally, the ortho-linear separating planes of the decision tree produced from a moderate size sample fit badly to the distribution and depends a lot on the sample. In this example, decision trees with general linear boundaries perform much better, and in fact the optimal separator of the two distributions is close to the diagonal line that separates the point sets with a largest possible margin(section 2.8.

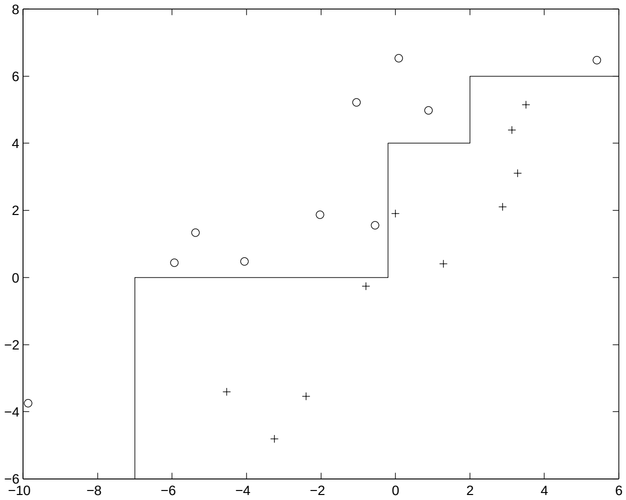

Figure 28: Decision tree with bad generalization power

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[22]) and by (Paass and Kindermann[75]). 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 possibly 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,

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

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 $(n_{i''j} + n_{i'j} = n_{ij})$:

$$\frac{p(D|T')}{p(D|T)} = \frac{\Gamma(n_{i.}+c)}{\Gamma(n_{i'.}+c)\Gamma(n_{i''.}+c)} \prod_j \frac{\Gamma(n_{i'j}+1)\Gamma(n_{i''j}+1)}{\Gamma(n_{ij}+1)} \tag{38}$$

A reasonably well generalizing decision tree can now be obtained top-down recursively by selecting the variable and decision boundary to maximize 38 until the maximum is less than 1, where we make a terminal deciding the majority of the labels of training cases falling in the box. This is a slight generalization of the 'change-point' problem describe in section 3.1.4.

3.5 Clustering and Mixture Modelling

Clustering is not a central topic in this course, but it is so closely related to mixture modeling and segmentation that a short discussion seems appropriate. Clustering is a practice-rich but theory-poor activity where one tries to find groups of points such that the distances between points are small within clusters and large between clusters. More correctly, one may say that the points of a cluster are closer to the cluster center than points in other clusters. With this view, clustering is apparently a form of mixture modeling where the components are characterized by a cluster center and a pdf that depends on the distance measure used - the density is a decreasing function of the distance to the center. In clustering, however, we are normally not interested in the interpretation in terms of mixtures and their posteriors, but the classes obtained by an algorithm are taken as 'true' - a kind of MAP estimation. The result of clustering must always be discussed directly in application terms, and quite a lot of work in preprocessing and defining the distance function is typically required before results are presentable to application owners.

The main algorithms for clustering are K-means and hierarchical clustering. K-means has the flavor of Expectation Maximization: The number of clusters and a distance function is determined, tentative cluster centers are determined, and then two refinement steps are iterated until convergence:

1. Assign each data point to the closest cluster center.
2. Recompute the cluster centers.

In hierarchical clustering one determines a method to compute the distance between two clusters (for one-point clusters, the point distance is used). Then one iteratively determines the two closest clusters and merges them. This process ends, with n data points, after $n-1$ steps, when there is only one cluster left. A tree representation of the merging process, and goodness measures for the clusterings on the way, can be used visually to guess which is the 'best' clustering proposal encountered on the way.

A considerable weakness with clustering compared to Gaussian mixture modeling is that components that overlap cannot be found. An extreme example is given in figure 29, where the two normal distributions are easily found with MCMC, but not with any distance-based clustering method.

On the other hand, if we want to model the data as a mixture of multivariate Gaussian distributions, we would write the probability for a given assignment of points to mixture components with known variances and means as follows:

$$p(\overline{x},\overline{\lambda},\overline{\mu},\overline{\Sigma},\overline{c}) = \prod_i \lambda_{c_i}(2\pi)^{-d/2}|\Sigma_{c_i}|^{-1/2}\exp\left(-\frac{1}{2}(x_i-\mu_{c_i})^T\Sigma_{c_i}^{-1}(x_i-\mu_{c_i})\right).$$

We can group the factors in the above product by mixture component to which data points are assigned, so $x_i^{(c)}$ is a sequence of data points for component c and n_c is the number of points assigned to component c:

$$(2\pi)^{-nd/2}\prod_k \lambda_k^{n_k}|\Sigma_k|^{-n_k/2}\exp\left(-\frac{1}{2}\sum_i(x_i^{(k)}-\mu_k)^T\Sigma_k^{-1}(x_i^{(k)}-\mu_k)\right).$$

The simple Rao-Blackwellization used for the univariate Gaussian mixture modeling (see Exercise 19) can not immediately be applied to the multivariate case since it is quite demanding to find a prior for the variance matrix of each component and to integrate it out as a nuisance parameter. However, with a little cheating we can integrate out the variance matrix. We do this cheating by assigning a prior on Σ_k that depends on the points actually assigned to component k. We prefer to do our integrals by only matching distributions and normalization constants for the beautiful but difficult to integrate distributions in classical tables. From a table, e.g., [16, App A] or [45], we find the *Inverse Wishart distribution* which is parametrized by a d by d positive definite and symmetric scale matrix S and a number of degrees of freedom ν:

$$p(W) = \left(2^{\nu d/2}\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma(\frac{\nu+1-i}{2})\right)^{-1}\times|S|^{\nu/2}|W|^{-(\nu+d+1)/2}\times\exp\left(-\frac{1}{2}\mathrm{tr}(SW^{-1})\right).$$

In order to integrate out a nuisance parameter, we must identify the exponentials in the two last formulas, and we must also identify Σ_k with W. This is done by first identifying $\mathrm{tr}(SW^{-1})$ with $\sum_i(x_i^{(k)}-\mu_k)^T\Sigma_k^{-1}(x_i^{(k)}-\mu_k)$ giving $S_k = q_k - \mu_k^T s_k - s_k^T\mu_k + n_k\mu_k^T\mu_k$, where $q_k = \sum x_i^{(k)}x_i^{(k)^T}$ and $s_k = \sum x_i^{(k)}$. We can now perform the integration over the covariance matrix, and the inverse normalization constant for the Inverse Wishart distribution is obtained by identifying the powers of $|W|$ leading to $\nu_k = n_k - d - 1$.

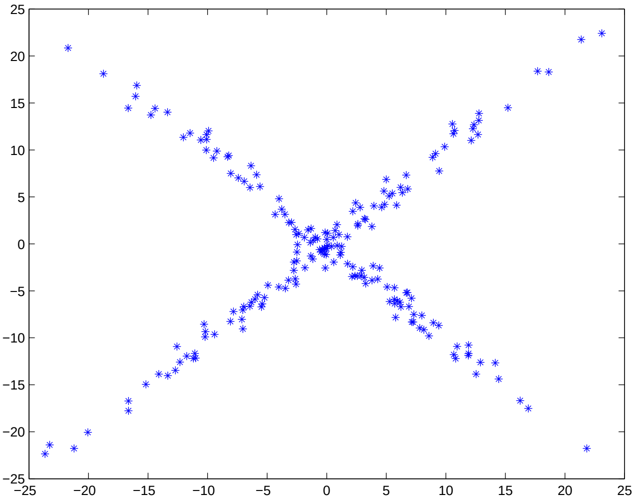

Figure 29: An extreme example of two overlapping Gaussian components in 2D. The components are readily identifiable with mixture modeling, but not with standard clustering methods since the geometrical distance concepts used in clustering do not capture the components.

