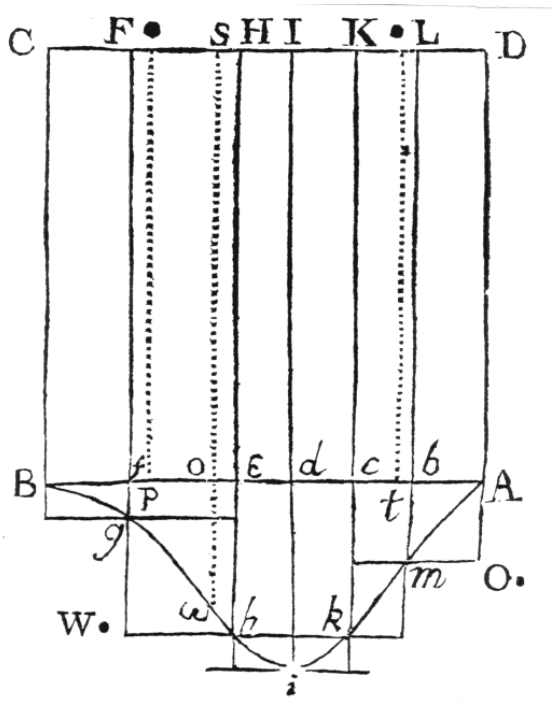
Bayesian Statistics



B.J.K. Kleijn

University of Amsterdam
Korteweg-de Vries institute for Mathematics

Spring 2009

Contents

Preface

These lecture notes were written for the course 'Bayesian Statistics', taught at University of Amsterdam in the spring of 2007. The course was aimed at first-year MSc.-students in statistics, mathematics and related fields. The aim was for students to understand the basic properties of Bayesian statistical methods; to be able to apply this knowledge to statistical questions and to know the extent (and limitations) of conclusions based thereon. Considered were the basic properties of the procedure, choice of the prior by objective and subjective criteria, Bayesian inference, model selection and applications. In addition, non-parametric Bayesian modelling and posterior asymptotic behaviour have received due attention and computational methods were presented.

An attempt has been made to make these lecture notes as self-contained as possible. Nevertheless the reader is expected to have been exposed to some statistics, preferably from a mathematical perspective. It is not assumed that the reader is familiar with asymptotic statistics; these lecture notes provide a general introduction to this topic. Where possible, definitions, lemmas and theorems have been formulated such that they cover parametric and nonparametric models alike. An index, references and an extensive bibliography are included.

Since Bayesian statistics is formulated in terms of probability theory, some background in measure theory is prerequisite to understanding these notes in detail. However the reader is not supposed to have all measure-theorical knowledge handy: appendix A provides an overview of relevant measure-theoretic material. In the description and handling of nonparametric statistical models, functional analysis and topology play a role. Of the latter two, however, only the most basic notions are used and all necessary detail in this respect will be provided during the course.

The author wishes to thank Aad van der Vaart for his contributions to this course and these lecture notes, concerning primarily (but not exclusively) the chapter entitled 'Numerical methods in Bayesian statistics'. For corrections to the notes, the author thanks C. Muris, ...

Bas Kleijn, Amsterdam, January 2007

Chapter 1

Introduction

The goal of inferential statistics is to understand, describe and estimate (aspects of) the randomness of measured data. Quite naturally, this invites the assumption that the data represents a sample from an unknown but fixed probability distribution. Based on that assumption, one may proceed to estimate this distribution directly, or to give estimates of certain characteristic properties (like its mean, variance, *etcetera*). It is this straightforward assumption that underlies frequentist statistics and markedly distinguishes it from the Bayesian approach.

1.1 Frequentist statistics

Any frequentist inferential procedure relies on three basic ingredients: the data, a model and an estimation procedure. The *data* is a measurement or observation which we denote by Y, taking values in a corresponding samplespace.

Definition 1.1.1. *The samplespace for an observation Y is a measurable space $(\mathscr{Y}, \mathscr{B})$ (see definition A.1.1) containing all values that Y can take upon measurement.*

Measurements and data can take any form, ranging from categorical data (sometimes referred to as nominal data where the samplespace is simply a (usually finite) set of points or labels with no further mathematical structure), ordinal data (sometimes called ranked data, where the samplespace is endowed with an total ordering), to interval data (where in addition to having an ordering, the samplespace allows one to compare differences or distances between points), to ratio data (where we have all the structure of the real line). Moreover Y can collect the results of a number of measurements, so that it takes its values in the form of a vector (think of an experiment involving repeated, stochastically independent measurements of the same quantity, leading to a so-called independent and identically distributed (or *i.i.d.*) sample). The data Y may even take its values in a space of functions or in other infinite-dimensional spaces.

The samplespace $\mathscr{Y}$ is assumed to be a measurable space to enable the consideration of probability measures on $\mathscr{Y}$, formalizing the uncertainty in measurement of Y. As was said in the opening words of this chapter, frequentist statistics hinges on the assumption that there exists a probability measure $P_0 : \mathscr{B} \to [0,1]$ on the samplespace $\mathscr{Y}$ representing the "true distribution of the data":

$$Y \sim P_0 \tag{1.1}$$

Hence from the frequentist perspective, inferential statistics revolves around the central question: "What is P_0?", which may be considered in parts by questions like, "What is the mean of P_0?", "What are the higher moments of P_0?", *etcetera*.

The second ingredient of a statistical procedure is a model, which contains all explanations under consideration of the randomness in Y.

Definition 1.1.2. *A statistical model $\mathscr{P}$ is a collection of probability measures $P : \mathscr{B} \to [0,1]$ on the samplespace $(\mathscr{Y}, \mathscr{B})$.*

The model $\mathscr{P}$ contains the candidate distributions for Y that the statistician finds "reasonable" explanations of the uncertainty he observes (or expects to observe) in Y. As such, it constitutes a choice of the statistician analyzing the data rather than a given. Often, we describe the model in terms of probability densities rather than distributions.

Definition 1.1.3. *If there exists a σ-finite measure $\mu : \mathscr{B} \to [0,\infty]$ such that for all $P \in \mathscr{P}$, $P \ll \mu$, we say that the model is dominated.*

The Radon-Nikodym theorem (see theorem A.4.2) guarantees that we may represent a dominated model $\mathscr{P}$ in terms of probability density functions $p = dP/d\mu : \mathscr{Y} \to \mathbb{R}$. Note that the dominating measure may not be unique and hence, that the representation of $\mathscr{P}$ in terms of densities depends on the particular choice of dominating measure μ. A common way of representing a model is a description in terms of a parameterization.

Definition 1.1.4. *A model $\mathscr{P}$ is parameterized with parameter space Θ, if there exists a surjective map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, called the parameterization of $\mathscr{P}$.*

Surjectivity of the parameterization is imposed so that for all $P \in \mathscr{P}$, there exists a $\theta \in \Theta$ such that $P_\theta = P$: unless surjectivity is required the parameterization may describe $\mathscr{P}$ only partially. Also of importance is the following property.

Definition 1.1.5. *A parameterization of a statistical model $\mathscr{P}$ is said to be identifiable, if the map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ is injective.*

Injectivity of the parameterization means that for all $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$ implies that $P_{\theta_1} \neq P_{\theta_2}$. In other words, no two different parameter values θ_1 and θ_2 give rise to the same distribution. Clearly, in order for $\theta \in \Theta$ to serve as a useful representation for the candidate distributions P_θ, identifiability is a first requirement. Other common conditions on the map

$\theta \mapsto P_\theta$ are continuity (with respect to a suitable (often metric) topology on the model), differentiability (which may involve technical subtleties in case Θ is infinite-dimensional) and other smoothness conditions.

Remark 1.1.1. *Although strictly speaking ambivalent, it is commonplace to refer to both $\mathscr{P}$ and the parameterizing space Θ as "the model". This practice is not unreasonable in view of the fact that, in practice, almost all models are parameterized in an identifiable way, so that there exists a bijective correspondence between Θ and $\mathscr{P}$.*

A customary assumption in frequentist statistics is that the model is well-specified.

Definition 1.1.6. *A model $\mathscr{P}$ is said to be well-specified if it contains the true distribution of the data P_0, i.e.*

$$P_0 \in \mathscr{P}. \tag{1.2}$$

If (1.2) does not hold, the model is said to be mis-specified.

Clearly if $\mathscr{P}$ is parameterized by Θ, (1.2) implies the existence of a point $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$; if, in addition, the model is identifiable, the parameter value θ_0 is unique.

Notwithstanding the fact that there may be inherent restrictions on the possible distributions for Y (like guaranteed positivity of the measurement outcome, or symmetries in the problem), the model we use in a statistical procedure constitutes a *choice* rather than a given: presented with a particular statistical problem, different statisticians may choose to use different models. The only condition is that (1.2) is satisfied, which is why we have to choose the model in a "reasonable way" given the nature of Y. However, since P_0 is unknown, (1.2) has the status of an assumption on the unknown quantity of interest P_0 and may, as such, be hard to justify depending on the comprehensiveness of $\mathscr{P}$. When choosing the model, two considerations compete: on the one hand, small models are easy to handle mathematically and parameters are usually clearly interpretable, on the other hand, for large models, assumption (1.2) is more realistic since they have a better chance of containing P_0 (or at least approximate it more closely). In this respect the most important distinction is made in terms of the dimension of the model.

Definition 1.1.7. *A model $\mathscr{P}$ is said to be parametric of dimension d, if there exists an identifiable parameterization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, where $\Theta \subset \mathbb{R}^d$ with non-empty interior $\mathring{\Theta} \neq \emptyset$.*

The requirement regarding the interior of Θ in definition 1.1.7 ensures that the dimension d really concerns Θ and not just the dimension of the space $\mathbb{R}^d$ of which Θ forms a subset.

Example 1.1.1. *The normal model for a single, real measurement Y, is the collection of all normal distributions on $\mathbb{R}$, i.e.*

$$\mathscr{P} = \left\{ N(\mu, \sigma^2) : (\mu, \sigma) \in \Theta \right\}$$

where the parameterizing space Θ equals $\mathbb{R} \times (0, \infty)$. The map $(\mu, \sigma) \mapsto N(\mu, \sigma^2)$ is surjective and injective, i.e. the normal model is a two-dimensional, identifiable parametric model. Moreover, the normal model is dominated by the Lebesgue measure on the samplespace $\mathbb{R}$ and can hence be described in terms of Lebesgue-densities:

$$p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Definition 1.1.8. *If an infinite-dimensional space Θ is needed to parameterize $\mathscr{P}$, then $\mathscr{P}$ is called a non-parametric model.*

For instance, the model consisting of *all* probability measures on $(\mathscr{Y}, \mathscr{B})$ (sometimes referred to as the full non-parametric model) is non-parametric unless the samplespace contains a finite number of points. Note that if the full non-parametric model is used, (1.2) holds trivially.

Example 1.1.2. *Let $\mathscr{Y}$ be a finite set containing $n \geq 1$ points $y_1, y_2, \ldots, y_n$ and let $\mathscr{B}$ be the power-set $2^{\mathscr{Y}}$ of $\mathscr{Y}$. Any probability measure $P : \mathscr{B} \to [0, 1]$ on $(\mathscr{Y}, \mathscr{B})$ is absolutely continuous with respect to the counting measure on $\mathscr{Y}$ (see example A.2.1). The density of P with respect to the counting measure is a map $p : \mathscr{Y} \to \mathbb{R}$ such that $p \geq 0$ and*

$$\sum_{i=1}^{n} p(y_i) = 1.$$

As such, P can be identified with an element of the so-called simplex S_n in $\mathbb{R}^n$, defined as follows

$$S_n = \Big\{ p = (p_1, \ldots, p_n) \in \mathbb{R}^n \,:\, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \Big\}.$$

This leads to an identifiable parameterization $S_n \to \mathscr{P} : p \mapsto P$ of the full non-parametric model on $(\mathscr{Y}, \mathscr{B})$, of dimension $n - 1$. Note that S_n has empty interior in $\mathbb{R}^n$, but can be brought in one-to-one correspondence with a compact set in $\mathbb{R}^{n-1}$ with non-empty interior by the embedding:

$$\Big\{ (p_1, \ldots, p_{n-1}) \in \mathbb{R}^{n-1} \,:\, p_i \geq 0, \sum_{i=1}^{n-1} p_i \leq 1 \Big\} \to S_n \,:$$

$$(p_1, \ldots, p_{n-1}) \mapsto \Big(p_1, \ldots, p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i \Big).$$

The third ingredient of a frequentist inferential procedure is an estimation method. Clearly not all statistical problems involve an explicit estimation step and of those that do, not all estimate the distribution P_0 directly. Nevertheless, one may regard the problem of point-estimation in the model $\mathscr{P}$ as prototypical.

Definition 1.1.9. *A point-estimator (or estimator) is a map $\hat{P} : \mathscr{Y} \to \mathscr{P}$, representing our "best guess" $\hat{P}(Y) \in \mathscr{P}$ for P_0 based on the data Y (and other known quantities).*

Note that a point-estimator is a *statistic*, *i.e.* a quantity that depends only on the data (and possibly on other known information): since a point-estimator must be calculable in practice, it may depend only on information that is *known* to the statistician after he has performed the measurement with outcome $Y = y$. Also note that a point-estimator is a stochastic quantity: $\hat{P}(Y)$ depends on Y and is hence random with its own distribution on $\mathscr{P}$ (as soon as a σ-algebra on $\mathscr{P}$ is established with respect to which $\hat{P}$ is measurable). Upon measurement of Y resulting in a realisation $Y = y$, the estimator $\hat{P}(y)$ is a definite point in $\mathscr{P}$.

Remark 1.1.2. *Obviously, many other quantities may be estimated as well and the definition of a point-estimator given above is too narrow in that sense. Firstly, if the model is parameterized, one may define a point-estimator $\hat{\theta} : \mathscr{Y} \to \Theta$ for θ_0, from which we obtain $\hat{P}(Y) = P_{\hat{\theta}(Y)}$ as an estimator for P_0. If the model is identifiable, estimation of θ_0 in Θ is equivalent to estimation of P_0 in $\mathscr{P}$. But if the dimension d of the model is greater than one, we may choose to estimate only one component of θ (called the parameter of interest) and disregard other components (called nuisance parameters). More generally, we may choose to estimate certain properties of P_0, for example its expectation, variance or quantiles, rather than P_0 itself. As an example, consider a model $\mathscr{P}$ consisting of distributions on $\mathbb{R}$ with finite expectation and define the linear functional $e : \mathscr{P} \to \mathbb{R}$ by $e(P) = PX$. Suppose that we are interested in the expectation $e_0 = e(P_0)$ of the true distribution. Obviously, based on an estimator $\hat{P}(Y)$ for P_0 we may define an estimator*

$$\widehat{e}(Y) = \int_{\mathscr{Y}} y \, d[\hat{P}(Y)](y) \tag{1.3}$$

to estimate e_0. But in many cases, direct estimation of the property of interest of P_0 can be done more efficiently than through $\hat{P}$.

For instance, assume that X is integrable under P_0 and $Y = (X_1, \ldots, X_n)$ collects the results of an i.i.d. experiment with $X_i \sim P_0$ marginally (for all $1 \leq i \leq n$), then the empirical expectation of X, defined simply as the sample-average of X,

$$\mathbb{P}_n X = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

provides an estimator for e_0. (Note that the sample-average is also of the form (1.3) if we choose as our point-estimator for P_0 the empirical distribution $\hat{P}(Y) = \mathbb{P}_n$ and $\mathbb{P}_n \in \mathscr{P}$.) The law of large numbers guarantees that $\mathbb{P}_n X$ converges to e_0 almost-surely as $n \to \infty$, and the central limit theorem asserts that this convergence proceeds at rate $n^{-1/2}$ (and that the limit distribution is zero-mean normal with $P_0(X - P_0 X)^2$ as its variance) if the variance of X under P_0 is finite. (More on the behaviour of estimators in the limit of large sample-size n can be found in chapter 4.) Many parameterizations $\theta \mapsto P_\theta$ are such that parameters coincide with expectations: for instance in the normal model, the parameter μ coincides with

the expectation, so that we may estimate

$$\hat{\mu}(Y) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

Often, other properties of P_0 can also be related to expectations: for example, if $X \in \mathbb{R}$, the probabilities $F_0(s) = P_0(X \leq s) = P_0 1\{X \leq s\}$ can be estimated by

$$\hat{F}(s) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq s\}$$

i.e. as the empirical expectation of the function $x \mapsto 1\{x \leq s\}$. This leads to a step-function with n jumps of size $1/n$ at samplepoints, which estimates the distribution function F_0. Generalizing, any property of P_0 that can be expressed in terms of an expectation of a P_0-integrable function of X, $P_0(g(X))$, is estimable by the corresponding empirical expectation, $\mathbb{P}_n g(X)$. (With regard to the estimator $\hat{F}$, the convergence $\hat{F}(s) \to F_0(s)$ does not only hold for all $s \in \mathbb{R}$ but even uniform in s, i.e. $\sup_{s\in\mathbb{R}} |\hat{F}(s) - F_0(s)| \to 0$, c.f. the Glivenko-Cantelli theorem.)