3.6 Markov Random Fields and Image segmentation

A quite useful statistical model for Bayesian modeling and inference is the Markov Random Field(MRF). In the discrete version (that is representable on a finite computer) it is a set of dependent random variables located on the vertices of a *neighborhood graph.* The joint distribution of the variables in a Markov random field can be defined using a set of distributions over its closed neighborhood sets (each such set consists of a node in the neighborhood graph together with its neighbors), such that whenever two such sets intersect, the marginal distributions over the intersection is the same for the two sets. The defining information for the random properties of the field can also be expressed with, for each node, its distribution conditional on the values of its neighbors. With these conditional distributions it is also possible to simulate it, i.e., to generate a large sample from the field. The MRF is actually the same family of statistical distributions defined by directed graphical models. However the name MRF is typically used when the graph is large but has a known structure, whereas graphical model and Bayesian network are the names used when the graphs are small or have unknown neighborhood relations. When graphs are dense we cannot, for example, used the decomposability methods because the graphs cannot be decomposed by small separators. The most prominent case of this is a rectangular grid which, despite every vertex having just four neighbors, cannot be separated in two equally large parts with a small separator.

The family of distributions known as *Gibbs Random Fields*, *GRF*, are defined using the same formula as the state distribution for a physical system: $P(\bar{\lambda}) \propto \exp(-E(\bar{\lambda}))$, where E is the energy, a function of the total state $\bar{\lambda}$. For a system with a neighborhood graph, the state is the vector of vertex states and the energy decomposes additively into functions of the state projected on the cliques C of the graph. In other words, there is an energy function E_c on each completely connected part $c \in C$ of the neighborhood graph, and E_c is a function of the states of its vertices, $\bar{\lambda_c}$. The distribution of a GRF is thus:

$$P(\bar{\lambda}) \propto \exp(\sum_{c \in C} E_c(\bar{\lambda}_c).$$

The following is the central characterization theorem of a MRF:

Proposition 4 (Hammersley, Clifford) *For an undirected graph G, a distribution of its vertices is a MRF of G if and only if it is a GRF of G.*

The implication of this theorem is that a MRF distribution has a simple characterization in terms of an energy function that falls apart in small parts. This has been used in one of the major applications of MRF, image analysis. In many imaging instruments, the signal cannot be robustly transformed into a meaningful image because the noise is too high and the image is underdetermined on a voxel/pixel basis. The solution to this problem is to apply a MRF prior as a regularization device. An example of this method is the SPECT image reconstruction described by Green[52].

SPECT and PET cameras aim to reconstruct the distribution in an organ of a radioactive substance, which is in itself a biochemical compound that binds to certain molecules in the organ, like oxygenized blood (haemoglobin) or molecules participating in the signaling system of the brain. Say our aim is to reconstruct, on a voxel basis, the concentration of radioactivity in the organ, x_s, where we use a one-dimensional indexing system (a la Matlab). The organ is surrounded by a set of detectors, and the coefficients a_{sd} measure the effective spatial angle from a voxel s that leads into a detector d. Many different camera geometries exist, like an array of detectors moving around the scene(organ), and determining the coefficients is a non-trivial matter supported by drawings of the camera and calibration by means of 'phantoms', synthetic scenes with known geometry.

The signal reaching detector d from voxel s is thus proportional in some sense to a_{sd}. specifically, the detector sees a Poisson variable with the intensity obtained by summing over the scene, $\text{Poisson}(\sum_{s\in S} a_{ds}x_s)$. The corresponding likelihood function is obtained from the registered counts at each detector, where $\lambda(x_s) = \sum_{s\in S} a_{ds}x_s$:

$$\log(P(Y|x_s)) = \sum_d -\log(y_d!) + \log(\lambda(y_d)) * y_d - \lambda(y_d)$$

The prior will inevitably have a certain ad hoc character, and in this case, since intensities are always non-negative, is usually on a logarithmic scale, with a scale factor γ that can be fixed or subject to inference, and a non-linear scaling ϕ with $\phi(u) = \delta(1-\delta)\log\cosh(u/\delta)$:

$$\log(P(X)) = -\sum_{s\sim r} \phi(\gamma(\log x_s - \log x_r)).$$

The MCMC procedure is repeatedly proposing changes in values of the intensities x_s. Each proposal is tested and conditionally accepted. It is important to structure the computation so that only those terms that change are recomputed, the remainder being stored and remembered. The parameters γ and δ can also be subject to inference, taking account of what is known about the film recording process. The simple structure of the computation makes it easy to adapt to new circumstances, such as, in the case of PET, modern sensors that can get an estimate of the location of annihilation by precisely measuring the time difference for detection in the two sensors.

The type of regularizing prior useful for noisy PET images has been used to good results in numerous applications with a large range of likelihood functions, like medical imaging with X-ray, MR and light, and for reconstructing the atmosphere using spectral absorption from stars photographed through it from satellites[54].

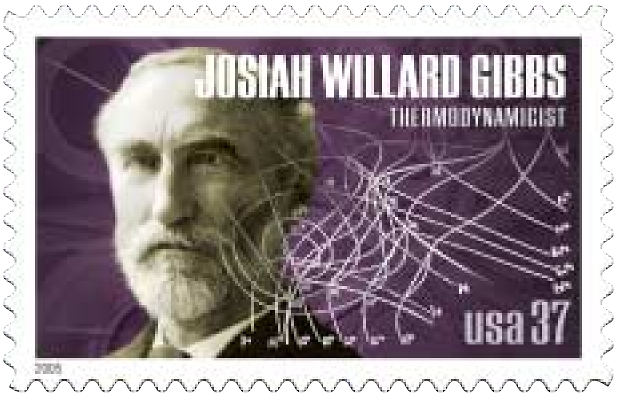

Figure 30: Josiah Willard Gibbs (1839-1903) founded and created many tools of thermodynamics and statistical mechanics. Several of these are used or adapted to use in applied statistics, sometimes mediated by Shannon's work in information theory and Jaynes' Maximum Entropy ideas. Quite many ideas from theoretical physics have been introduced in Computer Science and statistics, but it typically takes some time to find out how applicable they are and to give them a goal-oriented motivation in computer applications. The Gibbs distribution, a fundamental concept in statistical mechanics, is also useful in the analysis of directed graphical statistical models. It provides a simple characterization of the corresponding statistical distribution

3.7 Regression and the linear model

The most used model in estimation is regression, a model where batches of similar experiments are used to form inference on effects. One variable (the response) is assumed to be a linear function of a set of explanatory variables, with random noise added. The parameter is the set of coefficients of the linear function, and it is often of interest to decide if one of these, or a linear function of them, is different from zero. The probability model for the data is thus

$$y = X\beta + \epsilon, \tag{39}$$

where each component of ϵ is assumed to be from a common normal distribution with zero mean, with known or unknown variance. When an estimate $\hat{\beta}$ is available, the *residual vector* is $r = y - X\hat{\beta}$. The likelihood is the product of the probabilities of the residuals . The probability density for a batch is a function of the (Euclidean) norm of the residual vector, and the ML estimates of the coefficients are given by the least squares method which minimizes the (squared) length of the residual vector: $\hat{\beta} = \text{argmin}_\beta |y - X\beta|^2$, a quantity easy to compute by solving the normal equation: $\hat{\beta} = (X^TX)^{-1}X^Ty$, where ^{T} denotes transposition. It is easy to see that X^TX is invertible if the co-variates are linearly independent (and thus no more than the number of samples). If X^TX is not invertible, the coefficients are not uniquely estimable, and it is not possible, e.g., to test one of them for zero. However, a beautiful theory developed by Sheffé and others shows that it is possible to test any linear hypothesis $K\beta = m$, where the rows of K are linear combinations of the rows of the matrix of explanatory variables X. In other words, K must be expressible as AX, for some matrix A. A possible test statistic is y^TGy where $G = K^T(KK^T)^-K$, and where $^-$ is any generalized inverse, i.e., a matrix operator with $XX^-X = X$ for all X. The test statistic is distributed as $\sigma^2\chi^2_s$, where s is the rank of G. So testing depends on the often unknown σ^2. In the univariate case it is not very dangerous to estimate σ^2. But the dependence on σ^2 is better removed by using the test statistic $y^TGy/y^T(I-G)y$, which has an $F(s, n-s)$ distribution (this was pointed out, for functional neuro-science studies, in[65]). And of course the most severe problem is that we do not normally have any proof that the relationship between response and explanatory variables is linear - indeed, it is typically not, even when regression models seem to give good results. The response is to visualize the residuals or test them for normality.

When several (p) responses are considered we have *multivariate regression* and (39) is changed to

$$Y = XB + E, \tag{40}$$

where Y and B are now $n \times p$ matrices and each row of the $n \times p$ matrix E is from a multivariate normal distribution with zero mean and known covariance Σ. Many methods in univariate regression are specializations of methods in multivariate regression, like the estimation and testing methods described above. A possible use of multivariate regression is to make a joint regression on all N voxels in an experiment. Since Σ is not known, it has to be estimated, and since it has $N(N-1)/2$ different elements, it is difficult to estimate it with good confidence.

3.8 Bayesian regression

Many of the concepts in classical regression have their counterparts in Bayesian regression. The probability model is normally the same, except that one should have a prior for the coefficient set and for the noise variance. Estimation can be made directly for the coefficients even if X is rank-deficient, but of course rank-deficiency indicates that the coefficients get larger uncertainties, and in general their estimates and credible intervals will depend a lot on the priors. A particularly simple prior is a (improper) uniform prior for $(\beta, \log\sigma^2)$. Then the posterior for β is, see, e.g., [45],

$$\begin{aligned}
& \mathcal{N}(\beta; \hat{\beta}, V_\beta\sigma^2), && (41)\\
V_\beta = {} & (X^T X)^{-1} && (42)\\
\sigma^2 \sim {} & \text{Inv}-\chi^2(\sigma^2; n-m, s^2), && (43)\\
s^2 = {} & (y - X\hat{\beta})^T(y - X\hat{\beta}), && (44)
\end{aligned}$$

if X is $n \times m$. Since σ^2 is a random variable, this is not a normal distribution. The marginal distribution of the coefficients is not a normal distribution but the (in general) more vague multivariate t, with parameters mean $\hat{\beta}$, variance $s^2V_\beta/(n-m)$ and $n-m$ degrees of freedom. The above holds only for X of full rank, however, and one of the strengths of Bayesian regression in handling large sets of explanatory variables will only emerge after a proper prior distribution has been assumed for the coefficient set and covariance of ϵ.

In Bayesian applications it is usually required to find the probability of data given the parameters. This quantity (modulo a normalization constant) is, for the Bayesian regression model,

$$\begin{aligned}
p(Y|X, \beta, \Sigma) \quad & = \mathcal{N}(Y; X\beta, \Sigma) = && (45)\\
& c\exp(-\tfrac{1}{2}(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)) && (46)
\end{aligned}$$

The normalization constant c is $1/\sqrt{(2\pi)^n|\Sigma|}$.

A comprehensive analysis of Bayesian regression and related methods, with quite useful Matlab code, can be found in [33].

4 Approximate analysis with Metropolis-Hastings simulation

We have seen several simple examples of the Markov Chain Monte Carlo approach, and we will now explicate in detail their usages and mathematical justifications – as well as their limitations.

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[73]. 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 significant 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, as mentioned before, organized as follows: given the pdf $p(x)$ of a state variable x over a state space X, and an essentially arbitrary symmetric proposal function $q(x, x')$, a sequence of states is created in a Markov chain. In state x, draw x' according to the proposal function q. If $p(x')/p(x) > 1$, let x' be the new state. Otherwise, let x' be the new state with probability $p(x')/p(x)$, otherwise keep state x. We will in the next subsection verify that $p(x)$ 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. Recent developments aim at finding MCMC procedures which can detect when an unbiased sample has been obtained by using various variations of 'coupling from the past'[78]. These methods are however still complex and difficult to apply on the classification and graphical model problems.

Nevertheless, there have been great successes with MCMC methods 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 [34, 35, 37, 82].

4.1 Why MCMC and why does it work?

The main practical reason for using MCMC is that we often have an unnormalized probability distribution over a sometimes complex space, and we want a sample of it to get a feel for it, or we might want to estimate the average of a function on the space under the distribution. This section is a summary of [13, Ch 6], where a more complete account can be found. A typical way one obtains a distribution which is not normalized is by using (3), where the product on the right side is not a normalized distribution and is often impossible to normalize using analytical methods (even in the case where we have analytical expressions for the prior and the likelihood). So assume we have a distribution $\pi(x) = cf(x)$, with unknown normalization constant c, over a space X where we can evaluate f in every point of X (we will think of cf as a probability density function - a thorough measure-theoretic presentation is also possible, but this simplified view is appropriate in most applications). Given a set of N independent and identically distributed examples x_i with pdf $\pi(x)$, we can estimate the expected value of any function $v(x)$ over π by taking the average

$$f_N = \sum_i v(x_i)/N. \tag{47}$$

This is a well-known technique known as Monte Carlo estimation. Assuming that X and π are reasonably well-behaved, like a Euclidean n-dimensional space, the expected value is $I = \int_X \pi(x)v(x)\mathrm{d}x$ and the estimate f_N converges to I with probability 1 ('almost surely'). Moreover, if the variance of v can be bounded, the central limit theorem can be used to give a convergence estimate, namely that $\sqrt{N}(f_N - I)$ stays bounded (also with probability one) and indeed approaches the normal distribution with zero mean and the same variance as v.

It is not easy in general to generate an iid sample for f where its variation is large and irregular over X. If an upper bound M of $f(x)$ is known, we can use *rejection sampling.* In rejection sampling we generate the samples according to some distribution $g(x)$, where the support of g covers the support of f (in other words, g can generate any point with non-zero probability according to cf), and keep a subset of them (i.e., we reject some of them). Specifically, if a sample x' was generated, we keep it with probability $f(x')/(Mq(x))$. The set of kept samples will be distributed as $cf(x)$. Obviously, this method is practical only when the bound M is reasonably accurate and when $g(x)$ is a reasonable approximation of $cf(x)$, and if this is not true the rejection probability can be very close to 1 most of the time.

Another technique to obtain a sample from an irregular distribution is *importance sampling.* Here we do not get a simple sample but a *weighted sample*, i.e., each point is equipped with a weight and represented by a pair (x_i, w_i). If we can produce a sample of the (preferably close to cf) distribution g, then setting $w_i = f(x_i)/g(x_i)$ and normalizing the weights so that they sum to one, produced a weighted sample that corresponds to the distribution cf. The weighted sample can be used directly for estimation, changing (47) to $f_N = \sum_i v(x_i)w_i$. The weighted sample can also be changed to an unweighted sample by *resampling.* An unweighted sample of M elements is obtained by drawing, M times, an integer j between 1 and N, according to the probability distribution w, and selecting the corresponding sequence of x_j. Obviously, the information about f is filtered away in the sampling and resampling steps and the accuracy is less than with iid sampling according to cf. Typically, M should be at least $10N$ for obtaining reliable results.

The MCMC method produces a sequence (x_i) distributed asymptotically as cf, where c is unknown. It is based on a combination of the above sampling ideas, and it produces a sample in the form of a sequence of non-independent points with stationary (i.e., limiting) distribution $\pi(x) = cf$. We can use equation (47) for this sample, but only if it is long enough to remove the effect of local autocorrelation of the x_i.

Basically, a Markov Chain is represented with a transition kernel $K(x_{t-1}, x_t)$ giving the conditional distributions $p(x_t|x_{t-1})$. We want to design the transition kernel so that the sequences it produces have a stationary distribution $\pi(x) = cf(x)$, where we can evaluate f but not c (and thus not π either). In others words, $f(x) = \int_X K(x_{t-1}, x) f(x_{t-1}) \mathrm{d}x_{t-1}$. We will achieve this by designing the chain to be π-*reversible*, i.e., $K(x, y)f(x) = K(y, x)f(y)$. Reversibility is also called the *detailed balance condition*, a term which better explains why reversibility leads to stationarity. Namely, if we have two disjoint subsets A and B, the integral $\int_A \int_B K(x, y)cf(x)\mathrm{d}x\mathrm{d}y$ is the probability of a sample element in B being followed by one in A, and this is by the detailed balance condition equal to the probability that an element in B is followed by one in A. So if the chain is *ergodic*, i.e., its ensemble distribution is the same as its path distribution, then cf is a stationary distribution of the chain. A sufficient condition for this is that the chain can reach every region in the support of f, i.e., every region with non-zero probability can be reached from every point with non-zero probability. We will not go into the detailed conditions and proofs of convergence of Markov Chains, some more technical detail is given in [13] and a lot in [99, 98].

For simple (finite or of constant dimension) spaces X there is a simple way to construct a Markov Chain that is obviously both irreducible and satisfies the detailed balance condition if a relative (i.e., unnormalized as the function f above) density of the target distribution can be evaluated at any point, and thus has the required stationary distribution. The main problem, which has not yet a simple solution, is to design the chain in such a way that it has reasonably low autocorrelation. Such a chain is said to mix rapidly.