To estimate a probability distribution (or any of its properties or parameters), many different estimators may exist. Therefore, the use of any particular estimator constitutes (another) *choice* made by the statistician analyzing the problem. Whether such a choice is a good or a bad one depends on *optimality criteria*, which are either dictated by the particular nature of the problem (see section 2.4 which extends the purely inferential point of view), or based on more generically desirable properties of the estimator (note the use of the rather ambiguous qualification "best guess" in definition 1.1.9).

Example 1.1.3. *To illustrate what we mean by "desirable properties", note the following. When estimating P_0 one may decide to use an estimator $\hat{P}(Y)$ because it has the property that it is close to the true distribution of Y in total variation (see appendix A, definition A.2.1). To make this statement more specific, the property that make such an estimator $\hat{P}$ attractive is that there exists a small constant $\epsilon > 0$ and a (small) significance level $0 < \alpha < 1$, such that for all $P \in \mathscr{P}$,*

$$P\big(\|\hat{P}(Y) - P\| < \epsilon\big) > 1 - \alpha,$$

i.e. if $Y \sim P$, then $\hat{P}(Y)$ lies close to P with high P-probability. Note that we formulate this property "for all P in the model": since $P_0 \in \mathscr{P}$ is unknown, the only way to guarantee that this property holds under P_0, is to prove that it holds for all $P \in \mathscr{P}$, provided that (1.2) holds.

A popular method of estimation that satisfies common optimality criteria in many (but certainly not all!) problems is maximum-likelihood estimation.

Definition 1.1.10. *Suppose that the model $\mathscr{P}$ is dominated by a σ-finite measure μ. The likelihood principle says that one should pick $\hat{P} \in \mathscr{P}$ as an estimator for the distribution P_0 of Y such that*

$$\hat{p}(Y) = \sup_{P\in\mathscr{P}} p(Y).$$

thus defining the maximum-likelihood estimator (or MLE) $\hat{P}(Y)$ for P_0.

Remark 1.1.3. *Note that $\hat{P}$ does not depend on the particular dominating measure μ.*

A word of caution is in order: mathematically, the above "definition" of the MLE begs questions of existence and uniqueness: regarding $P \mapsto p(Y)$ as a (stochastic) map on the model (called the *likelihood*), there may not be any point in $\mathscr{P}$ where the likelihood takes on its supremal value nor is there any guarantee that such a maximal point is unique with P_0-probability equal to one.

Remark 1.1.4. *If $P_\cdot : \Theta \to \mathscr{P}$ parameterizes $\mathscr{P}$, the above is extended to the maximum-likelihood estimator $\hat{\theta}(Y)$ for θ_0, when we note that $\sup_{\theta\in\Theta} p_\theta(Y) = \sup_{P\in\mathscr{P}} p(Y)$.*

The above is only a very brief and rather abstract overview of the basic framework of frequentist statistics, highlighting the central premise that a P_0 for Y exists. It makes clear, however, that frequentist inference concerns itself primarily with the stochastics of the random variable Y and not with the *context* in which Y resides. Other than the fact that the model has to be chosen "reasonably" based on the nature of Y, frequentist inference does not involve any information regarding the background of the statistical problem in its procedures unless one chooses to use such information explicitly (see, for example, remark 2.2.7 on penalized maximum-likelihood estimation). In Bayesian statistics the use of background information is an integral part of the procedure unless one chooses to disregard it: by the definition of a prior measure, the statistician may express that he believes in certain points of the model more strongly than others. This thought is elaborated on further in section 1.2 (*e.g.* example 1.2.1).

Similarly, results of estimation procedures are sensitive to the context in which they are used: two statistical experiments may give rise to the same model formally, but the estimator used in one experiment may be totally unfit for use in the other experiment.

Example 1.1.4. *For example, if we interested in a statistic that predicts the rise or fall of a certain share-price on the stockmarket based on its value over the past week, the estimator we use does not have to be a very conservative one: we are interested primarily in its long-term performance and not in the occasional mistaken prediction. However, if we wish to predict the rise or fall of white-bloodcell counts in an HIV-patient based on last week's counts, overly optimistic predictions can have disastrous consequences.*

Although in the above example, data and model are very similar in these statistical problems, the estimator used in the medical application should be much more conservative than the estimator used in the stock-market problem. The inferential aspects of both questions are the same, but the context in which such inference is made calls for adaptation. Such considerations form the motivation for statistical decision theory, as explained further in section 2.4.

1.2 Bayesian statistics

The subject of these lecture notes is an alternative approach to statistical questions known as Bayesian statistics, after Rev. Thomas Bayes, the author of "*An essay towards solving a problem in the doctrine of chances*", (1763) [4]. Bayes considered a number of probabilistic questions in which data and parameters are treated on equal footing. The Bayesian procedure itself is explained in detail in chapter 2 and further chapters explore its properties. In this section we have the more modest goal of illustrating the conceptual differences with frequentist statistical analysis.

In Bayesian statistics, data and model form two factors of the same space, *i.e.* no formal distinction is made between measured quantities Y and parameters θ. This point of view may seem rather absurd in view of the definitions made in section 1.1, but in [4], Bayes gives examples in which this perspective is perfectly reasonable (see example 2.1.2 in these lecture notes). An element P_θ of the model is interpreted simply as the distribution of Y *given* the parameter value θ, *i.e.* as the conditional distribution of $Y|\theta$. The joint distribution of (Y, θ) then follows upon specification of the marginal distribution of θ on Θ, which is called the *prior*. Based on the joint distribution for the data Y and the parameters θ, straightforward conditioning on Y gives rise to a distribution for the parameters $\theta|Y$ called the *posterior* distribution on the model Θ. Hence, given the model, the data and a prior distribution, the Bayesian procedure leads to a posterior distribution that incorporates the information provided by the data.

Often in applications, the nature of the data and the background of the problem suggest that certain values of θ are more "likely" than others, even before any measurements are done. The model Θ describes possible probabilistic explanations of the data and, in a sense, the statistician believes more strongly in certain explanations than in others. This is illustrated by the following example, which is due to L. Savage [74].

Example 1.2.1. *Consider the following three statistical experiments:*

1. *A lady who drinks milk in her tea claims to be able to tell which was poured first, the tea or the milk. In ten trials, she determines correctly whether it was tea or milk that entered the cups first.*

2. *A music expert claims to be able to tell whether a page of music was written by Haydn or by Mozart. In ten trials conducted, he correctly determines the composer every time.*

3. *A drunken friend says that he can predict the outcome of a fair coin-flip. In ten trials, he is right every time.*

Let us analyse these three experiments in a frequentist fashion, e.g. we assume that the trials are independent and possess a definite Bernoulli distribution, c.f. (1.1). In all three experiments, $\theta_0 \in \Theta = [0, 1]$ is the per-trial probability that the person gives the right answer. We

test their respective claims posing the hypotheses:

$$H_0: \quad \theta_0 = \tfrac{1}{2}, \qquad H_1: \quad \theta_0 > \tfrac{1}{2}.$$

The total number of successes out of ten trials is a sufficient statistic for θ and we use it as our test-statistics, noting that its distribution is binomial with $n = 10$, $\theta = \theta_0$ under H_0. Given the data Y with realization y of ten correct answers, applicable in all three examples, we reject H_0 at p-value $2^{-10} \approx 0.1\%$. So there is strong evidence to support the claims made in all three cases. Note that there is no difference in the frequentist analyses: formally, all three cases are treated exactly the same.

Yet intuitively (and also in every-day practice), one would be inclined to treat the three claims on different footing: in the second experiment, we have no reason to doubt the expert's claim, whereas in the third case, the friend's condition makes his claim less than plausible. In the first experiment, the validity of the lady's claim is hard to guess beforehand. The outcome of the experiments would be as expected in the second case and remarkable in the first. In the third case, one would either consider the friend extremely lucky, or begin to doubt the fairness of the coin being flipped.

The above example convincingly makes the point that in our intuitive approach to statistical issues, we include *all* knowledge we have, even resorting to strongly biased estimators if the model does not permit a non-biased way to incorporate it. The Bayesian approach to statistics allows us to choose the prior such as to reflect this subjectivity: from the outset, we attach more prior mass to parameter-values that we deem more likely, or that we believe in more strongly. In the above example, we would choose a prior that concentrates more mass at high values of θ in the second case and at low values in the third case. In the first case, the absence of prior knowledge would lead us to remain objective, attaching equal prior weights to high and low values of θ. Although the frequentist's testing procedure can be adapted to reflect subjectivity, the Bayesian procedure incorporates it rather more naturally through the choice of a prior.

Subjectivist Bayesians view the above as an advantage; objectivist Bayesians and frequentists view it as a disadvantage. Subjectivist Bayesians argue that personal beliefs are an essential part of statistical reasoning, deserving of a explicit role in the formalism and interpretation of results. Objectivist Bayesians and frequentists reject this thought because scientific reasoning should be devoid of any personal beliefs or interpretation. So the above freedom in the choice of the prior is also the Achilles' heel of Bayesian statistics: fervent frequentists and objectivist Bayesians take the point of view that the choice of prior is an undesirable source of ambiguity, rather than a welcome way to incorporate "expert knowledge" as in example 1.2.1. After all, if the subjectivist Bayesian does not like the outcome of his analysis, he can just go back and change the prior to obtain a different outcome. Similarly, if two subjectivist Bayesians analyze the same data they may reach completely different conclusions, depending on the extent to which their respective priors differ.

To a certain extent, such ambiguity is also present in frequentist statistics, since frequentists make a choice for a certain point-estimator. For example, the use of either a maximum-likelihood or penalized maximum-likelihood estimator leads to differences, the size of which depends on the relative sizes of likelihood and penalty. (Indeed, through the maximum-a-posteriori Bayesian point-estimator (see definition 2.2.5), one can demonstrate that the log-prior-density can be viewed as a penalty term in a penalized maximum-likelihood procedure, *c.f.* remark 2.2.7.) Yet the natural way in which subjectivity is expressed in the Bayesian setting is more explicit. Hence the frequentist or objectivist Bayesian sees in this a clear sign that subjective Bayesian statistics lacks universal value unless one imposes that the prior should not express any bias (see section 3.2).

A second difference in philosophy between frequentist and Bayesian statisticians arises as a result of the fact that the Bayesian procedure does not require that we presume the existence of a "true, underlying distribution" P_0 of Y (compare with (1.1)). The subjectivist Bayesian views the model with (prior or posterior) distribution as his own, subjective explanation of the uncertainty in the data. For that reason, subjectivists prefer to talk about their (prior or posterior) "belief" concerning parameter values rather than implying objective validity of their assertions. On the one hand, such a point of view makes intrinsic ambiguities surrounding statistical procedures explicit; on the other hand, one may wonder about the relevance of strictly personal belief in a scientific tradition that emphasizes universality of reported results.

The philosophical debate between Bayesians and frequentist has raged with varying intensity for decades, but remains undecided to this date. In practice, the choice for a Bayesian or frequentist estimation procedure is usually not motivated by philosophical considerations, but by far more practical issues, such as ease of computation and implementation, common custom in the relevant field of application, specific expertise of the researcher or other forms of simple convenience. Recent developments [3] suggest that the philosophical debate will be put to rest in favour of more practical considerations as well.

1.3 The frequentist analysis of Bayesian methods

Since this point has the potential to cause great confusion, we emphasize the following: this course presents Bayesian statistics from a hybrid perspective, *i.e.* we consider Bayesian techniques but analyze them with frequentist methods.

We take the frequentist point of view with regard to the data, *e.g.* assumption (1.1); we distinguish between samplespace and model and we do not adhere to subjectivist interpretations of results (although their perspective is discussed in the main text). On the other hand, we endow the model with a prior probability measure and calculate the posterior distribution, *i.e.* we use concepts and definitions from Bayesian statistics. This enables us to assess Bayesian methods on equal footing with frequentist statistical methods and extends the range of interesting questions. Moreover, it dissolves the inherent ambiguity haunting the subjectivist interpretation of statistical results.

Note, however, that the derivation of expression (2.7) (for example), is the result of subjectivist Bayesian assumptions on data and model. Since these assumptions are at odds with the frequentist perspective, we shall take (2.7) as a *definition* rather than a derived form. This has the consequence that some basic properties implicit by derivation in the Bayesian framework, have to be imposed as conditions in the hybrid perspective (see remark 2.1.4).

Much of the material covered in these lecture notes does not depend on any particular philosophical point of view, especially when the subject matter is purely mathematical. Nevertheless, it is important to realize when philosophical issues may come into play and there will be points where this is the case. In particular when discussing asymptotic properties of Bayesian procedures (see chapter 4), adoption of assumption (1.1) is instrumental, basically because discussing convergence requires a limit-point.

Notation and conventions

Throughout these notes, we make use of notation that is common in the mathematical-statistics literature. In addition, the following notational conventions are used. The expectation of a random variable Z distributed according to a probability distribution P is denoted PZ. Samples are denoted Y with realization y, or in the case of n *i.i.d.*-P_0 observations, $X_1, \ldots, X_n$. The *sample-average* (or *empirical expectation*) for a sample $X_1, \ldots, X_n$, denoted $\mathbb{P}_n X$, is defined $\mathbb{P}_n X = n^{-1} \sum_{i=1}^n X_i$ (where it is assumed that X is P_0-integrable); the *empirical process* $\mathbb{G}_n$ is defined as $\mathbb{G}_n X = n^{1/2}(\mathbb{P}_n - P_0)X$ (where it is assumed that $P_0(X - P_0 X)^2 < \infty$). The distribution function of the standard normal distribution is denoted $\Phi : \mathbb{R} \to [0,1]$. The transpose of a vector $\ell \in \mathbb{R}^d$ is denoted ℓ^T; the transpose of a matrix I is denoted I^T. The formulation "$A(n)$ holds for large enough n" should be read as "there exists an $N \geq 1$ such that for all $n \geq N$, $A(n)$ holds".

1.4 Exercises

Exercise 1.1. *Let $Y \in \mathscr{Y}$ be a random variable with unknown distribution P_0. Let $\mathscr{P}$ be a model for Y, dominated by a σ-finite measure μ. Assume that the maximum-likelihood estimator $\hat{P}(Y)$ (see definition 1.1.10) is well-defined, P_0-almost-surely.*
Show that if ν is a σ-finite measure dominating μ and we calculate the likelihood using ν-densities, then the associated MLE is equal to $\hat{P}(Y)$. Conclude that the MLE does not depend on the dominating measure used, c.f. remark 1.1.3.

Exercise 1.2. *In the three experiments of example 1.2.1, give the Neyman-Person test for hypotheses H_0 and H_1 at level $\alpha \in (0,1)$. Calculate the p-value of the realization of 10 successes and 0 failures (in 10 Bernoulli trials according to H_0).*

Chapter 2

Bayesian basics

In this chapter, we consider the basic definitions and properties of Bayesian inferential and decision-theoretic methods. Naturally the emphasis lies on the posterior distribution, which we derive from the prior based on the subjectivist perspective. However, we also discuss the way prior and posterior should be viewed if one assumes the frequentist point of view. Furthermore, we consider point estimators derived from the posterior, credible sets, testing of hypotheses and Bayesian decision theory. Throughout the chapter, we consider frequentist methods side-by-side with the Bayesian procedures, for comparison and reference.