Once state space and target distribution has been defined, the only design choice is the *proposal distribution* $q(x'|x)$, a distribution for a proposed new state x' dependent on the current state x. We design m by giving a procedure both to generate a new proposal x' given current state x (using a random number generator), and another to evaluate the density given x and x'. Using the latter procedure and the unnormalized target density f, which can also be computed for any state, we can compute an *acceptance probability* for a proposed state that depends on the current state:

$$\alpha(x, z) = \min(1, \frac{f(z)q(x|z)}{f(x)q(z|x)}) \tag{48}$$

A trace $(x_0, \ldots)$ is now produce using the following algorithm:

```
Start with any $x_0$ in the support of $f$
for i=1, ... do
generate $z$ by the distribution $q(z|x_i)$
with probability $\alpha(x_i, z)$, set $x_{i+1} = z$,
otherwise set $x_{i+1} = x_i$
```

There are several ways to prove that this gives a chain with stable distribution cf, despite the fact that we do not know the value of c. The only really comprehensible argument can be found, e.g., in [13] and goes as follows:

First we need the useful consequence of (48) that

$$f(x_t)q(x_{t+1}|x_t)\alpha(x_t, x_{t+1}) = f(x_{t+1})q(x_t|x_{t+1})\alpha(x_{t+1}, x_t). \tag{49}$$

Equation (49) is proved by noting that exactly one of $\alpha(x_{t+1}, x_t)$ and $\alpha(x_t, x_{t+1})$ is less than one (unless both are equal to one). In each case, a simple algebraic expression without the minimum operator results from (48), that simplifies to (49) in both cases (as well as in the case where both alphas are one).

We next show that the resulting chain is π-reversible, where $\pi(x) = cf(x)$. The distribution of x_{t+1} given x_t is

$$\begin{aligned} p(x_{t+1}|x_t) = q(x_{t+1}|x_t)\alpha(x_t, x_{t+1}) + \\ \delta(x_{t+1} - x_t)(\int q(z|x_t)(1 - \alpha(x_t, z))\mathrm{d}z. \end{aligned}$$

The first term above comes from the case where a value for x_{t+1} is proposed and accepted, the second term (where δ is Dirac's delta function) has contributions from all possible proposal values z weighted by the probability that the value is generated and also rejected. Multiplying by $\pi(x_t)$ and using (49) gives

$$\begin{aligned} \pi(x_t)p(x_{t+1}|x_t) = \pi(x_t)q(x_{t+1}|x_t)\alpha(x_t, x_{t+1} + \\ \pi(x_t)\delta(x_{t+1} - x_t)(1 - \int q(z|x_t)\alpha(x_t, z)\mathrm{d}z) = \\ \dots \\ \pi(x_{t+1})p(x_t|x_{t+1}), \end{aligned}$$

the detailed balance equation which, together with aperiodicity and irreducibility means that every chain has $\pi = cf$ as a limiting distribution.

Exercise 25 *We believe that a sequence of real numbers, $(y_i)_{i=1}^N$ was generated from the mixture of two normal distributions, $f(y) = \sum_{i=1}^2 \lambda_i \mathcal{N}(\mu_i, \sigma_i)$ where $\lambda_1 + \lambda_2 = 1$.*

Design an MCMC procedure to make inference about the parameters of the normal distributions and the source (1 or 2) of each y_i. Check out what happens when the sample is too small for reliable inference.

4.1.1 MCMC burn-in and mixing

The theorems stating that a MCMC chain converges to an invariant distribution are only asymptotic. In practice some target distribution and proposal function pairs give well behaved chains where it is easy to estimate quantities of interest to a precision that can be reliably assessed. Such chains are said to have *good mixing behavior.*

In other cases there are phenomena in the trace that make it difficult to say that the chain has converged. There are quite many proposals in the literature suggesting different approaches to MCMC convergence assessment.

Kolmogorov-Smirnov's test is originally formulated as a way to test if a given sample has a given distribution. For a real-valued distribution, the cumulative distribution of the given distribution is compared with the cumulative empirical distribution of the sample, and the maximum absolute difference D_n is taken as the test statistic. Both functions compared have the range 0 to 1, and the statistic is thus in the interval $[0, 1]$. This statistic has under the null hypothesis that the sample is actually from the given distribution a distribution that is independent of which the latter distribution is. Moreover, it is asymptotically proportional to $1/\sqrt{n}$ for sample size n, and in a practical sample range it is a good rule of thumb that a statistic $\sqrt{n}D_n$ of 3 has a p-value of circa 0.022, whereas 4 corresponds to p-value 0.001. If the value is above 4 and n is above

2, then the p-value is less than 0.01%. The same idea can be used to test if two different samples can possibly be from the same distribution. Multiplying the largest absolute difference between the two cumulative distributions by $\sqrt{\min(n_1 n_2)}$ gives a fairly sample-size independent statistic. No matter what the sample sizes are, the hypothesis that the two samples are from the same distribution can be rejected at the 99% level if the computed statistic is 2.2 or more. The Kolmorov-Smirnov test is fairly strong. A major use today is for checking the convergence status of an MCMC chain.

The Matlab code for computing the statistic is as follows, where the test for a d-dimensional variable checks the marginal distribution along each dimension (checks for proper correlations can be obtained by preceding the test by a random rotation):

```
function stat=KS(tr1,tr2);
% Kolmogorov-Smirnov test
% stat=KS(tr1,tr2);
% tr1 and tr2 are d by n matrices
% containing two (not necessarily equal-sized)
% samples from d-dimensional distribution
% Output is d-vector of (scaled) Kolmogorov-Smirnov statistics,
% for any size samples. In each test, a statistic of 2.4
% corresponds to p-value > 0.005. If significantly more than 1 test
% in 1000  is larger than 2.4, the samples are probably not from
% the same distribution. The p-values are:
% p      stat
% 0.1    1.72
% 0.05   1.91
% 0.01   2.27
% 0.001  2.70
% 0.0001 3.20(?)
%
d=size(tr1,1);
n1=size(tr1,2); n2=size(tr2,2);
cums=[[tr1' ones(n1,1)/n1];[tr2' -ones(n2,1)/n2]];
stat=[];
ord=[ones(n1,1)/n1; -ones(n2,1)/n2];
tr=[tr1;tr2];
[tr,IX]=sort(tr);
[tr,IX]=sort(IX);
OM=ord(IX);
xx=sqrt(min(n1,n2))*cumsum(OM);
ix=find(xx(1:end-1,:)==xx(2:end,:));
xx(ix)=0;
stat=max(abs(xx));
```

There are several uses of the KS test in MCMC. One example is the breakpoint test of section 3.1.5, where of course a rejection of the hypothesis of uniform accident intensity can be followed by uniformity tests on parts of the time axis or other non-linear distribution tests. An even more important usage is for MCMC diagnostics, assessing the number of elements to reject at the beginning of a chain (burn-in) and the number of trace points required to estimate a quantity of interest to given precision.

An example of the use of the KS test in MCMC is shown in Figure 31. For four different chains of 10000 sample points of dimension 68, blocks of 100 and 200 were compared for empirical distribution[106]. A KS-diagnostic above 3.5 is a test for difference between the distributions. starting numbers of blocks are coordinates, and a red dot is a pair of blocks with length 100, black of length 200, with a low KS statistic (below 3.5). The chain 'long' has apparently two persistent modes indicated by white strips in the figure, the chain 'short' mixes slowly, and 'conv' mixes well. The chain 'X' is a control chain, generated with random numbers and should have no autocorrelation (which it also seems not to have). There are less drastic diagnostics for MCMC burn-in and convergence oriented to the case where there is a particular quantity of interest that the analysis wants to estimate, see *e.g.,* [106, 47].

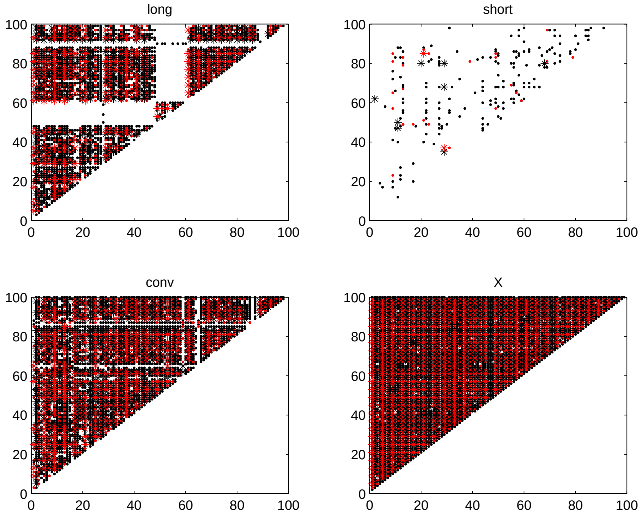

Figure 31: Diagnostics using KS test on four chains, 10000 points, 67 dimensional. Pairs of blocks of length 100 (red dot) and 2000 (black dot) are tested, starting coordinates for blocks in pair is the coordinate of the dot. The dot is shown where KS diagnostic is below 3.5. The first chain has two modes that switch slowly, the second has not at all converged, the third mixes well (many dots). The fourth is synthetic and has no autocorrelation, gives the appearance of good mixing.