It should be stressed that the material presented here covers only the most basic Bayesian concepts; further reading is recommended. Various books providing overviews of Bayesian statistics are recommended, depending on the background and interest of the reader: a highly theoretical treatment can be found in Le Cam (1986) [63], which develops a general, mathematical framework for statistics and decision theory, dealing with Bayesian methods as an important area of its application. For a more down-to-earth version of this work, applied only to smooth parametric models, the interested reader is referred to Le Cam and Yang (1990) [64]. The book by Van der Vaart (1998) [83] contains a chapter on Bayesian statistics focusing on the Bernstein-Von Mises theorem (see also section 4.4 in these notes). A general reference of a more decision-theoretic inclination, focusing on Bayesian statistics, is the book by Berger (1985) [8]; a more recent reference of a similar nature is Bernardo and Smith (1993) [13]. Both Berger and Bernardo and Smith devote a great deal of attention to philosophical arguments in favour of the Bayesian approach to statistics, staying rather terse with regard to mathematical detail and focusing almost exclusively on parametric models. Recommendable is also Robert's "The Bayesian choice" (2001) [72], which offers a very useful explanation on computational aspects of Bayesian statistics. Finally, Ripley (1996) [73] discusses Bayesian methods with a very pragmatic focus on pattern classification. The latter reference relates all material with applications in mind but does so based on a firm statistical and decision-theoretic background.

2.1 Bayes' rule, prior and posterior distributions

Formalizing the Bayesian procedure can be done in several ways. We start this section with considerations that are traditionally qualified as being of a "subjectivist" nature, but eventually we revert to the "frequentist" point of view. Concretely this means that we derive an expression for the posterior and prove regularity in the subjectivist framework. In a frequentist setting, this expression is simply used as a definition and properties like regularity and measurability are imposed. Ultimately the philosophical motivation becomes irrelevant from the mathematical point of view, once the posterior and its properties are established.

Perhaps the most elegant (and decidedly subjectivist) Bayesian framework unifies samplespace and model in a product space. Again, the measurement Y is a random variable taking values in a samplespace $\mathscr{Y}$ with σ-algebra $\mathscr{B}$. Contrary to the frequentist case, in the Bayesian approach the model Θ is assumed to be a measurable space as well, with σ-algebra $\mathscr{G}$. The model parameter takes values $\theta \in \Theta$ but is a random variable (denoted ϑ) in this context! We assume that on the product-space $\mathscr{Y} \times \Theta$ (with product σ-algebra $\mathscr{F} = \sigma(\mathscr{B} \times \mathscr{G})$) we have a probability measure

$$\Pi : \sigma(\mathscr{B} \times \mathscr{G}) \to [0,1], \tag{2.1}$$

which is *not* a product measure. The probability measure Π provides a joint probability distribution for (Y, ϑ), where Y is the observation and ϑ (the random variable associated with) the parameter of the model.

Implicitly the choice for the measure Π defines the model in Bayesian context, by the possibility to condition on $\vartheta = \theta$ for some $\theta \in \Theta$. The *conditional distribution* $\Pi_{Y|\vartheta} : \mathscr{B} \times \Theta \to [0,1]$ describes the distribution of the observation Y *given* the parameter ϑ. (For a discussion of conditional probabilities, see appendix A, *e.g.* definition A.6.3 and theorem A.6.1). As such, it defines the elements P_θ of the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, although the role they play in Bayesian context is slightly different from that in a frequentist setting. The question then arises under which requirements the conditional probability $\Pi_{Y|\vartheta}$ is a so-called regular conditional distribution.

Lemma 2.1.1. *Assume that Θ is a Polish space and that the σ-algebra $\mathscr{G}$ contains the Borel σ-algebra. Let Π be a probability measure, c.f. (2.1). Then the conditional probability $\Pi_{Y|\vartheta}$ is regular.*

Proof The proof is a direct consequence of theorem A.6.1. □

The measures $\Pi_{Y|\vartheta}(\,\cdot\,|\vartheta = \theta)$ form a (Π-almost-sure) version of the elements P_θ of the model $\mathscr{P}$:

$$P_\theta = \Pi_{Y|\vartheta}(\,\cdot\,|\,\vartheta = \theta\,) : \mathscr{B} \to [0,1] \tag{2.2}$$

Consequently, frequentist's notion of a model is only represented up to null-sets of the marginal distribution of ϑ, referred to in Bayesian context as the *prior* for the parameter ϑ.

Definition 2.1.1. *The marginal probability* Π *on* $\mathscr{G}$ *is the prior.*

The prior is interpreted in the subjectivist's philosophy as the "degree of belief" attached to subsets of the model *a priori*, that is, before any observation has been made or incorporated in the calculation. Central in the Bayesian framework is the conditional distribution for ϑ given Y.

Definition 2.1.2. *The conditional distribution*

$$\Pi_{\vartheta|Y} : \mathscr{G} \times \mathscr{Y} \to [0,1], \tag{2.3}$$

is called the posterior distribution.

The posterior is interpreted as a data-amended version of the prior, that is to say, the subjectivist's original "degree of belief", corrected by observation of Y through conditioning, *i.e.* the distribution for ϑ *a posteriori* (that is, after observations have been incorporated).

Assuming that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure on $\mathscr{Y}$, the above can also be expressed in terms of μ-densities $p_\theta = dP_\theta/d\mu : \mathscr{Y} \to \mathbb{R}$. Using Bayes' rule (*c.f.* lemma A.6.2), we obtain the following expression for the posterior distribution:

$$\Pi(\,\vartheta \in G \,|\, Y\,) = \frac{\displaystyle\int_G p_\theta(Y)\, d\Pi(\theta)}{\displaystyle\int_\Theta p_\theta(Y)\, d\Pi(\theta)}, \tag{2.4}$$

where $G \in \mathscr{G}$ is a measurable subset of the model $\mathscr{P}$. Note that when expressed through (2.4), the posterior distribution can be calculated based on a choice for the model (which specifies p_θ) with a prior Π and the data Y (or a realisation $Y = y$ thereof).

Based on the above definitions, two remarks are in order with regard to the notion of a *model* in Bayesian statistics. First of all, one may choose a large model $\mathscr{P}$, but if for a subset $\mathscr{P}_1 \subset \mathscr{P}$ the prior assigns mass zero, then for all practical purposes $\mathscr{P}_1$ does not play a role, since omission of $\mathscr{P}_1$ from $\mathscr{P}$ does not influence the posterior. As long as the model is parametric, *i.e.* $\Theta \subset \mathbb{R}^d$, we can always use priors that dominate the Lebesgue measure, ensuring that $\mathscr{P}_1$ is a "small" subset of $\mathbb{R}^d$. However, in non-parametric models null-sets of the prior and posterior may be much larger than expected intuitively (for a striking example, see section 4.2, specifically the discussion of Freedman's work).

Example 2.1.1. *Taking the above argument to the extreme, consider a normal location model* $\mathscr{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$ *with a prior* $\Pi = \delta_{\theta_1}$ *(see example A.2.2), for some* $\theta_1 \in \Theta$*, defined on the Borel* σ*-algebra* $\mathscr{B}$*. Then the posterior takes the form:*

$$\Pi(\vartheta \in A|Y) = \int_A p_\theta(Y)\, d\Pi(\theta) \,\Big/\, \int_\Theta p_\theta(Y)\, d\Pi(\theta) = \frac{p_{\theta_1}(Y)}{p_{\theta_1}(Y)}\Pi(A) = \Pi(A).$$

for any $A \in \mathscr{B}$*. In other words, the posterior* equals *the prior, concentrating all its mass in the point* θ_1*. Even though we started out with a model that suggests estimation of location,*

effectively the model consists of only one point, $\theta_1 \in \Theta$ due to the choice of the prior. In subjectivist terms, the prior belief is fully biased towards θ_1, leaving no room for amendment by the data when we condition to obtain the posterior.

This example raises the question which part of the model proper $\mathscr{P}$ plays a role. In that respect, it is helpful to make the following definition.

Definition 2.1.3. *In addition to $(\Theta, \mathscr{G}, \Pi)$ being a probability space, let $(\Theta, \mathscr{T})$ be a topological space. Assume that $\mathscr{G}$ contains the Borel σ-algebra $\mathscr{B}$ corresponding to the topology $\mathscr{T}$. The support* $\mathrm{supp}(\Pi)$ *of the prior Π is defined as:*

$$\mathrm{supp}(\Pi) = \bigcap \{G \in \mathscr{G} : G \text{ closed}, \ \Pi(G) = 1\}.$$

The viability of the above definition is established in the following lemma.

Lemma 2.1.2. *For any topological space Θ with σ-algebra $\mathscr{G}$ that contains the Borel σ-algebra $\mathscr{B}$ and any (prior) probability measure $\Pi : \mathscr{G} \to [0, 1]$, $\mathrm{supp}(\Pi) \in \mathscr{G}$ and $\Pi(\mathrm{supp}(\Pi)) = 1$.*

N ote that $\mathrm{supp}(\Pi)$ is closed, as it is an intersection of closed sets, $\mathrm{supp}(\Pi) \in \mathscr{B} \subset \mathscr{G}$. The proof that the support has measure 1 is left as exercise 2.7. □

In example 2.1.1, the model $\mathscr{P}$ consists of all normal distributions of the form $N(\theta, 1)$, $\theta \in \mathbb{R}$, but the support of the prior $\mathrm{supp}(\Pi)$ equals the singleton $\{N(\theta_1, 1)\} \subset \mathscr{P}$.

Note that the support of the prior is defined based on a topology, the Borel σ-algebra of which must belong to the domain of the prior measure. In parametric models this assumption is rarely problematic, but in non-parametric models finding such a prior may be difficult and the support may be an ill-defined concept. Therefore we may choose to take a less precise but more generally applicable perspective: the model is viewed as the support of the prior Π, but only *up to* Π*-null-sets* (*c.f.* the Π-almost-sure nature of the identification (2.2)). That means that we may add to or remove from the model at will, as long as we make sure that the changes have prior measure equal to zero: the model itself is a Π-almost-sure concept. (Since the Bayesian procedure involves only integration of integrable functions with respect to the prior, adding or removing Π-null-sets to/from the domain of integration will not have unforeseen consequences.)

To many who have been introduced to statistics from the frequentist point of view, including the parameter θ for the model as a random variable ϑ seems somewhat unnatural because the frequentist role of the parameter is entirely different from that of the data. The following example demonstrates that in certain situations the Bayesian point of view is not unnatural at all.

Example 2.1.2. *In the posthumous publication of "An essay towards solving a problem in the doctrine of chances" in 1763 [4], Thomas Bayes included an example of a situation in which the above, subjectivist perspective arises quite naturally. It involves a number of red balls and one white ball placed on a table and has become known in the literature as Bayes' billiard.*

We consider the following experiment: unseen by the statistician, someone places n red balls and one white ball on a billiard table of length 1. Calling the distance between the white ball and the bottom cushion of the table X and the distances between the red balls and the bottom cushion Y_i, $(i = 1, \ldots, n)$, it is known to the statistician that their joint distribution is:

$$(X; Y_1, \ldots, Y_n) \sim U[0,1]^{n+1}, \tag{2.5}$$

i.e. all balls are placed independently with uniform distribution. The statistician will be reported the number K of red balls that is closer to the cushion than the white ball (the data*, denoted Y in the rest of this section) and is asked to give a distribution reflecting his beliefs concerning the position of the white ball X (the* parameter*, denoted ϑ in the rest of this section) based on K. His prior knowledge concerning X (i.e. without knowing the observed value $K = k$) offers little information: the best that can be said is that $X \sim U[0,1]$, the marginal distribution of X, i.e. the prior. The question is how this distribution for X changes when we incorporate the observation $K = k$, that is, when we use the observation to arrive at our* posterior *beliefs based on our* prior *beliefs.*

Since for every i, Y_i and X are independent c.f. (2.5), we have,

$$P(Y_i \leq X | X = x) = P(Y_i \leq x) = x,$$

So for each of the red balls, determining whether it lies closer to the cushion than the white ball amounts to a Bernoulli experiment with parameter x. Since in addition the positions $Y_1, \ldots, Y_n$ are independent, counting the number K of red balls closer to the cushion than the white ball amounts to counting "successes" in a sequence of independent Bernoulli experiments. We conclude that K has a binomial distribution $\mathrm{Bin}(n; x)$*, i.e.*

$$P(K = k | X = x) = \frac{n!}{k!(n-k)!} x^k (1-x)^{n-k}.$$

It is possible to obtain the density for the distribution of X conditional on $K = k$ from the above display using Bayes' rule (c.f. lemma A.6.2):

$$p(x | K = k) = P(K = k | X = x) \frac{p(x)}{P(K = k)}, \tag{2.6}$$

but in order to use it, we need the two marginal densities $p(x)$ and $P(K = k)$ in the fraction. From (2.5) it is known that $p(x) = 1$ and $P(K = k)$ can be obtained by integrating

$$P(K = k) = \int_0^1 P(K = k | X = x)\, p(x)\, dx$$

Substituting in (2.6), we find:

$$p(x | K = k) = \frac{P(K = k | X = x)\, p(x)}{\int_0^1 P(K = k | X = x)\, p(x)\, dx} = B(n,k)\, x^k (1-x)^{n-k}.$$

where $B(n,k)$ is a normalization factor. The x-dependence of the density in the above display reveals that $X | K = k$ is distributed according to a Beta-distribution, $B(k+1, n-k+1)$, so that the normalization factor $B(n,k)$ must equal $B(n,k) = \Gamma(n+2)/\Gamma(k+1)\Gamma(n-k+1)$.

This provides the statistician with distributions reflecting his beliefs concerning the position of the white ball for all possible values k for the observation K. Through conditioning on $K = k$, the prior distribution of X is changed: if a relatively small number of red balls is closer to the cushion than the white ball (i.e. in case k is small compared to n), then the white ball is probably close to the cushion; if k is relatively large, the white ball is probably far from the cushion (see figure 2.1).

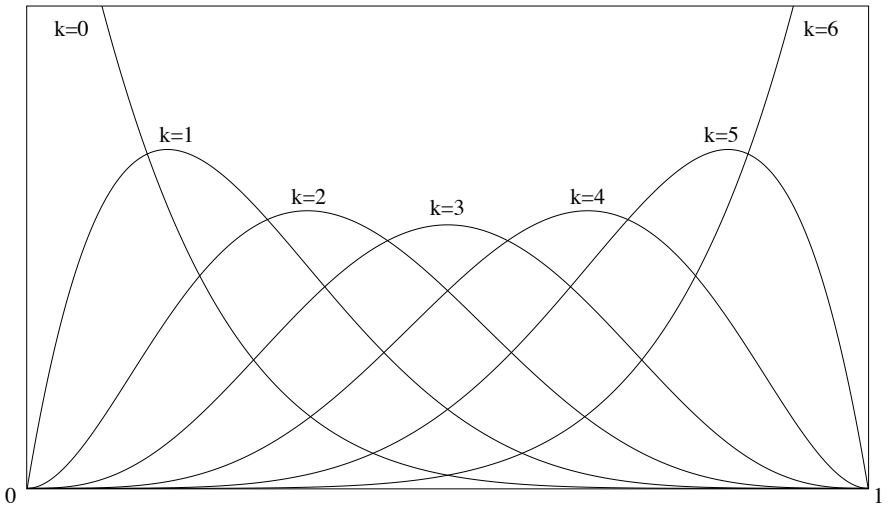


FIGURE 2.1 Posterior densities for the position X of the white ball, given the number k of red balls closer to the cushion of the billiard (out of a total of $n = 6$ red balls). For the lower values of k, the white ball is close to the cushion with high probability, since otherwise more red balls would probably lie closer to the cushion. This is reflected by the posterior density for $X|K = 1$, for example, by the fact that it concentrates much of its mass close to $x = 0$.

In many experiments or observations, the data consists of a sample of n repeated, stochastically independent measurements of the same quantity. To accommodate this situation formally, we choose $\mathscr{Y}$ equal to the n-fold product of a sample space $\mathscr{X}$ endowed with a σ-algebra $\mathscr{A}$, so that the observation takes the form $Y = (X_1, X_2, \ldots, X_n)$. The additional assumption that the sample is *i.i.d.* (presently a statement concerning the *conditional independence* of the observations given $\vartheta = \theta$) then reads:

$$\Pi_{Y|\vartheta}(\, X_1 \in A_1, \ldots, X_n \in A_n \,|\, \vartheta = \theta \,) = \prod_{i=1}^{n} \Pi_{Y|\vartheta}(\, X_i \in A_i \,|\, \vartheta = \theta \,) = \prod_{i=1}^{n} P_\theta(X_i \in A_i),$$

for all $(A_1, \ldots, A_n) \in \mathscr{A}^n$. Assuming that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ on $\mathscr{X}$, the above can also be expressed in terms of μ-densities $p_\theta = dP_\theta/d\mu : \mathscr{X} \to \mathbb{R}$. Using Bayes' rule, we obtain the following expression for the posterior

distribution:

$$\Pi_n(\,\vartheta \in G \,|\, X_1, X_2, \ldots, X_n\,) = \frac{\displaystyle\int_G \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}, \tag{2.7}$$

where $G \in \mathscr{G}$ is a measurable subset of the model $\mathscr{P}$.

Remark 2.1.1. *In a dominated model, the Radon-Nikodym derivative (see theorem A.4.2) of the posterior with respect to the prior is the likelihood function, normalized to be a probability density function:*

$$\frac{d\Pi(\,\cdot \,|\, X_1, \ldots, X_n)}{d\Pi}(\theta) = \prod_{i=1}^n p_\theta(X_i) \Big/ \int_\Theta \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta), \quad (P_0^n - a.s.). \tag{2.8}$$

The latter fact explains why such strong relations exist (e.g. the Bernstein-Von Mises theorem, theorem 4.4.1) between Bayesian and maximum-likelihood methods. Indeed, the proportionality of the posterior density and the likelihood provides a useful qualitative picture of the posterior as a measure that concentrates on regions in the model where the likelihood is relatively high. This may serve as a direct motivation for the use of Bayesian methods in a frequentist context, c.f. section 1.3. Moreover, this picture gives a qualitative explanation of the asymptotic behaviour of Bayesian methods: under suitable continuity-, differentiability- and tail-conditions, the likelihood remains relatively high in small neighbourhoods of P_0 and drops off steeply outside in the large-sample limit. Hence, if the prior mass in those neighbourhoods is not too small, the posterior concentrates its mass in neighbourhoods of P_0, leading to the asymptotic behaviour described in chapter 4.

Returning to the distributions that play a role in the subjectivist Bayesian formulation, there exists also a marginal for the observation Y.

Definition 2.1.4. *The distribution $P^\Pi : \mathscr{B} \to [0,1]$ defined by*

$$P_n^\Pi(\,X_1 \in A_1, \ldots, X_n \in A_n\,) = \int_\Theta \prod_{i=1}^n P_\theta(A_i)\, d\Pi(\theta) \tag{2.9}$$

is called the prior predictive distribution.

Strictly speaking the prior predictive distribution describes a subjectivist's expectations concerning the observations $X_1, X_2, \ldots, X_n$ based only on the prior Π, *i.e.* without involving the data. More readily interpretable is the following definition.

Definition 2.1.5. *For given $n, m \geq 1$, the distribution $P_{n,m}^\Pi$ defined by*

$$P_{n,m}^\Pi(\,X_{n+1} \in A_{n+1}, \ldots, X_{n+m} \in A_{n+m} \,|\, X_1, \ldots, X_n\,) = \int_\Theta \prod_{i=1}^m P_\theta(A_{n+i})\, d\Pi(\theta \,|\, X_1, \ldots, X_n\,))$$

is called the posterior predictive distribution.

The prior predictive distribution is subject to correction by observation through substitution of the prior by the posterior: the resulting posterior predictive distribution is interpreted as the Bayesian's expectation concerning the distribution of the observations $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$ given the observations $X_1, X_2, \ldots, X_n$ and the prior Π.

Remark 2.1.2. *The form of the prior predictive distribution is the subject of de Finetti's theorem (see theorem A.2.2), which says that the distribution of a sequence $(X_1, \ldots, X_n)$ of random variables is of the form on the r.h.s. of the above display (with uniquely determined prior Π) if and only if the sample $(X_1, \ldots, X_n)$ is exchangeable, that is, if and only if the joint distribution for $(X_1, \ldots, X_n)$ equals that of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ for all permutations π of n elements.*

Remark 2.1.3. *We conclude the discussion of the distributions that play a role in Bayesian statistics with the following important point: at no stage during the derivation above, was an "underlying distribution of the data" used or needed! For comparison we turn to assumption (1.1), which is fundamental in the frequentist approach. More precisely, the assumption preceding (2.1) (c.f. the subjectivist Bayesian approach) is at odds with (1.1), unless*

$$P_0^n = P_n^\Pi = \int_\Theta P_\theta^n \, d\Pi(\theta),$$

Note, however, that the l.h.s. is a product-measure, whereas on the r.h.s. only exchangeability is guaranteed! (Indeed, the equality in the above display may be used as the starting point for definition of a goodness-of-fit criterion for the model and prior (see section 3.3). The discrepancy in the previous display makes the "pure" (e.g. subjectivist) Bayesian reluctant to assume the existence of a distribution P_0 for the sample.)

The distribution P_0 could not play a role in our analysis if we did not choose to adopt assumption (1.1). In many cases we shall assume that Y contains an *i.i.d.* sample of observations $X_1, X_2, \ldots, X_n$ where $X \sim P_0$ (so that $Y \sim P_0^n$). Indeed, if we would not make this assumption, asymptotic considerations like those in chapter 4 would be meaningless. However, adopting (1.1) leaves us with questions concerning the background of the quantities defined in this section because they originate from the subjectivist Bayesian framework.

Remark 2.1.4. (Bayesian/frequentist hybrid approach) *Maintaining the frequentist assumption that $Y \sim P_0$ for some P_0 requires that we revise our approach slightly: throughout the rest of these lecture notes, we shall assume (1.1) and require the model $\mathscr{P}$ to be a probability space $(\mathscr{P}, \mathscr{G}, \Pi)$ with a probability measure Π referred to as the prior. So the prior is introduced as a measure on the model, rather than emergent as a marginal to a product-space measure. Model and sample space are left in the separate roles they are assigned by the frequentist. We then proceed to* define *the posterior by expression (2.7). Regularity of the posterior is* imposed *(for a more detailed analysis, see Schervish (1995) [75] and Barron, Schervish and Wasserman (1999) [7]). In that way, we combine a frequentist perspective on statistics with Bayesian*

methodology: we make use of Bayesian quantities like prior and posterior, but analyze them from a frequentist perspective.

Remark 2.1.5. *In places throughout these lecture notes, probability measures P are decomposed into a P_0-absolutely-continuous part $P_{\|}$ and a P_0-singular part $P_{\perp}$. Following Le Cam, we use the convention that if P is not dominated by P_0, the Radon-Nikodym derivative refers to the P_0-absolutely-continuous part only: $dP/dP_0 = dP_{\|}/dP_0$. (See theorem A.4.2.) With this in mind, we write the posterior as follows*

$$\Pi(\,\vartheta \in A \,|\, X_1, X_2, \ldots, X_n\,) = \frac{\displaystyle\int_A \prod_{i=1}^n \frac{dP_\theta}{dP_0}(X_i)\, d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^n \frac{dP_\theta}{dP_0}(X_i)\, d\Pi(\theta)}, \quad (P_0^n - a.s.) \tag{2.10}$$

Since the data $X_1, X_2, \ldots$ are i.i.d.-P_0-distributed, the P_0-almost-sure version of the posterior in the above display suffices. Alternatively, any σ-finite measure that dominates P_0 may be used instead of P_0 in (2.10) while keeping the definition P_0^n-almost-sure. Such P_0-almost sure representations are often convenient when deriving proofs.

In cases where the model is not dominated, (2.10) may be used as the definition of the posterior measure but there is no guarantee that (2.10) leads to sensible results!

Example 2.1.3. *Suppose that the samplespace is $\mathbb{R}$ and the model $\mathscr{P}$ consists of all measures of the form (see example A.2.2):*

$$P = \sum_{j=1}^m \alpha_j \delta_{x_j}, \tag{2.11}$$

for some $m \geq 1$, with $\alpha_1, \ldots, \alpha_m$ satisfying $\alpha_j \geq 0$, $\sum_{j=1}^m \alpha_j = 1$ and $x_1, \ldots, x_m \in \mathbb{R}$. A suitable prior for this model exists: distributions drawn from the so-called Dirichlet process prior are of the form (2.11) with (prior) probability one. There is no σ-finite dominating measure for this model and hence the model can not be represented by a family of densities, c.f. definition 1.1.3. In addition, if the true distribution P_0 for the observation is also a convex combination of Dirac measures, distributions in the model are singular with respect to P_0 unless they happen to have support-points in common with P_0. Consequently definition (2.10) does not give sensible results in this case. We have to resort to the subjectivist definition (2.3) in order to make sense of the posterior distribution.

To summarize, the Bayesian procedure consists of the following steps

(i) Based on the background of the data Y, the statistician chooses a model $\mathscr{P}$, usually with some measurable parameterization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$.

(ii) A prior measure Π on $\mathscr{P}$ is chosen, based either on subjectivist or objectivist criteria. Usually a measure on Θ is defined, inducing a measure on $\mathscr{P}$.

(iii) Based on (2.3), (2.4) or in the case of an *i.i.d.* sample $Y = (X_1, X_2, \ldots, X_n)$, on:

$$d\Pi_n(\,\theta \,|\, X_1, X_2, \ldots, X_n\,) = \frac{\prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}{\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)},$$

we calculate the posterior density or posterior as a function of the data.

(iv) We observe a realization of the data $Y = y$ and use it to calculate a realisation of the posterior.

The statistician may then infer properties of the parameter θ from the posterior $\Pi(\,\cdot\,|\,Y = y\,)$, giving them a subjectivist or objectivist interpretation. One important point: when reporting the results of any statistical procedure, one is obliged to also reveal all relevant details concerning the procedure followed and the data. So in addition to inference on θ, the statistician should report on the nature and size of the sample used and, in the Bayesian case, should always report choice of model and prior as well, with a clear motivation.

2.2 Bayesian point estimators

When considering questions of statistical estimation, the outcome of a frequentist procedure is of a different nature than the outcome of a Bayesian procedure: a point-estimator (the frequentist outcome) is a point in the model, whereas the posterior is a distribution on the model. A first question, then, concerns the manner in which to compare the two. The connection between Bayesian procedures and frequentist (point-)estimation methods is provided by point-estimators derived from the posterior, called Bayesian point-estimators. Needless to say, comparison of frequentist and Bayesian point-estimators requires that we assume the "hybrid" point of view presented in remark 2.1.4.

We think of a reasonable Bayesian point-estimators as a point in the model around which posterior mass is accumulated most, a point around which the posterior distribution is concentrated in some way. As such, any reasonable Bayesian point-estimator should represent the *location* of the posterior distribution. But as is well known from probability theory, there is no unique definition of the "location" of a distribution. Accordingly, there are many different ways to define Bayesian point-estimators.

Remark 2.2.1. *Arguably, there are distributions for which even the* existence *of a "location" is questionable. For instance, consider the convex combination of point-masses* $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$ *on* $(\mathbb{R}, \mathscr{B})$. *Reasonable definitions of location, like the mean and the median of* P, *all assign as the location of* P *the point* $0 \in \mathbb{R}$. *Yet small neighbourhoods of* 0 *do not receive* any P*-mass, so* 0 *can hardly be viewed as a point around which* P *concentrates its mass. The intuitition of a distribution's location can be made concrete without complications of the above*

nature, if we restrict attention to unimodal distributions. However, it is common practice to formulate the notion for all distributions by the same definitions.

One quantity that is often used to represent a distribution's location is its expectation. This motivates the first definition of a Bayesian point-estimator: the posterior mean.

Definition 2.2.1. *Consider a statistical problem involving data Y taking values in a samplespace $(\mathscr{Y}, \mathscr{B})$ and a model $(\mathscr{P}, \mathscr{G})$ with prior Π. Assume that the maps $P \mapsto P(B)$, $(B \in \mathscr{B})$ are measurable with respect to $\mathscr{G}$ and that the posterior $\Pi(\,\cdot\,|Y)$ is regular, P_0^n-almost-surely. The posterior mean (or posterior expectation) is a probability measure $\hat{P} : \mathscr{B} \to [0,1]$, defined*

$$\hat{P}(B) = \int_{\mathscr{P}} P(B)\; d\Pi(\,P\,|\,Y\,), \tag{2.12}$$

P_0-almost-surely, for every event $B \in \mathscr{B}$.

Remark 2.2.2. *In order to justify the above definition, we have to show that $\hat{P}$ is a probability measure, P_0-almost-surely. Since the posterior is a regular conditional distribution, the map $B \mapsto \hat{P}(B)$ is defined P_0-almost-surely. Obviously, for all $B \in \mathscr{B}$, $0 \leq \hat{P}(B) \leq 1$. Let $(B_i)_{i\geq 1} \subset \mathscr{B}$ be a sequence of disjoint events. Since $(P,i) \mapsto P(B_i)$ is non-negative and measurable, Fubini's theorem applies in the third equality below:*

$$\begin{aligned}\hat{P}\Big(\bigcup_{i\geq 1} B_i\Big) &= \int_{\mathscr{P}} P\Big(\bigcup_{i\geq 1} B_i\Big)\, d\Pi(\,P\,|\,Y\,) = \int_{\mathscr{P}} \sum_{i\geq 1} P(B_i)\, d\Pi(\,P\,|\,Y\,) \\ &= \sum_{i\geq 1} \int_{\mathscr{P}} P(B_i)\, d\Pi(\,P\,|\,Y\,) = \sum_{i\geq 1} \hat{P}(B_i),\end{aligned}$$

which proves σ-additivity of $\hat{P}$.

Remark 2.2.3. *Note that, unless $\mathscr{P}$ happens to be convex, $\hat{P} \in \mathscr{P}$ is* not *guaranteed! In other words, the posterior mean may lie outside the model!*

In many practical situations, the model $\mathscr{P}$ is parametric with parameterization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$. In that case a different definition of the posterior mean can be made.

Definition 2.2.2. *Let $\mathscr{P}$ be a model parameterized by a convex subset Θ of $\mathbb{R}^d$. Let Π be a prior defined on Θ. If ϑ is integrable with respect to the posterior, the parametric posterior mean is defined*

$$\hat{\theta}_1(Y) = \int_{\Theta} \theta\; d\Pi(\,\theta\,|\,Y\,) \in \Theta, \tag{2.13}$$

P_0^n-almost-surely.

Remark 2.2.4. *The distinction between the posterior mean and the* parametric *posterior mean, as made above, is non-standard: it is customary in the Bayesian literature to refer to either as "the posterior mean". See, however, example 2.2.1.*

In definition 2.2.2, convexity of Θ is a condition (instead of an afterthought, as with definition 2.2.1): if Θ is not convex there is no guarantee that $\hat{\theta}_1 \in \Theta$, in which case $P_{\hat{\theta}_1}$ is not defined since $\hat{\theta}_1$ does not lie in the domain of the parameterization. Definition 2.2.2 can be extended to non-parametric models, *i.e.* models with an infinite-dimensional Θ. In that case, regularity of the posterior reappears as a condition and the condition of "integrability" of ϑ requires further specification.

It is tempting to assume that there is no difference between the posterior mean and the parametric posterior mean if the model is parametric and priors are brought in correspondence. This is not the case, however, as demonstrated by the following (counter)example.

Example 2.2.1. *Consider a normal location model in two dimensions for an observation Y, where the location $\mu \in \mathbb{R}^2$ lies on the unit circle and the covariance Σ is fixed and known:*

$$\mathscr{P} = \big\{P_\theta = N(\mu(\theta), \Sigma) \,:\, \mu(\theta) = (\cos\theta, \sin\theta),\, \theta \in [0, 2\pi)\big\}.$$

This is an identifiable, one-dimensional parametric model with convex parameterizing space $\Theta = [0, 2\pi)$. Assume that Ξ is the uniform distribution on Θ (Ξ plays the role of the posterior; it does not matter what shape the posterior really has, all we need is a counterexample). We define the corresponding measure Ξ' on $\mathscr{P}$ by applying Ξ to the pre-image of the parameterization. By rotational symmetry of Ξ and Fubini's theorem, the expectation of Y under $\hat{P}$ is

$$\int Y\, d\hat{P} = \int_{\mathscr{P}} PY\, d\Xi'(P) = \int_\Theta P_\theta Y\, d\Xi(\theta) = \frac{1}{2\pi}\int_0^{2\pi} \mu(\theta)\, d\theta = (0, 0).$$

Note that none of the distributions in $\mathscr{P}$ has the origin as its expectation. We can also calculate the expectation of Y under $P_{\hat{\theta}}$ in this situation:

$$\hat{\theta}_1(Y) = \int_\Theta \theta\, d\Xi(\theta) = \frac{1}{2\pi}\int_0^{2\pi} \theta\, d\theta = \pi,$$

which leads to $P_{\hat{\theta}}Y = P_\pi Y = (-1, 0)$. Clearly, the posterior mean does not equal the point in the model corresponding to the parametric posterior mean. In fact, we see from the above that $\hat{P} \notin \mathscr{P}$.