4.1.2 The particle filter

The *particle filter* is a computation scheme adapting MCMC to the recursive and dynamic inference problems (5,6). It is described concisely in [13, Ch 6] and a state-of-the-art survey is given in [36]. In the first case (5), we try to estimate a static quantity and expect to get better estimates as more data flows in. In the second case (6), we try to track a moving target. The method maintains a sample that is updated for each time step and can be visualized as a moving swarm of particles. The sample represents our information about the posterior. The particle filter can handle highly non-Gaussian measurement and process noise, and the likelihoods can be chosen in very general ways, the only essential restriction being that relative likelihoods can be computed reasonably efficiently, and that process states can be simulated stochastically from one time step to the next. It is not necessary to have constant time steps, so jittered observations can be easily handled provided the measurement times are known. It is not certain that the method performs better than sophisticated non-linear numerical methods, the attractivity of the methods stems more from ease of implementation and flexibility.

In the steady state our inference of the state at time t is represented by a set of particles $(x_i^{(t)})_{i=1}^N$. Each particle is brought to time $t+1$ by drawing state $u_i^{(t+1)}$ using the process kernel $f(u_i^{(t+1)}|x_i^{(t)})$. In practice, we produce many (say 10) $u_i^{(t+1)}$ from each $x_i^{(t)}$, so that the resampling will produce a better result. Now the data d_{t+1} are considered and used to weight the $u_i^{(t+1)}$ particles: $w_i^{(t+1)} = f_{t+1}(d_{t+1}|u_i^{(t+1)})$. In order to get back to the situation at time $t+1$, we now use the weights to draw N indices $(i_1, \ldots, i_N)$ from the probability distribution defined by the (normalized) weights. Finally, the set of particles defining our belief of system state at time $t+1$ is the set $(x_{i_j})_{j=1}^N$.

4.2 Basic classification with categorical attributes

For data matrices with categorical attributes, the categorical model of Cheeseman and Stutz is appropriate. We assume that rows are generated as a finite mixture with a uniform Dirichlet prior for the component (class) probabilities, and each mixture component has its rows generated independently, according to a discrete distribution also with a uniform Dirichlet prior for each column and class. Assume that the number of classes C is given, the number of rows is n and the number of columns is K, and that there are d_k different values that can occur in column k. For a given classification, the data probability can be computed; let $n_i^{(c,k)}$ be the number of occurrences of value i in column k of rows belonging to class c. Let $x_i^{(c,k)}$ be the probability of class c having the value i in column k. Let $n_{\bar{i}}^{(c)}$ be the number of occurrences in class c of the row $\bar{i}$, and $n^{(c)}$ the number of rows of class c. By (19) the probability of the class assignment depends only on the number of classes and the table size, $\Gamma(n+1)\Gamma(C)/\Gamma(n+C)$. The probability of the data in class c is, if $\bar{i} = (i_1, \ldots, i_k)$:

$$\int \prod_{k\bar{i}} (x_{i_k}^{(c,k)})^{n_{\bar{i}}^{(c)}} \mathrm{d}\overline{x} = \prod_{k=1}^{K} \int \prod_{i=1}^{d_k} (x_i^{(c,k)})^{n_i^{(c,k)}} \mathrm{d}\overline{x}^{(c,k)}$$

The right side integral can be evaluated using the normalization constant of the Dirichlet distribution, giving the total data probability of a classification:

$$\frac{\Gamma(n+1)\Gamma(C)}{\Gamma(n+C)} \prod_{c,k} frac \prod_i \Gamma(n_i^{(c,k)}+1)\Gamma(n^{(c)}+d_k)$$

The posterior class assignment distribution is obtained normalizing over all class assignments. This distribution is intractable, but can be approximated by searching for a number of local maxima and estimating the weight of the neighborhood of each maximum.

Here the *EM algorithm* is competitive to MCMC-calculations in many cases, because of the difficulty of tuning the proposal distribution of MCMC-computations to avoid getting stuck in local minima. The procedure is to randomly assign a few cases to classes, estimating parameters $x_i^{(c,k)}$, assign remaining cases to optimum classes, recomputing the distribution of each case over classes, reclassifying each case to optimum class, and repeating until convergence. Repeating this procedure one typically finds after a while a single most probable class assignment for each number of classes. The set of local optima so obtained can be used to guide a MCMC simulation giving more precise estimates of the probabilities of the classifications possible. But in practice, a set of high-probability classifications is normally a starting point for application specialists trying to give application meaning to the classes obtained.

5 Two cases

5.1 Case 1: Individualized media distribution - significance of profiles

The application concerns individualized presentation of news items(and is currently confidential in its details). The data used are historical records of individual subscribers to an electronic news service. The purpose of the investigation is to design a presentation strategy where an individual is first treated as the 'average customer', then as his record increases he can be included in one of a set of 'customer types', and finally he can also get a profile of his own. Several methodological problems arise when the data base is examined: customers are identified by an electronic address, and this address can sometimes be used by several persons, either in a household or among the customers of a store or library where the terminal is located. Similarly, the same customer might be using the system from several sites, and unless he has different roles on different sites, he might be surprised when finding has his profile is different depending on where he is. After a proposed switch to log-in procedures for users, these problems may disappear, but on the other hand the service may also become less attractive. An individual can also be in different modes with different interests, and we have no channel for a customer to signal this, except by seeing which links he follows. On the other hand, recording the access patterns from different customers is very easily automated and the total amount of data is overwhelming although many customers have few accesses recorded. The categorization of the items to be offered is also by necessity coarse. Only two basic mechanisms are available for evaluating an individuals interest in an item: To which degree has he been interested in similar items before, and to which degree have similar individuals been interested in this item? This suggest two applications of the material of this chapter: Segmentation or classification of customers into types, and evaluating a customer record against a number of types, to find out whether or not they can confidently be said to differ. We will address these problems here, and we will leave out many practical issues in the implementation.

The data base consists of a set of news items with coarse classification (into news, sport, culture, entertainment, economy and with a coarse 'hotness indicator' and a location indicator of local, national or international, all in the perspective of Stockholm, Sweden); a set of customers, each with a type, a 'profile' and a list of rejected and accepted items; and a set of customer types, each with a list of members. Initially, we have no types or profiles, but only classified news items and the different individuals access records. The production of customer types is a fairly manual procedure: even if many automatic classification programs can make a reasonable initial guess, it is inevitable that the type list will be scrutinized and modified by media professionals – the types are of course also used for direct advertising. The AUTOCLASS model has the attractive property that it produces classes where the attributes are independent: we will not get a class where a customer of the class is either interested in international sports events or local culture items, but not both. In the end, there is no unique objective criterion that the type list is 'right', but if we have two proposed type lists we can distinguish them by estimating the fraction of accepted offers, and saying that the one giving a higher estimate is 'better'. In other words, a type set giving either high or low acceptance probability to each item is good. But this can be easily accomplished with many types, leading to impreciseness in assigning customers to types – a case of over-fitting.

The assignment of new individuals to types cannot be done manually because of the large volumes. Our task is thus now to say, for a new individual with a given access list, to which type he belongs. The input to this problem is a set of tables, containing for each type as well as for the new individual, the number of rejected and accepted offers of items from each class. The modeling assumptions required are that for each news category, there is a probability of accepting the item for the new individual or for an average member of a type. Our question is now if these data support the conclusion that the individual has the same probability table as one of the types, or if we can say that he is different from every type (and thus should get a profile of his own).