The fact that the expectations of $\hat{P}$ and $P_{\hat{\theta}}$ in example 2.2.1 differ makes it clear that

$$\hat{P} \neq P_{\hat{\theta}},$$

unless special circumstances apply: if we consider a parameterization $\theta \mapsto P_\theta$ from a (closed, convex) parameterizing space Θ with posterior measure $\Pi(d\theta)$ onto a space of probability measures $\mathscr{P}$ (with induced posterior $\Pi(dP)$), it makes a difference whether we consider the posterior mean as defined in (2.12), or calculate $P_{\hat{\theta}}$. The parametric posterior mean $P_{\hat{\theta}}$ lies in the model $\mathscr{P}$; $\hat{P}$ lies in the closed convex hull $\overline{\text{co}}(\mathscr{P})$ of the model, but not necessarily $\hat{P} \in \mathscr{P}$.

Since there are multiple ways of defining the location of a distribution, there are more ways of obtaining point-estimators from the posterior distribution. For example in a one-dimensional parametric model, we can consider the *posterior median* defined by

$$\tilde{\theta}(Y) = \inf\{s \in \mathbb{R} \,:\, \Pi(\vartheta \leq s|Y) \geq 1/2\},$$

i.e. the smallest value for θ such that the posterior mass to its left is greater than or equal to $1/2$. (Note that this definition simplifies in case the posterior has a continuous, strictly monotone distribution function: in that case the median equals the (unique) point where this distribution function equals $1/2$.) More generally, we consider the following class of point-estimators [63].

Definition 2.2.3. *Let $\mathscr{P}$ be a model with metric $d : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ and a prior Π on $\mathscr{G}$ containing the Borel σ-algebra corresponding to the metric topology on $\mathscr{P}$. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex loss-function $\ell : \mathbb{R} \to \mathbb{R}$. The* formal Bayes estimator *is the minimizer of the function:*

$$\mathscr{P} \to \mathbb{R} \,:\, P \mapsto \int_{\mathscr{P}} \ell(d(P,Q))\, d\Pi(\, Q \,|\, Y\,),$$

over the model $\mathscr{P}$ (provided that such a minimizer exists and is unique).

The heuristic idea behind formal Bayes estimators is decision-theoretic (see section 2.4). Ideally, one would like to estimate by a point P in $\mathscr{P}$ such that $\ell(d(P, P_0))$ is minimal; if $P_0 \in \mathscr{P}$, this would lead to $P = P_0$. However, lacking specific knowledge on P_0, we choose to represent it by averaging over $\mathscr{P}$ weighted by the posterior, leading to the notion in definition 2.2.3. Another useful point estimator based on the posterior is defined as follows.

Definition 2.2.4. *Let the data Y with model $\mathscr{P}$, metric d and prior Π be given. Suppose that the σ-algebra on which Π is defined contains the Borel σ-algebra generated by the metric topology. For given $\epsilon > 0$, the small-ball estimator is defined to be the maximizer of the function*

$$P \mapsto \Pi(\, B_d(P,\epsilon) \,|\, Y\,), \tag{2.14}$$

over the model, where $B_d(P,\epsilon)$ is the d-ball in $\mathscr{P}$ of radius ϵ centred on P (provided that such a maximizer exists and is unique).

Remark 2.2.5. *Similarly to definition 2.2.4, for a fixed value p such that $1/2 < p < 1$, we may define a Bayesian point estimator as the centre point of the smallest d-ball with posterior mass greater than or equal to p (if it exists and is unique (see also, exercise 2.6)).*

If the posterior is dominated by a σ-finite measure μ, the posterior density with respect to μ can be used as a basis for defining Bayesian point estimators.

Definition 2.2.5. *Let $\mathscr{P}$ be a model with prior Π. Assume that the posterior is absolutely continuous with respect to a σ-finite measure μ on $\mathscr{P}$. Denote the μ-density of $\Pi(\,\cdot\,|Y)$ by $\theta \mapsto \pi(\theta|Y)$. The maximum-a-posteriori estimator (or MAP-estimator, or posterior mode)*

$\hat{\theta}_2$ for θ is defined as the point in the model where the posterior density takes on its maximal value (provided that such a point exists and is unique):

$$\pi(\hat{\theta}_2|Y) = \sup_{\theta\in\Theta} \pi(\theta|Y). \tag{2.15}$$

Remark 2.2.6. *The MAP-estimator has a serious weak point: a different choice of dominating measure μ leads to a different MAP estimator! A MAP-estimator is therefore unspecified unless we specify also the dominating measure used to obtain a posterior density. It is with respect to this dominating measure that we define our estimator, so a motivation for the dominating measure used is inherently necessary (and often conspicuously lacking). Often the Lebesgue measure is used without further comment, or objective measures (see section 3.2) are used. Another option is to use the prior measure as the dominating measure, in which case the MAP estimator equals the maximum-likelihood estimator (see remark 2.2.7).*

All Bayesian point estimators defined above as maximizers or minimizers over the model suffer from the usual existence and uniqueness issues associated with extrema. However, there are straightforward methods to overcome such issues. We illustrate using the MAP-estimator. Questions concerning the existence and uniqueness of MAP-estimators should be compared to those of the existence and uniqueness of M-estimators in frequentist statistics. Although it is hard to formulate conditions of a general nature to guarantee that the MAP-estimator exists, often one can use the following lemma to guarantee existence.

Lemma 2.2.1. *Consider a parameterized model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ If the Θ is compact*[1] *and the posterior density $\theta \mapsto \pi(\theta|Y)$ is upper-semi-continuous (P_0^n-almost-surely) then the posterior density takes on its maximum in some point in Θ, P_0^n-almost-surely.*

To prove uniqueness, one has to be aware of various possible problems, among which are identifiability of the model (see section 1.1, in particular definition 1.1.5). Considerations like this are closely related to matters of consistency of M-estimators, *e.g.* Wald's consistency conditions for the maximum-likelihood estimator. The crucial property is called *well-separatedness* of the maximum, which says that outside neighbourhoods of the maximum, the posterior density must be uniformly strictly below the maximum. The interested reader is referred to chapter 5 of van der Vaart (1998) [83], *e.g.* theorems 5.7 and 5.8.

Remark 2.2.7. *There is an interesting connection between (Bayesian) MAP-estimation and (frequentist) maximum-likelihood estimation. Referring to formula (2.7) we see that in an i.i.d. experiment with parametric model, the MAP-estimator maximizes:*

$$\Theta \to \mathbb{R} : \theta \mapsto \prod_{i=1}^{n} p_\theta(X_i)\, \pi(\theta),$$

[1] Compactness of the model is a requirement that may be unrealistic or mathematically inconvenient in many statistical problems, especially when the model is non-parametric. However in a Bayesian context Ulam's theorem (see theorem A.2.3) offers a way to relax this condition.

where it is assumed that the model is dominated and that the prior has a density π with respect to the Lebesgue measure μ. If the prior had been uniform, the last factor would have dropped out and maximization of the posterior density is *maximization of the likelihood. Therefore, differences between ML and MAP estimators are entirely due to non-uniformity of the prior. Subjectivist interpretation aside, prior non-uniformity has an interpretation in the frequentist setting as well, through what is called penalized maximum likelihood estimation (see, Van de Geer (2000) [39]): Bayes' rule (see lemma A.6.2) applied to the posterior density $\pi_n(\theta|X_1,\ldots,X_n)$ gives:*

$$\log \pi_n(\theta|X_1,\ldots,X_n) = \log \pi_n(X_1,\ldots,X_n|\theta) + \log \pi(\theta) + D(X_1,\ldots,X_n),$$

where D is a (θ-independent, but stochastic) normalization constant. The first term equals the log-likelihood and the logarithm of the prior plays the role of a penalty term when maximizing over θ. Hence, maximizing the posterior density over the model Θ can be identified with maximization of a penalized likelihood over Θ. So defining a penalized MLE $\hat{\theta}_n$ with the logarithm of the prior density $\theta \mapsto \log \pi(\theta)$ in the role of the penalty, the MAP-estimator coincides with $\hat{\theta}_n$. The above offers a direct connection between Bayesian and frequentist methods of point-estimation. As such, it provides an frequentist interpretation of the prior as a penalty in the ML procedure. The asymptotic behaviour of the MAP-estimator is discussed in chapter 4 (see theorem 4.4.2).

2.3 Credible sets and Bayes factors

Besides point-estimation, frequentist statistics has several other inferential techniques at its disposal. The two most prominent are the analysis of confidence intervals and the testing of statistical hypotheses. Presently, it is assumed that the reader is familiar with these methods, but the essential reasoning is summarized for reference and comparison. The goal of this section is to formulate Bayesian analogs, so-called credible sets and Bayes factors respectively, and to compare them with aforementioned frequentist techniques.

Before we consider the Bayesian definitions, we briefly review the frequentist procedures. We assume that we have data Y and a parameterized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ such that $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. For simplicity, we assume that $\Theta \subset \mathbb{R}$ whenever the dimension of Θ is of importance.

We start with the central ideas and definitions that play a role in the Neyman-Pearson approach to statistical hypothesis testing. In this context, the hypotheses are presumptions one can make concerning the distribution of the data. Since the model contains all distributions the statistician is willing to consider as possibilities for P_0, the hypotheses are formulated in terms of a partition of the model (or its parameterization) into two disjoint subsets. One of them corresponds to the so-called null hypothesis and the other to the alternative hypothesis, which do *not* play a symmetric role in the Neyman-Pearson procedure. The goal of

Neyman-Pearson hypothesis testing is to find out whether or not the data contains "enough" evidence to reject the null hypothesis as a likely explanation when compared to alternative explanations. Sufficiency of evidence is formulated in terms of statistical significance.

For simplicity, we consider a so-called simple null hypothesis (*i.e.* a hypothesis consisting of only one point in the model, which is assumed to be identifiable): let a certain point $\theta_1 \in \Theta$ be given and consider the hypotheses:

$$H_0: \quad \theta_0 = \theta_1, \qquad H_1: \quad \theta_0 \neq \theta_1,$$

where H_0 denotes the null-hypothesis and H_1 the alternative. By no means does frequentist hypothesis testing equate to the corresponding *classification* problem, in which one would treat H_0 and H_1 symmetrically and make a choice for one or the other based on the data (for more on frequentist and Bayesian classification, see section 2.4).

To assess both hypotheses using the data, the simplest version of the Neyman-Pearson method of hypothesis testing seeks to find a test-statistic $T(Y) \in \mathbb{R}$ displaying different behaviour depending on whether the data Y is distributed according to (a distribution in) H_0 or in H_1. To make this distinction more precise, we define a so-called *critical region* $K \subset \mathbb{R}$, such that $P_{\theta_1}(T \in K)$ is "small" and $P_\theta(T \notin K)$ is "small" for all $\theta \neq \theta_1$. What we mean by "small" probabilities in this context is a *choice* for the statistician, a so-called significance level α is to be chosen to determine when these probabilities are deemed "small". That way, upon realization $Y = y$, a distribution in the hypothesis H_0 makes an outcome $T(y) \in K$ improbable compared to H_1.

Definition 2.3.1. *Let $\Theta \to \mathscr{P} : \theta \to P_\theta$ be a parameterized model for a sample Y. Formulate two hypotheses H_0 and H_1 by introducing a two-set partition $\{\Theta_0, \Theta_1\}$ of the model Θ:*

$$H_0: \quad \theta_0 \in \Theta_0, \qquad H_1: \quad \theta_0 \in \Theta_1.$$

We say that a test for these hypotheses based on a test-statistic T with critical region K is of level $\alpha \in (0,1)$ if the power function $\pi : \Theta \to [0,1]$, defined by

$$\pi(\theta) = P_\theta\big(T(Y) \in K\big),$$

is uniformly small over Θ_0:

$$\sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha. \tag{2.16}$$

From the above definition we arrive at the conclusion that if $Y = y$ and $T(y) \in K$, hypothesis H_0 is improbable enough to be rejected, since H_0 forms an "unlikely" explanation of observed data (at said significance level). The degree of "unlikeliness" can be quantified in terms of the so-called *p-value*, which is the lowest significance level at which the realised value of the test statistic $T(y)$ would have led us to reject H_0. Of course there is the possibility that our decision is wrong and H_0 is actually true but $T(y) \in K$ nevertheless, so that our rejection of the null hypothesis is unwarranted. This is called a *type-I error*; a *type-II error* is

made when we do *not* reject H_0 while H_0 is not true. The significance level α thus represents a fixed upper-bound for the probability of a type-I error. Having found a collection of tests displaying the chosen significance level, the Neyman-Pearson approach calls for subsequent minimization of the Type-II error probability, *i.e.* of all the pairs (T, K) satisfying (2.16), one prefers a pair that minimizes $P_\theta(T(Y) \notin K)$, ideally uniformly in $\theta \in \Theta_1$. However, generically such uniformly most-powerful tests do not exist due to the possibility that different (T, K) pairs are most powerful over distinct subsets of the alternative. The famed Neyman-Pearson lemma [60] asserts that a most powerful test exists in the case Θ contains only two points and can be extended to obtain uniformly most powerful tests in certain models.

We consider the Neyman-Pearson approach to testing in some more detail in the following example while also extending the argument to the asymptotic regime. Here Y is an *i.i.d.* sample and the test-statistic and critical region are dependent on the size n of this sample. We investigate the behaviour of the procedure in the limit $n \to \infty$.

Example 2.3.1. *Suppose that the data Y forms an i.i.d. sample from a distribution $P_0 = P_{\theta_0}$ and that $P_\theta X = \theta$ for all $\theta \in \Theta$. Moreover, assume that $P_\theta X^2 < \infty$ for all θ. Due to the law of large numbers, the sample-average*

$$T_n(X_1, \ldots, X_n) = \mathbb{P}_n X,$$

converges to θ under P_θ (for all $\theta \in \Theta$) and seems a suitable candidate for the test-statistic, at least in the regime where the sample-size n is large (i.e. asymptotically). The central limit theorem allows us to analyze matters in greater detail, for all $s \in \mathbb{R}$:

$$P_\theta^n\big(\mathbb{G}_n X \leq s\sigma(\theta)\big) \to \Phi(s), \qquad (n \to \infty). \tag{2.17}$$

where $\sigma(\theta)$ denotes the standard deviation of X under P_θ. For simplicity, we assume that $\theta \mapsto \sigma(\theta)$ is a known quantity in this derivation. The limit (2.17) implies that

$$P_\theta^n\big(T_n(X_1, \ldots, X_n) \leq \theta + n^{-1/2}\sigma(\theta)\, s\big) \to \Phi(s), \qquad (n \to \infty).$$

Assuming that H_0 holds, i.e. that $\theta_0 = \theta_1$, we then find that, given an asymptotic significance level $\alpha \in (0, 1)$ and with the standard-normal quantiles denoted s_α,

$$P_0^n\big(T_n(X_1, \ldots, X_2) \leq \theta_1 + n^{-1/2}\sigma(\theta_1)s_{\alpha/2}\big) \to 1 - \tfrac{1}{2}\alpha,$$

For significance levels close to zero, we see that under the null-hypothesis, it is improbable to observe $T_n > \theta_1 + n^{-1/2}\sigma(\theta_1)s_{\alpha/2}$. It is equally improbable to observe $T_n < \theta_1 - n^{-1/2}\sigma s_{\alpha/2}$, which means that we can take as our critical region $K_{n,\alpha}$

$$K_{n,\alpha} = \mathbb{R} \setminus [\theta_1 - n^{-1/2}\sigma(\theta_1)s_{\alpha/2}, \theta_1 + n^{-1/2}\sigma(\theta_1)s_{\alpha/2}],$$

(Note that this choice for the critical region is not unique unless we impose that it be an interval located symmetrically around θ_1.) Then we are in a position to formulate our decision on the null hypothesis, to reject H_0 or not:

(i) if $T_n \in K_{n,\alpha}$, we reject H_0 at significance level α, and,

(ii) if $T_n \notin K_{n,\alpha}$, we do not see enough evidence in the data to reject H_0 at significance level α.

Beware of a very common philosophical pitfall in the last case: even under case (ii), we do not *draw the conclusion that H_0 is accepted. The data does not provide enough evidence to reject the null hypothesis, but that does not imply that we have enough evidence to accept it!*

Note the behaviour of the procedure with varying sample-size: keeping the significance level fixed, the width of the critical regions $K_{n,\alpha}$ is of order $O(n^{-1/2})$, so smaller and smaller critical regions can be used as more information concerning the distribution P_0 (read, data) comes available. Similarly, if instead we keep the critical region fixed, the probability for a Type-I error (sometimes called the p-value if no fixed significance level is used) decreases with growing sample-size.

Strictly speaking the reasoning we follow here is not exact, because in practice n is finite and we are using a criterion based on the limit $n \to \infty$. At any finite n, the distribution of $\mathbb{G}_n X$ may not be close to $N(0, \sigma^2)$. In general we do not know which minimal sample-size n should be used in order for these distributions to be "sufficiently close". Nevertheless, it is common practice to use asymptotic tests like this one, in cases where the sample-size is deemed to be large enough and the test-statistic is expected to assume its asymptotic behaviour behaviour in close approximation.

It is important to stress that our criterion for tests is entirely geared at minimizing the probability of rejecting H_0 when in fact H_0 contains the true distribution. As such, the testing procedure we follow can only lead to one definite conclusion, *rejection* of the null hypothesis. The inverse conclusion, *acceptance* of the null hypothesis, is never the result. Therefore, it is crucial that we choose the null hypothesis to be an assertion that we would like to disprove. In practice, one also tries to find a test such that the probability of *not* rejecting H_0 when it is *not* valid is also small. Before we formalize the latter, we generalize the concepts introduced above in this section somewhat.

Note that the indicators for the events $\{Y \in \mathscr{Y} : T_n(Y) \in K_n\}$ form a (bounded, positive) sequence of random variables, on which we base the decision to reject H_0 or not. The power functions $\pi_n : \Theta \to [0,1]$ are simply the P_θ-expectations of these random variables.

Definition 2.3.2. *Let $\mathscr{P}$ be a model for a sample $X_1, X_2 \ldots$ taking values in $\mathscr{X}$ and assume that the true distribution of the data lies in the model, $(X_1, X_2, \ldots) \sim P_0 \in \mathscr{P}$. Formulate two hypotheses H_0 and H_1 by introducing a two-set partition $\{\mathscr{P}_0, \mathscr{P}_1\}$ of the model $\mathscr{P}$:*

$$H_0 : \quad P_0 \in \mathscr{P}_0, \qquad H_1 : \quad P_0 \in \mathscr{P}_1.$$

A test sequence $(\phi_n)_{n\geq 1}$ is a sequence of statistics $\phi_n : \mathscr{X}^n \to [0,1]$, (for all $n \geq 1$). An asymptotic test is defined as a criterion for the decision to reject H_0 or not, based on (a realization of) $\phi_n(X_1, \ldots, X_n)$ and is studied in the limit $n \to \infty$.

An example of such a criterion is the procedure given in definition 2.3.1 and example 2.3.1, where test-functions take on the values zero or one depending on the (realized) test-statistic and the critical region. When we replace indicators by test functions as in definition 2.3.2 criteria may vary depending on the nature of the test functions used.

Definition 2.3.3. *Extending definition 2.3.2, we define the power function sequence of the test sequence* (ϕ_n) *as a map* $\pi_n : \mathscr{P} \to [0,1]$ *on the model defined by:*

$$\pi_n(P) = P\phi_n.$$

Like in definition 2.3.1, the quality of the test depends on the behaviour of the power sequence on $\mathscr{P}_0$ and $\mathscr{P}_1$ respectively. If we are interested exclusively in rejection of the null hypothesis, we could reason like in definition 2.3.1 and set a significance level α to select only those test sequences that satisfy

$$\sup_{P \in \mathscr{P}_0} \pi_n(P) \leq \alpha.$$

Subsequently, we prefer test sequences that have high power on the alternative. For example, if we have two test sequences (ϕ_n) and (ψ_n) and a point $P \in \mathscr{P}_1$ such that

$$\lim_{n\to\infty} P\phi_n \geq \lim_{n\to\infty} P\psi_n, \tag{2.18}$$

then (ϕ_n) is said to be asymptotically more powerful than (ψ_n) at P. If (2.18) holds for *all* $P \in \mathscr{P}_1$, the test sequence (ϕ_n) is said to be uniformly asymptotically more powerful than (ψ_n). If one can show that this holds for all test sequences (ψ_n), then (ϕ_n) is said to be uniformly asymptotically most powerful. Note, however, that the above ordering of test sequences is not complete: it is quite possible that (ϕ_n) is asymptotically more powerful than (ψ_n) on a subset of $\mathscr{P}_1$, whereas on its complement in $\mathscr{P}_1$, (ψ_n) is asymptotically more powerful. As a result, uniformly most powerful tests do not exist in many problems.

Besides providing a criterion for rejection of a null hypothesis, test sequences may be used to indicate whether the true distribution of the data resides in $\mathscr{P}_0$ or $\mathscr{P}_1$ (where now, $\mathscr{P}_0$ and $\mathscr{P}_1$ are disjoint but may not cover all of $\mathscr{P}$). This requires that we treat H_0 and H_1 on a symmetrical footing, much like in a classification problem. For that purpose, one would like to consider test sequences (ϕ_n) such that the quantity

$$\sup_{P \in \mathscr{P}_0} P\phi_n + \sup_{P \in \mathscr{P}_1} P(1 - \phi_n), \tag{2.19}$$

(which is sometimes also referred to as "the power function") is "small" in the limit $n \to \infty$, possibly quantified by introduction of a significance level pertaining to both type-I and type-II errors simultaneously. In many proofs of Bayesian limit theorems (see chapter 4), a test sequence (ϕ_n) is needed such that (2.19) goes to zero, or is bounded by a sequence (a_n) decreasing to zero (typically $a_n = e^{-nD}$ for some $D > 0$). The existence of such test sequences forms the subject of section 4.5.

Closely related to hypothesis tests are confidence intervals. Suppose that pose our inferential problem differently: our interest now lies in using the data $Y \sim P_0$ to find a data-dependent subset $C(Y)$ of the model that contains P_0 with "high" probability. Again, "high" probability requires quantification in terms of a level α, called the confidence level.

Definition 2.3.4. *Let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ be a parameterized model; let $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. Choose a confidence level $\alpha \in (0,1)$. Let $C(Y)$ be subset of Θ dependent only on the data Y. Then $C(Y)$ is a confidence region for θ of confidence level α, if*

$$P_\theta\big(\theta \in C(Y)\big) \geq 1 - \alpha, \tag{2.20}$$

for all $\theta \in \Theta$.

The dependence of C on the data Y is meant to express that $C(Y)$ can be calculated once the data has been observed. The confidence region may also depend on other quantities that are known to the statistician, so $C(Y)$ is a *statistic* (see definition 1.1.9). Note also that the dependence of $C(Y)$ on the data Y makes $C(Y)$ a *random* subset of the model. Compare this to point estimation, in which the data-dependent estimator is a *random* point in the model.

Like hypothesis testing, confidence regions can be considered from an asymptotic point of view, as demonstrated in the following example.

Example 2.3.2. *We consider the experiment of example 2.3.1, i.e. we suppose that the data Y forms an i.i.d. sample from a distribution $P_0 = P_{\theta_0}$ or $\mathbb{R}$ and that $P_\theta X = \theta$ for all $\theta \in \Theta$. Moreover, we assume that for some known constant $S > 0$, $\sigma^2(\theta) = \mathrm{Var}_\theta X \leq S^2$, for all $\theta \in \Theta$. Consider the sample-average $T_n(X_1, \ldots, X_n) = \mathbb{P}_n X$. Choose a confidence level $\alpha \in (0,1)$. The limit (2.17) can be rewritten in the following form:*

$$P_\theta^n\big(|T(X_1, \ldots, X_n) - \theta| \leq n^{-1/2}\sigma(\theta)s_{\alpha/2}\big) \to 1 - \alpha, \qquad (n \to \infty). \tag{2.21}$$

Define C_n by

$$C_n(X_1, \ldots, X_n) = \big[T(X_1, \ldots, X_n) - n^{-1/2}Ss_{\alpha/2}, T(X_1, \ldots, X_n) + n^{-1/2}Ss_{\alpha/2}\big].$$

Then, for all $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta^n\big(\theta \in C_n(X_1, \ldots, X_n)\big) \geq 1 - \alpha.$$

Note that the θ-dependence of $\sigma(\theta)$ would violate the requirement that C_n be a statistic: since the true value θ_0 of θ is unknown, so is $\sigma(\theta)$. Substituting the (known) upper-bound S for $\sigma(\theta)$ enlarges the $\sigma(\theta)$-interval that follows from (2.21) to its maximal extent, eliminating the θ-dependence. In a realistic situation, one would not use S but substitute $\sigma(\theta)$ by an estimator $\hat{\sigma}(Y)$, which amounts to the use of a plug-in version of (2.21). As a result, we would also have to replace the standard-normal quantiles s_α by the quantiles of the Student t-distribution.

Clearly, confidence regions are not unique, but of course small confidence regions are more informative than large ones: if, for some confidence level α, two confidence regions $C(Y)$ and

$D(Y)$ are given, both satisfying (2.20) for all $\theta \in \Theta$, and $C(Y) \subset D(Y)$, P_θ-almost-surely for all θ, then $C(Y)$ is preferred over $D(Y)$.

The Bayesian analogs of tests and confidence regions are called Bayes factors and credible regions, both of which are derived from the posterior distribution. We start by considering credible sets. The rationale behind their definition is exactly the same one that motivated confidence regions: we look for a subset D of the model that is as small as possible, while receiving a certain minimal probability. Presently, however, the word "probability" is in line with the Bayesian notion, *i.e.* probability according to the posterior distribution.

Definition 2.3.5. *Let $(\Theta, \mathscr{G})$ be a measurable space parameterizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data Y, with prior $\Pi : \mathscr{G} \to [0,1]$. Choose a level $\alpha \in (0,1)$. Let $D \in \mathscr{G}$ be a subset of Θ. Then D is a level-α credible set for ϑ, if*

$$\Pi\big(\vartheta \in D \;\big|\; Y\big) \geq 1 - \alpha. \tag{2.22}$$

In a Bayesian setting, one interprets $\Pi(\vartheta \in D|Y)$ as the probability of finding ϑ in D, given the data. Note that credible sets are *random* sets, since they are defined based on the posterior which, in turn, depends on the sample: this data-dependence can be made explicit by writing credible sets as $D(Y)$ instead of D. In practice, one calculates the posterior distribution from the prior and the data and, based on that, proceeds to derive a subset $D(Y)$ such that (2.22) is satisfied. A credible set is sometimes referred to as a credible region, or, if D is an interval in a one-dimensional parametric model, a credible interval.

Remark 2.3.1. *In smooth, parametric models for i.i.d. data there is an close, asymptotic relation between Bayesian credible sets and frequentist confidence intervals centred on the maximum-likelihood estimator: the Bernstein-von Mises theorem (see section 4.4) implies that level-α credible regions coincide with abovementioned level-α confidence intervals asymptotically! In situations where it is hard to calculate the ML estimator or to construct the corresponding confidence interval explicitly, it is sometimes relatively easy to obtain credible regions (based on a simulated sample from the posterior, as obtained from the MCMC procedure (see section 6.1)). In such cases, one can calculate credible regions and conveniently interpret them as confidence intervals centred on the MLE, due to theorem 4.4.1.*

Definition 2.3.5 suffices to capture the concept of a credible set, but offers too much freedom in the choice of D: given a level $\alpha > 0$, many sets will satisfy (2.22), just like confidence regions can be chosen in many different ways. Note that, also here, we prefer smaller sets over large ones: if, for some level α, two different level-α credible sets F and G are given, both satisfying (2.22) and $F \subset G$, then F is preferred over G. If the posterior is dominated with density $\theta \mapsto \pi(\theta|Y)$, we can be more specific. We define, for every $k \geq 0$, the level-set

$$D(k) = \big\{\theta \in \Theta \,:\, \pi(\theta|Y) \geq k\big\}, \tag{2.23}$$

and consider the following.

Definition 2.3.6. *Let $(\Theta, \mathscr{G})$ a measurable space parameterizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior $\Pi : \mathscr{G} \to [0,1]$. Assume that the posterior is dominated by a σ-finite measure μ on $(\Theta, \mathscr{G})$, with density $\pi(\cdot|Y) : \Theta \to \mathbb{R}$. Choose $\alpha \in (0,1)$. A level-α HPD-credible set (from* highest posterior density*) for ϑ is the subset $D_\alpha = D(k_\alpha)$, where k_α equals:*

$$k_\alpha = \sup\big\{k \geq 0 \,:\, \Pi(\vartheta \in D(k)|Y) \geq 1-\alpha\big\}.$$

In other words, D_α is the smallest level-set of the posterior density that receives posterior mass greater than or equal to $1-\alpha$. Note that HPD-credible sets depend on the choice of dominating measure: if we had chosen to use a different measure μ, HPD-credible sets would have changed as well! One may wonder what happens if the posterior is dominated by the prior and we use the density of the posterior with respect to the prior to define HPD-credible regions.

Lemma 2.3.1. *Let $(\Theta, \mathscr{G})$ a measurable space parameterizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data Y taking values in a measurable space $(\mathscr{Y}, \mathscr{B})$. Assume that the model is dominated by some σ-finite measure $\nu : \mathscr{B} \to \mathbb{R}$, with $p_\theta : \mathscr{Y} \to \mathbb{R}$ is the ν-density of P_θ for every $\theta \in \Theta$. Let $\Pi_1, \Pi_2 : \mathscr{G} \to [0,1]$ be two priors, such that $\Pi_1 \ll \Pi_2$ and $\Pi_2 \ll \Pi_1$. Denote the posterior densities with respect to Π_1, Π_2 as $\pi_1(\cdot|Y), \pi_2(\cdot|Y) : \mathscr{Y} \to \mathbb{R}$ and corresponding HPD-credible sets as $D_{1,\alpha}$, $D_{2,\alpha}$. Then*

$$D_{1,\alpha} = D_{2,\alpha},$$

for all $\alpha \in (0,1)$.