We can formulate the model choice problem by a transformation of the access tables to a dependency problem for data tables that we have already treated in depth. For a type t with a_i accepts and r_i rejects for a news category i, we imagine a table with three columns and $\sum(a_i + r_i)$ rows: a 1 in column 1 to indicate an access of a type, the category number i in column 2 of $(a_i + r_i)$ rows, a_i of which contain 1 (for accept) and r_i a 0 (for reject) in column 3. We add a similar set of rows for the access list of the individual, marked with 0 in column 1. If the probability of a 0 (or 1) in column 3 depends on the category (column 2) but not on column 1, then the user cannot be distinguished from the type. But columns 1 and 2 may be dependent if the user has seen a different mixture of news categories compared to the type. In graphical modeling terms we could use the model choice algorithm. The probability of the customer belonging to type t is thus equal to the probability of model $M4$ against $M3$, where variable C in figure 25 corresponds to the category variable (column 2).

Although this is an elegant way to put our problem in a graphical model choice perspective, it is a little over-ambitious and a simpler method may be adequate: the types can in practice be considered as sets of reasonably precise accept probabilities, one for each news category, and an extension of the analysis of the coin tossing experiment is adequate. The customers can be described as an outcome of a sampling, where we record the number of accepted and rejected items. If type i accepts an offer for category j with probability p_{ij}, and a customer has accepted a_j out of n_j offers for category j, then the probability of this is, if the customer belongs to type i, $p_{ij}^{a_j}(1-p_{ij})^{(n_j-a_j)}$. For the general model H_u of an unbalanced probability distribution (the accept probability is unknown but constant, and is uniformly distributed between 0 and 1), the probability of the outcome is $a_j!(n_j-a_j)!/(n_j+1)!$. Taking the product over category, we can now find the data probabilities $\{p(D|H_j)\}$ and $p(D|H_u)$, and a probability distribution over the types and the unknown type by evaluating formula (2). In a prototype implementation we have the following customer types described by their accept probabilities:

category	Typ1	Typ2	Typ3	Typ4
news-int	0.9	0.06	0.82	0.23
news-loc	0.88	0.81	0.34	0.11
sports-int	0.16	0	0.28	0.23
sports-loc	0.09	0.06	0.17	0.21
cult-int	0.67	0.24	0.47	0.27
cult-loc	0.26	0.7	0.12	0.26
tourism-int	0.08	0.2	0.11	0.11
tourism-loc	0.08	0.14	0.2	0.13
entert	0.2	0.25	0.74	0.28

Three new customers have arrived, with the following access records of presented(accepted) offers:

category	Ind1	Ind2	Ind3
news-int	3(3)	32(25)	17(8)
news-loc	1(1)	18(9)	25(14)
sports-int	1(1)	7(2)	7(3)
sports-loc	0(0)	5(5)	6(1)
cult-int	2(2)	11(4)	14(6)
cult-loc	1(1)	6(2)	10(3)
tourism-int	0(0)	4(4)	8(8)
tourism-loc	1(1)	5(1)	8(3)
entert	1(1)	17(13)	15(6)
sum	10	105	110

We compute the data probabilities under the hypotheses that each new customer belongs to either type 1 to 4, or to H_u. Then we compute the probability distribution of the individuals types, under the uniform prior assumption:

	Typ1	Typ2	Typ3	Typ4	H_u
Ind1	0.0639	0.0000	0.0686	0.0001	0.8675
Ind2	0.0000	0.0000	0.0002	0.0000	0.9980
Ind3	0.0000	0.0000	0.0000	0.0000	1.0000

Dropping the H_u possibility, we have:

	Typ1	Typ2	Typ3	Typ4
Ind1	0.4818	0.0000	0.5176	0.0005
Ind2	0.0000	0.0000	1.0000	0.0000
Ind3	0.0000	0.0000	0.9992	0.0008

Clearly, quite few observations are adequate for putting a new customer confidently into a type. The profile of the third individual is significantly not in one of the four types($p(H_u) = 1$). For the first individial, the access record is too short to confidently put him in a class, but the first and third classes fit reasonably.

Actually, the model for putting individuals into types may be inappropriate when considering what the application models: A type should really be considered a set of individuals with a clustered span of profiles, whereas the model we used considers a type to be the average accept profile over the individuals in the type. This is a too precise concept. We can make the types broader in two ways: either by spanning them with a set of individuals and computing the probability that the new individual has the same profile as one of the individuals defining the type, or we can define a type as a probability distribution over accept profiles. The second option is quite attractive, since the type accept probabilities can be defined with a sample that defines a set of Beta distributions of accept probabilities. In graphical modeling terms, we test the adequacy of a type for an individual by putting their access records into a three column table with values individual/type, category of item and accept indicator. The adequacy of the type for the individual is measured by the conditional independence of type and accept given category, evaluated by formula (37), with news category corresponding to variable C. The computation is of course done on the contingency table without actually expanding the data table. The results in our example, if the types are defined by a sample of 100 items of each category, is:

	Typ1	Typ2	Typ3	Typ4
Ind1	0.2500	0.2500	0.2500	0.2500
Ind2	0.0000	0.0000	1.0000	0.0000
Ind3	0.0000	0.0001	0.9854	0.0145

It is now even more clear than before that the access record for individual 1 is inadequate, and that the third individual is not quite compatible with any type. It should be noted that we have throughout this example worked with uniform priors. These priors have no canonic justification but should be regarded as conventional. If specific information justifying other priors is available they can easily be used, but this is seldom the case. The choice of prior will affect the assignment of individual to type in rare cases, but only when the access records are very short and when the individual does not really fit to any type.

5.2 Case 2: Schizophrenia research–hunting for causal relationships in complex systems

This application is directed at understanding a complex system - the human brain. Similar methods have been applied to understanding of complex engineered systems like paper mills, and micro-economic systems. Many investigations on normal subjects have brought immense new knowledge about the normal function of the human brain, but mental disorders still escape understanding of their causes and cures (despite a large number of theories, it is not known why mental disorders develop except in special cases, and it is not known which physiological and/or psychological processes cause them). In order to get handles on the complex relationships between psychology, psychiatry, and physiology of the human brain, a data base is being built with many different types of variables measured for a large population of schizophrenia patients and control subjects. For each volunteering subject, a large number of variables are obtained or measured, like age, gender, age of admission to psychiatric care; volumes of gray and white matter as well as cerebrospinal fluid in several regions of the brain (obtained from MR images), genetic characterization and measurements of concentrations in the blood of large numbers of substances and metabolites. For the affected subjects, a detailed clinical characterization is recorded.

In this application one can easily get lost. There is an enormous amount of knowledge on the relationships and possible significances of these quite many (ca 150) variables in the medical profession. At the same time, the data collection process is costly, so the total number of subjects is very small compared, for example, with national registers that have millions of persons but relatively few variables for each of them. This means that statistical significance problems become important. A test set of 144 subjects, 83 controls and 61 affected by schizophrenia, was obtained. This set was investigated with most methods described in this course, giving an understanding of the strongest relationships (graphical model), possible classifications into different types of the disease, *etc.*. In order to find possible causal chains, we tried to find variables and variable pairs with a significant difference in co-variation with the disease, *i.e.,* variables and tuples of variables whose joint distribution is significantly different for affected person relative to control subjects. This exercise exhibited a very large number of such variables and tuples, many of which were known before, others not. All these associations point to probable mechanisms involved in the disease, which seems to permeate every part of the organism. But it is not possible to see which is the effect and what is the cause, and many of the effects can be related to the strong medication given to all schizophrenia patients. In order to single out the more promising effects, a graphical model approach was tried: Each variable was standardized around its mean, separately for affected and controls. Then the pairs of variables were detected giving the highest probability to the leftmost graph in figure 32. Here D stands for the diagnosis (classification of the subject as affected or control), and A and B are the two variables compared. In the middle graph, the relationship can be described as affecting the two variables independently, whereas in the left graph the relationship can be described as the disease affecting one of them but with a relationship to the other that is the same for affected and for controls. Most of the pairs selecting the first graph involved some parts of the vermis (a part of cerebellum). Particularly important was the pair subject age and posterior superior vermis volume. As shown in figure 33, this part of the brain decreases in size with age for normal subjects. But for the affected persons the size is smaller and does not change with age. Neither does the size depend significantly on the duration of the disease or the medication received. Although these findings could be explained by confounding effects, the more likely explanation presently is that the reduction occurred before outbreak of the disease and that processes leading to the disease involve disturbing the development of this part of the brain.

Several other variables were linked to the vermis in the same way: there was an association for control subjects but not for the affected ones, indicating that the normal co-variation is broken by the mechanisms of the disease. For variables that were similarly co-varying with the vermis for controls and affected, there is a possibility that these regions are affected by the disease similarly as the vermis.