Proof Under the conditions stated, the densities $\theta \mapsto \pi_1(\theta|Y)$ and $\theta \mapsto \pi_2(\theta|Y)$ are both of the form (2.8). Note that both $\pi_1(\theta|Y)$ and $\pi_1(\theta|Y)$ are almost-sure expressions with respect to their respective priors, but since $\Pi_1 \ll \Pi_2$ and $\Pi_2 \ll \Pi_1$ by assumption, Π_1-almost-sureness and Π_2-almost-sureness are equivalent. From (2.8), we see that

$$\frac{\pi_1(\theta|Y)}{\pi_2(\theta|Y)} = \frac{\displaystyle\int_\Theta p_\theta(Y)\, d\Pi_2(\theta)}{\displaystyle\int_\Theta p_\theta(Y)\, d\Pi_1(\theta)} = K(Y) > 0,$$

almost-surely with respect to both priors (and P_0). So the fraction of posterior densities is a positive constant as a function of θ. Therefore, for all $k \geq 0$,

$$D_1(k) = \big\{\theta \in \Theta \,:\, \pi_1(\theta|Y) \geq k\big\} = \big\{\theta \in \Theta \,:\, \pi_2(\theta|Y)\, K(Y) \geq k\big\} = D_2\big(K(Y)^{-1}k\big).$$

and, hence, for all $\alpha \in (0,1)$,

$$k_{1,\alpha} = \sup\big\{k \geq 0 \,:\, \Pi(\vartheta \in D_2(K(Y)^{-1}k)|Y) \geq 1-\alpha\big\} = K(Y)\, k_{2,\alpha}.$$

To conclude,

$$D_{1,\alpha} = D_1(k_{1,\alpha}) = D_2\big(K(Y)^{-1}k_{1,\alpha}\big) = D_2(k_{2,\alpha}) = D_{2,\alpha}.$$

□

The above lemma proves that using the posterior density with respect to the prior leads to HPD-credible sets that are independent of the choice of prior. This may be interpreted further, by saying that *only the data* is of influence on HPD-credible sets based on the posterior density with respect to the prior. Such a perspective is attractive to the objectivist, but rather counterintuitive from a subjectivist point of view: a prior chosen according to subjectivist criteria places high mass in subsets of the model that the statistician attaches "high belief" to. Therefore, the density of the posterior with respect to the prior can be expected to be relatively *small* in those subsets! As a result, those regions may end up in D_α only for relatively high values of α. However, intuition is to be amended by mathematics in this case: when we say above that only the data is of influence, this is due entirely to the likelihood factor in (2.8). Rather than incorporating both prior knowledge and data in HPD credible sets, the above construction emphasizes the *differences* between prior and posterior beliefs, which lie entirely in the data and are represented in the formalism by the likelihood. (We shall reach a similar conclusion when considering the difference between posterior odds and Bayes factors later in this section). To present the same point from a different perspective, HPD credible regions based on the posterior density with respect to the prior coincide with levelsets of the likelihood and centre on the ML estimate if the likelihood is smooth enough and has a well-separated maximum (as a function on the model). We shall see that the coincidence between confidence regions and credible sets becomes more pronounced in the large-sample limit when we study the Bernstein-Von Mises theorem (see chapter 4 for more on large-sample limiting behaviour of the posterior).

Bayesian hypothesis testing is formulated in a far more straightforward fashion than frequentist methods based on the Neyman-Pearson approach. The two hypotheses H_0 and H_1 correspond to a two-set partition $\{\Theta_0, \Theta_1\}$ of the model Θ and for each of the parts, we have both posterior and prior probabilities. Based on the proportions between those, we shall decide which hypothesis is the more likely one. It can therefore be remarked immediately that in the Bayesian approach, the hypotheses are treated on *equal* footing, a situation that is more akin to classification than to Neyman-Pearson hypothesis testing. To introduce Bayesian hypothesis testing, we make the following definitions.

Definition 2.3.7. *Let $(\Theta, \mathscr{G})$ a measurable space parameterizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior $\Pi : \mathscr{G} \to [0,1]$. Let $\{\Theta_0, \Theta_1\}$ be a partition of Θ such that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$. The prior and posterior odds ratios are defined by $\Pi(\Theta_0)/\Pi(\Theta_1)$ and $\Pi(\Theta_0|Y)/\Pi(\Theta_1|Y)$ respectively. The Bayes factor in favour of Θ_0 is defined to be*

$$B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)}.$$

When doing Bayesian hypothesis testing, we have a choice of which ratio to use and that choice will correspond directly with a choice for subjectivist or objectivist philosophies. In

the subjectivist's view, the posterior odds ratio has a clear interpretation: if

$$\frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} > 1,$$

then the probability of $\vartheta \in \Theta_0$ is greater than the probability of $\vartheta \in \Theta_0$ and hence, the subjectivist decides to adopt H_0 rather than H_1. If, on the other hand, the above display is smaller than 1, the subjectivist decides to adopt H_1 rather than H_0. The objectivist would object to this, saying that the relative prior weights of Θ_0 and Θ_1 can introduce a heavy bias in favour of one or the other in this approach (upon which the subjectivist would answer that that is exactly what he had in mind). Therefore, the objectivist would prefer to use a criterion that is less dependent on the prior weights of Θ_0 and Θ_1. We look at a very simple example to illustrate the point.

Example 2.3.3. *Let Θ be a dominated model that consists of only two points, θ_0 and θ_1 and let $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$, corresponding to simple null and alternative hypotheses H_0, H_1. Denote the prior by Π and assume that both $\Pi(\{\theta_0\}) > 0$ and $\Pi(\{\theta_1\}) > 0$. By Bayes rule, the posterior weights of Θ_0 and Θ_1 are*

$$\Pi(\vartheta \in \Theta_i|Y) = \frac{p_{\theta_i}(Y)\Pi(\Theta_i)}{p_{\theta_0}(Y)\Pi(\Theta_0) + p_{\theta_1}(Y)\Pi(\Theta_1)},$$

for $i = 0, 1$. Therefore, the posterior odds ratio takes the form:

$$\frac{\Pi(\vartheta \in \Theta_0|Y)}{\Pi(\vartheta \in \Theta_1|Y)} = \frac{p_{\theta_0}(Y)\Pi(\Theta_0)}{p_{\theta_1}(Y)\Pi(\Theta_1)},$$

and the Bayes factor equals the likelihood ratio:

$$B = \frac{p_{\theta_0}(Y)}{p_{\theta_1}(Y)}.$$

We see that the Bayes factor does not depend on the prior weights assigned to Θ_0 and Θ_1 (in this simple example), but the posterior odds ratio does. Indeed, suppose we stack the prior odds heavily in favour of Θ_0, by choosing $\Pi(\Theta_0) = 1-\epsilon$ and $\Pi(\Theta_1) = \epsilon$ (for some small $\epsilon > 0$). Even if the likelihood ratio $p_{\theta_0}(Y)/p_{\theta_1}(Y)$ is much smaller than one (but greater than $\epsilon/1-\epsilon$), the subjectivist's criterion favours H_0. In that case, the data clearly advocates hypothesis H_1 but the prior odds force adoption of H_0. The Bayes factor B equals the likelihood ratio (in this example), so it does not suffer from the bias imposed on the posterior odds.

The objectivist prefers the Bayes factor to make a choice between two hypotheses: if $B > 1$ the objectivist adopts H_0 rather than H_1; if, on the other hand, $B < 1$, then the objectivist adopts H_1 rather than H_0. In example 2.3.3 the Bayes factor is independent of the choice of the prior. In general, the Bayes factor is not completely independent of the prior, but it does not depend on the relative prior weights of Θ_0 and Θ_1. We prove this using the following decomposition of the prior:

$$\Pi(A) = \Pi(A|\Theta_0)\,\Pi(\Theta_0) + \Pi(A|\Theta_1)\,\Pi(\Theta_1), \tag{2.24}$$

for all $A \in \mathscr{G}$ (where it is assumed that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$). In the above display, $\Pi(\,\cdot\,|\Theta_i)$ can be any probability measure on Θ_i $(i = 0, 1)$, and since $\Pi(\Theta_0) + \Pi(\Theta_1) = 1$, Π is decomposed as a convex combination of two probability measures on Θ_0 and Θ_1 respectively. The Bayes factor is then rewritten using Bayes' rule (see lemma A.6.1):

$$B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)} = \frac{\Pi(Y|\Theta_0)}{\Pi(Y|\Theta_1)},$$

where, in a dominated model,

$$\Pi(Y|\Theta_i) = \int_{\Theta_i} p_\theta(Y)\, d\Pi(\theta|\Theta_i),$$

for $i = 0, 1$. In terms of the decomposition (2.24), B depends on $\Pi(\,\cdot\,|\Theta_0)$ and $\Pi(\,\cdot\,|\Theta_1)$, but not on $\Pi(\Theta_0)$ and $\Pi(\Theta_1)$. So using Bayes factors instead of posterior odds exactly eliminates the bias introduced by non-zero prior odds.

Remark 2.3.2. *The condition that both Θ_0 and Θ_1 receive prior mass strictly above zero is important since Bayes factors and odds ratios are based on conditioning of ϑ. Bayesian hypothesis testing is sensible* only *if both Θ_0 and Θ_1 receive non-zero prior mass. This remark plays a role particularly when comparing a* simple *null hypothesis to an alternative, as illustrated in exercise 2.10.*

2.4 Decision theory and classification

Many practical problems require that we make an observation and based on the outcome, make a decision of some kind. For instance when looking for the diagnosis for a patient, a doctor will observe variables like the patients temperature, blood-pressure and appearance, in addition to the results of chemical and physical scans to come to a decision regarding the affliction the patient is probably suffering from. Another example concerns the financial markets, in which past stock- and option-prices are considered by analysts to decide whether to buy or sell stocks and derivatives. In a chemical plant, regulation of a chemical process amounts to a succession of decisions to control and optimize conditions, based on the measurement of thermo-dynamical quantities and concentrations of chemicals involved in the reaction. In this section, we look at problems of this nature, first from a frequentist perspective and then with the Bayesian approach.

Practical problems like those described above usually involve optimality criteria that are prescribed by the context of the problem itself: for example, when a doctor makes the wrong diagnosis for a patient suffering from cancer the consequences can be most serious, whereas the misdiagnosis of a case of influenza is usually no more than unfortunate. In any useful statistical procedure meant to assist in medical diagnosis, such differences should be reflected in the decision-making procedure. That is certainly not the case for the methods that we have discussed thus far. Up to this point, we have used optimality criteria of a more general nature,

like the accuracy of an estimation procedure, coverage probabilities for confidence intervals or the probability of Type-I and type-II errors in a testing procedure.

The distinction lies in the nature of the optimality criteria: so far we have practiced what is called statistical inference, in which optimality is formulated entirely in terms of the stochastic description of the data. For that reason, it is sometimes said that statistical inference limits itself to those questions that "summarize the data". By contrast, *statistical decision theory* formalizes the criteria for optimality by adopting the use of a so-called loss-function to quantify the consequences of wrong decisions in a way prescribed by the context of the statistical problem.

In statistical decision theory the nomenclature is slightly different from that introduced earlier. We consider a system that is in an unknown *state* $\theta \in \Theta$, where Θ is called the *state-space*. The observation Y takes its values in the *samplespace* $\mathscr{Y}$, a measurable space with σ-algebra $\mathscr{B}$. The observation is stochastic, its distribution $P_\theta : \mathscr{B} \to [0,1]$ being dependent on the state θ of the system. The observation does not reveal the state of the system completely or with certainty. Based on the outcome $Y = y$ of the observation, we take a *decision* $a \in \mathscr{A}$ (or perform an *action* a, as some prefer to say), where $\mathscr{A}$ is the called the *decision-space*. For each state θ of the system there may be an optimal or prescribed decision, but since observation of Y does not give us the state θ of the system with certainty, the decision is stochastic and may be wrong. The goal of statistical decision theory is to arrive at a rule that decides in the best possible way given only the data Y.

The above does not add anything new to the approach we were already following: aside from the names, the concepts introduced here are those used in the usual problem of statistically estimating $a \in \mathscr{A}$. Decision theory distinguishes itself through its definition of optimality in terms of a so-called loss-function.

Definition 2.4.1. *Any lower-bounded function $L : \Theta \times \mathscr{A} \to \mathbb{R}$ may serve as a loss-function. The utility-function is $-L : \Theta \times \mathscr{A} \to \mathbb{R}$.*

(Although statisticians talk about loss-functions, people in applied fields often prefer to talk of utility-functions, which is why the above definition is given both in a positive and a negative version.) The interpretation of the loss-function is the following: if a particular decision a is taken while the state of the system is θ, then a loss $L(\theta, a)$ is incurred which can be either positive (loss) or negative (profit). To illustrate, in systems where observation of the state is direct (*i.e.* $Y = \theta$) and non-stochastic, the optimal decision $a(\theta)$ given the state θ is the value of a that minimizes the loss $L(\theta, a)$. However, the problem we have set is more complicated because the state θ is unknown and can not be measured directly. All we have is the observation Y.

Definition 2.4.2. *Let $\mathscr{A}$ be a measurable space with σ-algebra $\mathscr{H}$. A measurable $\delta : \mathscr{Y} \to \mathscr{A}$ is called a decision rule.*

A decision-rule is an automated procedure to arrive at a decision $\delta(y)$, given that the observation is $Y = y$. We denote the collection of all decision rules under consideration by Δ. Clearly our goal will be to find decision rules in Δ that "minimize the loss" in an appropriate sense. The above basic ingredients of decision-theoretic problems play a role in both the frequentist and Bayesian analysis. We consider the frequentist approach first and then look at decision theory from a Bayesian perspective.

In frequentist decision theory we assume that $Y \sim P_{\theta_0}$ for some state $\theta_0 \in \Theta$ and we analyze the expectation of the loss.

Definition 2.4.3. *The risk-function $R : \Theta \times \Delta \to \mathbb{R}$ is defined as the expected loss under $Y \sim P_\theta$ when using δ,*

$$R(\theta, \delta) = \int L(\theta, \delta(Y))\, dP_\theta. \tag{2.25}$$

Of interest to the frequentist is only the expected loss under the true distribution $Y \sim P_{\theta_0}$. But since θ_0 is unknown, we are forced to consider *all* values of θ, *i.e.* look at the risk-*function* $\theta \mapsto R(\theta, \delta)$ for each decision rule δ.

Definition 2.4.4. *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. Choose $\delta_1, \delta_2 \in \Delta$. The decision rule δ_1 is R-better than δ_2, if*

$$\forall_{\theta \in \Theta} : \quad R(\theta, \delta_1) < R(\theta, \delta_2). \tag{2.26}$$

A decision rule δ is admissible if there exists no $\delta' \in \Delta$ that is R-better than δ (and inadmissible if such a δ' does exist).

It is clear that the definition of R-better decision-rules is intended to order decision rules: if the risk-function associated with a decision-rule is relatively small, then that decision rule is preferable. Note, however, that the ordering we impose by definition 2.4.4 may be partial rather than complete: pairs δ_1, δ_2 of decision rules may exist such that neither δ_1 nor δ_2 is R-better than the other. This is due to the fact that δ_1 may perform better (in the sense that $R(\theta, \delta_1) \leq R(\theta, \delta_2)$) for values of θ in some $\Theta_1 \subset \Theta$, while δ_2 performs better in $\Theta_2 = \Theta \setminus \Theta_1$, resulting in a situation where (2.26) is true for neither. For that reason, it is important to find a way to compare risks (and thereby decision rules) in a θ-independent way and thus arrive at a complete ordering among decision rules. This motivates the following definition.

Definition 2.4.5. (Minimax decision principle) *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. The function*

$$\Delta \to \mathbb{R} : \delta \mapsto \sup_{\theta \in \Theta} R(\theta, \delta)$$

is called the minimax risk. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule δ_1 is minimax-preferred to δ_2, if

$$\sup_{\theta \in \Theta} R(\theta, \delta_1) < \sup_{\theta \in \Theta} R(\theta, \delta_2).$$

If $\delta^M \in \Delta$ *minimizes* $\delta \mapsto \sup_\theta R(\theta, \delta)$*, i.e.*

$$\sup_{\theta\in\Theta} R(\theta, \delta^M) = \inf_{\delta\in\Delta} \sup_{\theta\in\Theta} R(\theta, \delta). \tag{2.27}$$

then δ^M *is called a minimax decision-rule.*

Regarding the existence of minimax decision rules, it is noted that the Minimax theorem (see Strasser (1985) [81]) asserts existence of δ^M and moreover, that

$$\inf_{\delta\in\Delta} \sup_{\theta\in\Theta} R(\theta, \delta) = \sup_{\theta\in\Theta} \inf_{\delta\in\Delta} R(\theta, \delta).$$

under the conditions that R is convex on Δ, concave on Θ and that the topology on Δ is such that Δ is compact, $\delta \mapsto R(\theta, \delta)$ is continuous for all θ. Since many loss-functions used in practice satisfy the convexity requirements, the Minimax theorem has broad applicability in statistical decision theory and many other fields. In some cases, use of the minimax theorem requires that we extend the class Δ to contain more general decision rules. Particularly, it is often necessary to consider the class of all so-called *randomized* decision rules. Randomized decision rules are not only stochastic in the sense that they depend on the data, but also through a further stochastic influence: concretely, this means that after realisation $Y = y$ of the data, uncertainty in the decision remains. To give a formal definition, consider a measurable space $(\Omega, \mathscr{F})$ with data $Y : \Omega \to \mathscr{Y}$ and a decision rule $\delta : \Omega \to \mathscr{A}$. The decision rule δ is a randomized decision rule whenever $\sigma(\delta)$ is not a subset of $\sigma(Y)$, *i.e.* δ is not a function of Y. An example of such a situation is that in which we entertain the possibility of using one of two different non-randomized decision rules $\delta_1, \delta_2 : \mathscr{Y} \to \mathscr{A}$. After the data is realised as $Y = y$, δ_1 and δ_2 give rise to two decisions $\delta_1(y)$, $\delta_2(y)$, which may differ. In that case, we flip a coin with outcome $C \in \{0, 1\}$ to decide which decision to use. The extra stochastic element introduced by the coin-flip has then "randomized" our decision rule. The product space $\mathscr{Y} \times \{0, 1\}$ endowed with the product σ-algebra may serve as the measurable space $(\Omega, \mathscr{F})$ with $\delta : \Omega \to \mathscr{Y}$ defined by,

$$(y, c) \mapsto \delta(y, c) = c\,\delta_1(Y) + (1 - c)\,\delta_2(y),$$

for all $y \in \mathscr{Y}$ and $c \in \{0, 1\}$. Perhaps a bit counterintuitively (but certainly in accordance with the fact that minimization over a larger set produces a lower infimum), in some decision problems the minimax risk associated with such randomized decision rules lies strictly below the minimax risks of both non-randomized decision rules. We return to the Minimax theorem in section 4.3.