In order to get an idea of which physiological and anatomical variables that give best predictors of the disease, a decision tree was produced using the Bayesian approach of section 3.4.1, choosing the most significant variable and decision boundary in a top-down fashion. This tree is shown in figure 34. Note that it does not accurately classify all cases, but leaves 6 of the 144 cases misclassified. A bigger decision tree would classify all cases correctly but also over-fit to the training set, giving probably less accuracy on new cases.

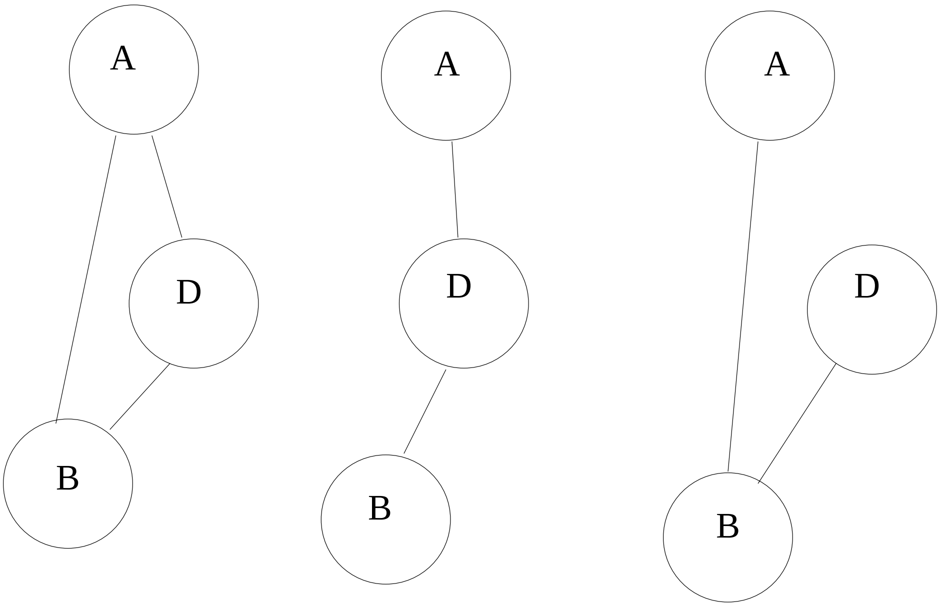

Figure 32: Graphical models detecting co-variation

6 Conclusions

I hope that you now agree with me that applied statistics in computer science is a cool topic, and that you are familiar enough with it to start applying it. Large numbers of computer systems with uncertainty management will be realized the coming decades. These systems differ a lot in their requirements: in many cases standard application of known methods are adequate, but for many, significant development of method and theory will be required.

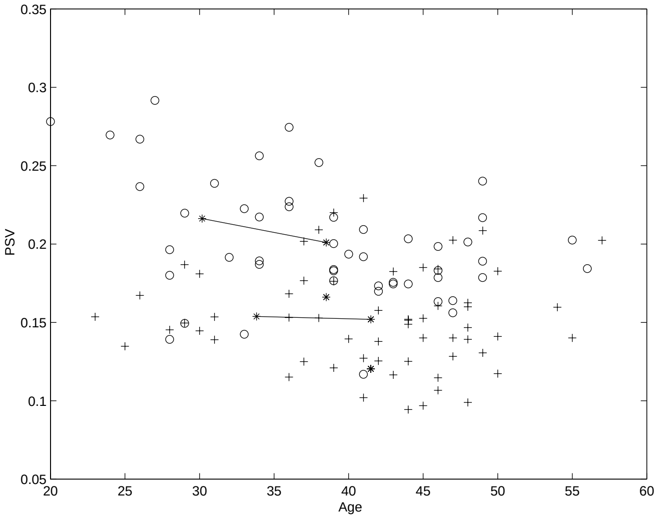

Figure 33: Association between age and posterior inferior vermis volume depends on diagnosis. The principal directions of variation for controls(o) and affected subjects (+) are shown.

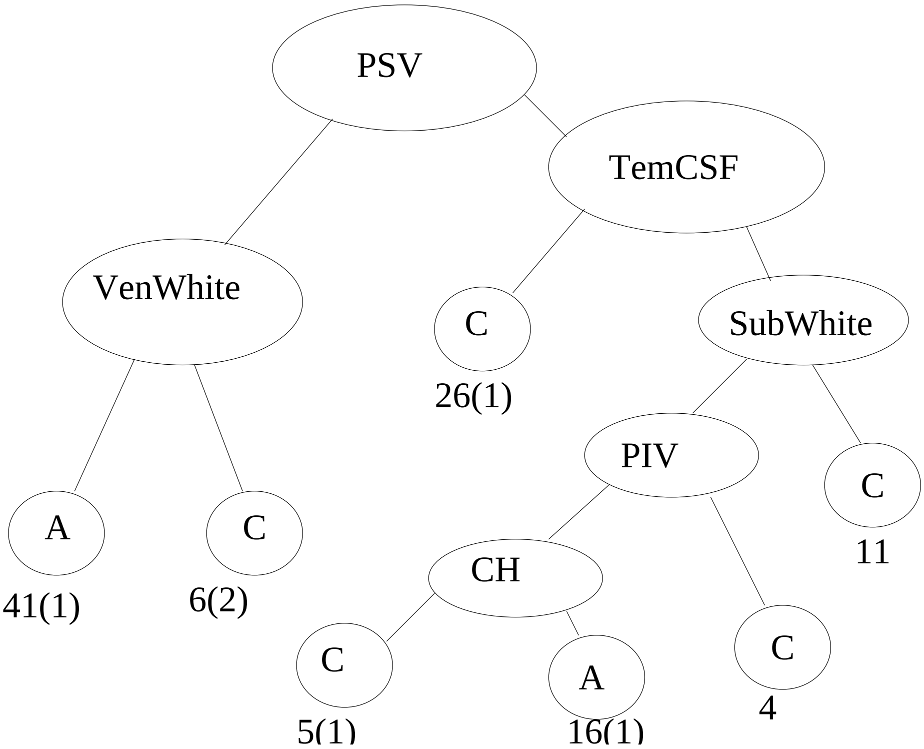

Figure 34: Robust decision tree. Numbers on terminals indicate number of cases in the training set, misclassified subjects in parentheses.

Index

References

[1] Pablo Arambel, Matthew Antone, Constantino Rago, Herbert Landau, and Thomas Strat. A multiple-hypothesis tracking of multiple ground targets from aerial video with dynamic sensor control. In Per Svensson and Johan Schubert, editors, *Proceedings of the Seventh International Conference on Information Fusion*, volume II, pages 1080–1087, Mountain View, CA, Jun 2004. International Society of Information Fusion.

[2] S. Arnborg. In J. Wang, editor, *Data Mining: Opportunities and Challenges*, chapter 1: A Survey of Bayesian Data Mining. Idea Group Publishing, 2003.

[3] S. Arnborg. Robust Bayesian fusion: Imprecise and paradoxical reasoning. In *FUSION 2004*, pages 407–414, 2004.

[4] S. Arnborg and G. Sjödin. Bayes rules in finite models. In *Proc. European Conference on Artificial Intelligence*, pages 571–575, Berlin, 2000.

[5] S. Arnborg and G. Sjödin. On the foundations of Bayesianism. In Ali Mohammad-Djarafi, editor, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 20th International Workshop, Gif-sur-Yvette, 2000*, pages 61–71. American Institute of Physics, 2001.

[6] Stefan Arnborg. Robust Bayesianism: Relation to evidence theory. *ISIF Journal of Advances in Information Fusion*, 1(1):75–90, 2006.

[7] M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* Phd thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[8] S. Benferhat, D. Dubois, and H. Prade. Nonmonotonic reasoning, conditional objects and possibility theory. *Artificial Intelligence*, 92:259–276, 1997.

[9] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. of the Royal statistical Society B*, 57:289–300, 1995.

[10] Y. Benjamini and D. Yeuketeli. The control of the FDR multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

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

[12] J. O. Berger. An overview of robust Bayesian analysis (with discussion). *Test*, 3:5–124, 1994.

[13] N. Bergman. *Recursive Bayesian Estimation.* PhD thesis, Linköping University, 1999.

[14] G. Berkeley. The Analyst. In James Newman, editor, *The World of Mathematics*, New York, 1956. Simon and Schuster.

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

[16] J.M. Bernardo and A.F. Smith. *Bayesian Theory.* Wiley, 1994.

[17] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. Svm and kernel methods matlab toolbox. Perception Systemes et Information, INSA de Rouen, Rouen, France, 2005.

[18] Barbara Caputo. *A new kernel method for object recognition:spin glass-Markov random fields.* PhD thesis, KTH, Sweden, October 11 2004.

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

[20] G. Casella and C. P. Robert. Rao-blackwellization of sampling schemes. Technical report, Cornell University, 1995.

[21] 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.

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

[23] V. Chvátal. *Linear Programming.* Freeman and Company, 1983.

[24] J. Corrander and M. Sillanpää. A unified approach to joint modeling of multiple quantitative and qualitative traits in gene mapping. *J.theor. Biol.*, 218:435–446, 2002.

[25] R. Courant and D. Hilbert. *Methods of Mathematical Physics V.I.* Interscience Publishers, New York, 1953.

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

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

[28] N. Cristianini and J. Shawe-Taylor, editors. *Support Vector Machines and other kernel based methods.* Cambridge University Press, 2000.

[29] A. Dale. *A history of Inverse Probability: from Thomas Bayes to Karl Pearson.* Springer, Berlin, 1991.

[30] B. de Finetti. *Theory of Probability.* London:Wiley, 1974.

[31] P. de Laplace. On probability. In James Newman, editor, *The World of Mathematics*, New York, 1956. Simon and Schuster.

[32] A.P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[33] David G. T. Denison, Christopher C. Holmes, Bani K. Mallick, and Adrian F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression.* Wiley, 2002.

[34] 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.

[35] 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.

[36] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice.* Springer, 2001.

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

[38] W. Feller. *An introduction to probability theory and its applications*, volume 1. Wiley, New York, 1956.

[39] S. Fienberg. When did Bayesian inference become "Bayesian"? *Bayesian analysis*, 1(1):1–40, 2005.

[40] R.A. Fisher. Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535, 1930.

[41] R.A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398, 1935.

[42] D. Fixsen and R.P.S. Mahler. The modified Dempster-Shafer approach to classification. *IEEE Trans. SMC-A*, 27(1):96–104, January 1997.

[43] A. Gammerman and V. Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):164–172, 2007.

[44] B. Gärtner. Fast and robust smallest enclosing balls. In *ESA*, volume 1643 of *LNCS*, pages 325–338, Berlin, 1999. Springer.

[45] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (Second Edition).* Chapman & Hall, New York, 2003.

[46] N. Gershenfeld and A. Weigend. The future of time series: Learning and understanding. In A. Weigend and N. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 1–66. Addison-Wesley, 1993.

[47] S.G. Giakoumatos, I.D. Vrontos, P. Dellaportas, and D.N. Politis. A Markov Chain Monte Carlo convergence diagnostic using subsampling. *Journal of Computational and Graphical Statistics*, 8:431–451, 1999.

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

[49] C. Glymour and G. Cooper, editors. *Computation, Causation and Discovery.* MIT Press, 1999.

[50] I.R. Goodman, R. Mahler, and H.T. Nguyen. *The Mathematics of Data Fusion.* Kluwer Academic Publishers, 1997.

[51] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[52] P. J. Green. Markov Chain Monte Carlo in image analysis. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[53] P.D. Grünwald and A.P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4), 2004.

[54] H. Haario, M. Laine, M. Lehtinen, E. Saksman, and J. Tamminen. Markov Chain Monte Carlo methods for high dimensional inversion in remote sensing. *J. of the Royal statistical Society B*, 66:591–607, 2004.

[55] David J. Hand. Rejoinder: Classifier technology and the illusion of progress. *Statistical Science*, 21:30, 2006.

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

[57] Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–803, 1988.

[58] A. Holst. *The Use of a Bayesian Neural Network Model for Classification Tasks.* Norstedts tryckeri AB, 1997. *Ph D Thesis*, Stockholm University.

[59] E. Hüllermeier. Toward a probabilistic formalization of case-based inference. In D. Thomas, editor, *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99-Vol1)*, pages 248–253, S.F., July 31–August 6 1999. Morgan Kaufmann Publishers.

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

[61] E.T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[62] F. B. Jensen. *Bayesian Networks and Decision Graphs.* Springer, 2001.

[63] H. W. Kuhn and A. Tucker. Non-linear programming. In H. W. Kuhn and A. W. Tucker, editors, *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilities*, pages 481–492. University of California Press, 1951.

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

[65] A. Ledberg. *Measuring Brain Functions: Statistical Tests for Neuroimaging Data.* Phd thesis, Karolinska Institute, Stockholm, Sweden, September 2001.

[66] 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.

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

[68] Heikki Mannila, Hannu Toivonen, Atte Korhola, and Heikki Olander. Learning, mining, or modeling? — a case study from paleoecology.

[69] V. Megalooikonomou, J. Ford, Li Shen, F. Makedon, and A. Saykin. Data mining in brain imaging. *Statistical Methods in Medical Research*, 9:359–394, 2000.

[70] G. Mendel. Experiments in plant-hybridization. *Verh. naturf. Ver. in Brunn*, 4, 1865.

[71] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, Mass, 1969.

[72] T.M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[73] 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.

[74] M.-S. Oh and A. Raftery. Model-based clustering with dissimilarities: A Bayesian approach. Technical Report 441, Dept of Statistics, University of Washington, 2003.

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

[76] Z. Pawlak. *Rough Sets.* Kluwer, Dordrecht, 1992.

[77] Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, England, 2000.

[78] J.G. Propp and D.B. Wilson. How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *Journal of Algorithms*, 27(2):170–217, May 1998.

[79] Lennart Råde and Bertil Westergren. *Beta, Mathematical Handbook.* Chartwell-Bratt, Old Orchard, Bickley Road, Bromley, Kent BR1 2NE, England, second edition, 1990.

[80] A. E. Raftery and V.E. Akman. Bayesian analysis of a Poisson process with a change-point. *Biometrika*, 73:85–89, 1986.

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

[82] 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.

[83] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

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

[85] D. J. Rose, R. E. Tarjan, and G. S. Lueker. Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.*, 5:266–283, 1976.

[86] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[87] R. Royall. On the probability of observing misleading statistical evidence (with discussion). *J. American Statistical Ass.*, 95:760–780, 2000.

[88] C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schoelkopf, and A. Smola. Support vector machine - reference manual. Technical Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1998.

[89] L.J. Savage. *Foundations of Statistics*. John Wiley & Sons, New York, 1954.

[90] Valeriu Savcenko. Bayesian methods for mixture modelling. Master's thesis, Royal Institute of Technology, 2004.

[91] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

[92] G Shafer, A. Gammerman, and V. Vovk. *Algorithmic Learning in a Random World*. Springer, 2005.

[93] G Shafer and V. Vovk. *Probability Theory – it's only a Game*. MIT Press, 2001.

[94] H. Sidenbladh. Probabilistic tracking and reconstruction of 3D human motion in monocular video sequences. Doctoral Dissertation, ISRN KTH/NA/P–01/14–SE, Numerical Analysis and Computer Science, October 2001.

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

[96] P. Smets and R Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[97] F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L. S. Young, editors, *Dynamical systems and turbulence*, pages 366–381. Spinger-Verlag, New York, 1980.

[98] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

[99] L. Tierney. Theoretical perspectives. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[100] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[101] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

[102] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[103] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[104] F. Voorbraak. Partial probability: Theory and applications. In Gert de Cooman, Fabio G. Cozman, Serafin Moral, and Peter Walley, editors, *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*. Gent University, 1999.

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

[106] Helga Westerlind. Diagnosing mcmc burn-in and mixing. Master's thesis, Royal Institute of Technology, 2008.

[107] B. Widrow and M. E. Hoff. Adaptive switching circuits. In *Proceedings WESCON*, pages 96–104, 1960.

[108] N. Wilson. Extended probability. In *Proceedings of the 12th European Conference on Artificial Intelligence*, pages 667–671. John Wiley and Sons, 1996.

[109] M. W. Woolrich, C. F. Beckmann, M. Jenkinson, and S. M. Smith. Multilevel linear modelling for fMRI group analysis using Bayesian inference. *Neuroimage*, 21(4):1732–1747, 2004.

[110] L. A. Zadeh. The roles of fuzzy logic and soft computing in the conception, design and deployment of intelligent systems. *Lecture Notes in Computer Science*, 1198:183–210, 1997.