Example 2.4.1. (Decision theoretic L_2-estimation) *The decision-theoretic approach can also be used to formulate estimation problems in a generalized way, if we choose the decision space* $\mathscr{A}$ *equal to the state-space* $\Theta = \mathbb{R}$*. Let* $Y \sim N(\theta_0, 1)$ *for some unknown* $\theta_0 \in \Theta$*. Choose* $L : \Theta \times \mathscr{A} \to \mathbb{R}$ *equal to the quadratic difference*

$$L(\theta, a) = (\theta - a)^2,$$

a choice referred to as an L_2-loss (or squared-error loss*). Consider the decision-space*

$$\Delta = \{\delta_c : \mathscr{Y} \to \mathscr{A} \,:\, \delta_c(y) = c\,y,\ c \geq 0\}.$$

Note that Δ plays the role of a family of estimators for θ_0 here. The risk-function takes the form:

$$\begin{aligned} R(\theta, \delta_c) &= \int L(\theta, \delta_c(Y))\, dP_\theta = \int_{\mathbb{R}} (\theta - cy)^2 dN(\theta, 1)(y) \\ &= \int_{\mathbb{R}} \big(c(\theta - y) + (1-c)\theta\big)^2 dN(\theta, 1)(y) \\ &= \int_{\mathbb{R}} \Big(c^2(y-\theta)^2 + 2c(1-c)\theta(\theta - y) + (1-c)^2\theta^2\Big) dN(\theta, 1)(y) \\ &= c^2 + (1-c)^2\theta^2. \end{aligned}$$

It follows that δ_1 is R-better than all δ_c for $c > 1$, so that for all $c > 1$, δ_c is inadmissible. If we had restricted c to be greater than or equal to 1, δ_1 would have been admissible. However, since c may lie in $[0, 1)$ as well, admissibility in the uniform sense of (2.26) does not apply to any δ_c. To see this, note that $R(\theta, \delta_1) = 1$ for all θ, whereas for $c < 1$ and some $\theta > c/(1-c)$, $R(0, \delta_c) < 1 < R(\theta, \delta_c)$. Therefore, there is no admissible decision rule in Δ.

The minimax criterion does give rise to a preference. However, in order to guarantee its existence, we need to bound (or rather, compactify) the parameter space: let $M > 0$ be given and assume that $\Theta = [-M, M]$. The minimax risk for δ_c is given by

$$\sup_{\theta \in \Theta} R(\theta, \delta_c) = c^2 + (1-c)^2 M^2,$$

which is minimal iff $c = M^2/(1 + M^2)$, i.e. the (unique) minimax decision rule for this problem (or, since we are using decision theory to estimate a parameter in this case, the minimax estimator with respect to L_2-loss) is therefore,

$$\delta^M(Y) = \frac{M^2}{1 + M^2} Y.$$

Note that if we let $M \to \infty$, this estimator for θ converges to the MLE for said problem.

As demonstrated in the above example, uniform admissibility of a decision rule (*c.f.* (2.26)) is hard to achieve, but in many such cases a minimax decision rule does exist. One important remark concerning the use the minimax decision principle remains: considering (2.27), we see that the minimax principle chooses the decision rule that minimizes the *maximum* of the risk $R(\,\cdot\,, \delta)$ over Θ. As such, the minimax criterion takes into account *only* the worst-case scenario and prefers decision rules that perform well under those conditions. In practical problems, that means that the minimax principle tends to take a rather pessimistic perspective on decision problems.

Bayesian decision theory presents a more balanced perspective because instead of maximizing the risk function over Θ, the Bayesian has the prior to integrate over Θ. Optimization

of the resulting integral takes into account more than just the worst case, so that the resulting decision rule is based on a less pessimistic perspective than the minimax decision rule.

Definition 2.4.6. *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. In addition, assume that Θ is a measurable space with σ-algebra $\mathscr{G}$ and prior $\Pi : \mathscr{G} \to \mathbb{R}$. The function*

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta)\, d\Pi(\theta), \tag{2.28}$$

is called the Bayesian risk function. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule δ_1 is Bayes-preferred to δ_2, if

$$r(\Pi, \delta_1) < r(\Pi, \delta_2).$$

If $\delta^\Pi \in \Delta$ minimizes $\delta \mapsto r(\Pi, \delta)$, i.e.

$$r(\Pi, \delta^\Pi) = \inf_{\delta \in \Delta} r(\Pi, \delta). \tag{2.29}$$

then δ^Π is called a Bayes rule. The quantity $r(\Pi, \delta^\Pi)$ is called the Bayes risk.

Lemma 2.4.1. *Let $Y \in \mathscr{Y}$ denote data in a decision theoretic problem with state space Θ, decision space $\mathscr{A}$ and loss $L : \Theta \times \mathscr{A} \to \mathbb{R}$. For any prior Π and all decision rules $\delta : \mathscr{Y} \to \mathscr{A}$,*

$$r(\Pi, \delta) \leq \sup_{\theta \in \Theta} R(\theta, \delta),$$

i.e. the Bayesian risk is always upper bounded by the minimax risk.

The proof of this lemma follows from the fact that the minimax risk is an upper bound for the integrand in the Bayesian risk function.

Example 2.4.2. (continuation of example 2.4.1) *Let $\Theta = \mathbb{R}$ and $Y \sim N(\theta_0, 1)$ for some unknown $\theta_0 \in \Theta$. Choose the loss-function $L : \Theta \times \mathscr{A} \to \mathbb{R}$ and the decision space Δ as in example 2.4.1. We choose a prior $\Pi = N(0, \tau^2)$ (for some $\tau > 0$) on Θ. Then the Bayesian risk function is give by:*

$$\begin{aligned} r(\Pi, \delta_c) &= \int_\Theta R(\theta, \delta_c)\, d\Pi(\theta) = \int_{\mathbb{R}} \big(c^2 + (1-c)^2\theta^2\big) dN(0, \tau^2)(\theta) \\ &= c^2 + (1-c)^2\tau^2, \end{aligned}$$

which is minimal iff $c = \tau^2/(1+\tau^2)$. The (unique) Bayes rule for this problem and corresponding Bayes risk are therefore,

$$\delta^\Pi(Y) = \frac{\tau^2}{1+\tau^2} Y, \qquad r(\Pi, \delta^\Pi) = \frac{\tau^2}{1+\tau^2}.$$

In the Bayesian case, there is no need for a compact parameter space Θ, since we do not maximize but integrate over Θ.

In the above example, we could find the Bayes rule by straightforward optimization of the Bayesian risk function, because the class Δ was rather restricted. If we extend the class Δ to contain *all* non-randomized decision rules, the problem of finding the Bayes rule seems to be far more complicated at first glance. However, as we shall see in theorem 2.4.1, the following definition turns out to be the solution to this question.

Definition 2.4.7. (The conditional Bayes decision principle) *Let the state-space Θ, states P_θ, $(\theta \in \Theta)$, decision space $\mathscr{A}$ and loss L be given. In addition, assume that Θ is a measurable space with σ-algebra $\mathscr{G}$ and prior $\Pi : \mathscr{G} \to \mathbb{R}$. We define the decision rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ to be such that for all $y \in \mathscr{Y}$,*

$$\int_\Theta L(\theta, \delta^*(y))\, d\Pi(\theta|Y = y) = \inf_{a \in \mathscr{A}} \int_\Theta L(\theta, a)\, d\Pi(\theta|Y = y), \tag{2.30}$$

i.e. point-wise for every y, the decision rule $\delta^*(y)$ minimizes the posterior *expected loss.*

The above defines the decision rule δ^* implicitly as a point-wise minimizer, which raises the usual questions concerning existence and uniqueness, of which little can be said in any generality. However, if the existence of δ^* is established, it is optimal.

Theorem 2.4.1. *Let the state-space Θ, states P_θ, $(\theta \in \Theta)$, decision space $\mathscr{A}$ and loss L be given. In addition, assume that Θ is a measurable space with σ-algebra $\mathscr{G}$ and prior $\Pi : \mathscr{G} \to \mathbb{R}$. Assume that there exists a σ-finite measure $\mu : \mathscr{B} \to \mathbb{R}$ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$. If the decision rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ is well-defined, then δ^* is a Bayes rule.*

Proof Denote the class of all decision rules for this problem by Δ throughout the proof. We start by rewriting the Bayesian risk function for a decision rule $\delta : \mathscr{Y} \to \mathscr{A}$.

$$\begin{aligned} r(\Pi, \delta) &= \int_\Theta R(\theta, \delta)\, d\Pi(\theta) = \int_\Theta \int_{\mathscr{Y}} L(\theta, \delta(y))\, dP_\theta(y)\, d\Pi(\theta) \\ &= \int_{\mathscr{Y}} \int_\Theta L(\theta, \delta(y))\, p_\theta(y)\, d\Pi(\theta)\, d\mu(y) \\ &= \int_{\mathscr{Y}} \Big(\int_\Theta p_\theta(y)\, d\Pi(\theta) \Big) \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y)\, d\mu(y). \end{aligned}$$

where we use definitions (2.28) and (2.25), the Radon-Nikodym theorem (see theorem A.4.2), Fubini's theorem (see theorem A.4.1) and the definition of the posterior, *c.f.* (2.7). Using the prior predictive distribution (2.9), we rewrite the Bayesian risk function further:

$$r(\Pi, \delta) = \int_{\mathscr{Y}} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y)\, dP^\Pi(y). \tag{2.31}$$

By assumption, the conditional Bayes decision rule δ^* exists. Since δ^* satisfies (2.30) point-wise for all $y \in \mathscr{Y}$, we have

$$\int_\Theta L(\theta, \delta^*(y))\, d\Pi(\theta|Y = y) = \inf_{\delta \in \Delta} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y).$$

Substituting this in (2.31), we obtain

$$\begin{aligned} r(\Pi, \delta^*) &= \int_{\mathscr{Y}} \inf_{\delta \in \Delta} \int_{\Theta} L(\theta, \delta(y))\, d\Pi(\theta|Y=y)\, dP^{\Pi}(y) \\ &\leq \inf_{\delta \in \Delta} \int_{\mathscr{Y}} \int_{\Theta} L(\theta, \delta(y))\, d\Pi(\theta|Y=y)\, dP^{\Pi}(y) \\ &= \inf_{\delta \in \Delta} r(\Pi, \delta). \end{aligned}$$

which proves that δ^* is a Bayes rule. □

To conclude, it is noted that randomization of the decision is not needed when optimizing with respect to the Bayes risk. The conditional Bayes decision rule is non-randomized and optimal.

Example 2.4.3. (Classification and Bayesian classifiers) *Many decision-theoretic questions take the form of a classification problem: under consideration is a population Ω of objects that each belong to one of a finite number of classes $\mathscr{A} = \{1, 2, \ldots, L\}$. The class K of the object is the unknown quantity of interest. Observing a vector Y of features of the object, the goal is to* classify *the object, i.e. estimate which class it belongs to. We formalize the problem in decision-theoretic terms: the population is a probability space $(\Omega, \mathscr{F}, P)$; both the feature vector and the class of the object are random variables, $Y : \Omega \to \mathscr{Y}$ and $K : \Omega \to \mathscr{A}$ respectively. The state-space in a classification problem equals the decision space $\mathscr{A}$: the class can be viewed as a "state" in the sense that the distribution $P_{Y|K=k}$ of Y* given *the class $K = k$ depends on k. Based on the feature vector Y, we decide to classify in class $\delta(Y)$, i.e. the decision rule (or classifier, as it is usually referred to in the context of classification problems) maps features to classes by means of a map $\delta : \mathscr{Y} \to \mathscr{A}$. A classifier δ can be viewed equivalently as a finite partition of the feature-space $\mathscr{Y}$: for every $k \in \mathscr{A}$, we define*

$$\mathscr{Y}_k = \{y \in \mathscr{Y} : \delta(y) = k\}$$

and note that if $k \neq l$, then $\mathscr{Y}_k \cap \mathscr{Y}_l = \emptyset$ and $\mathscr{Y}_1 \cup \mathscr{Y}_2 \cup \ldots \cup \mathscr{Y}_L = \mathscr{Y}$. The partition of the feature space is such that if $Y = y \in \mathscr{Y}_k$ for certain $k \in \mathscr{A}$, then we classify the object in class k.

Depending on the context of the classification problem, a loss-function $L : \mathscr{A} \times \mathscr{A} \to \mathbb{R}$ is defined (see the examples in the introduction to this section, e.g. the example on medical diagnosis). Without context, the loss function in a classification problem can be chosen as follows

$$L(k, l) = 1_{\{k \neq l\}}.$$

i.e. we incur a loss equal to one for each misclassification.

Using the minimax decision principle, we look for a classifier $\delta^M : \mathscr{Y} \to \mathscr{A}$ that minimizes:

$$\delta \mapsto \sup_{k \in \mathscr{A}} \int_{\mathscr{Y}} L(k, \delta(y))\, dP(y|K=k) = \sup_{k \in \mathscr{A}} P\big(\delta(Y) \neq k \;\big|\; K = k\big),$$

i.e. the minimax decision principle prescribes that we minimize the probability of misclassification uniformly over all classes.

In a Bayesian context, we need a prior on the state-space, which equals $\mathscr{A}$ in classification problems. Note that if known (or estimable), the marginal probability distribution for K is to be used as the prior for the state k, in accordance with definition 2.1.1. In practical problems, frequencies of occurrence for the classes $\{1, \ldots, L\}$ in Ω are often available or easily estimable; in the absence of information on the marginal distribution of K equal prior weights can be assigned. Here, we assume that the probabilities $P(K = k)$ are known and use them to define the prior density with respect to the counting measure on the (finite) space $\mathscr{A}$:

$$\pi(k) = P(K = k).$$

The Bayes rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ for this classification problem is defined to as the minimizer of

$$\delta \mapsto \int_{\mathscr{A}} L(k, \delta(y))\, d\Pi(k|Y = y) = \sum_{k=1}^{L} \Pi\big(\delta(y) \neq K \;\big|\; Y = y\big)$$

for every $y \in \mathscr{Y}$. According to theorem 2.4.1, the classifier δ^* minimizes the Bayes risk, which in this situation is given by:

$$\begin{aligned} r(\Pi, \delta) &= \int_{\mathscr{A}} R(k, \delta)\, d\Pi(\theta) = \sum_{k \in \mathscr{A}} \int_{\mathscr{Y}} L(k, \delta(y))\, dP(y|K = k)\, \pi(k) \\ &= \sum_{k \in \mathscr{A}} P\big(k \neq \delta(Y) \;\big|\; K = k\big)\, P(K = k) = P\big(K \neq \delta(Y)\big). \end{aligned}$$

Summarizing, the Bayes rule δ^* minimizes the overall probability of misclassification, i.e. without referring to the class of the object. (Compare this with the minimax classifier.)

Readers interested in the statistics of classification and its applications are encouraged to read B. Ripley's "Pattern recognition and neural networks" (1996) [73].

To close the chapter, the following remark is in order: when we started our comparison of frequentist and Bayesian methods, we highlighted the conflict in philosophy. However, now that we have seen some of the differences in more detail by considering estimation, testing and decision theory in both schools, we can be far more specific. Statistical problems can be solved in both schools; whether one chooses for a Bayesian or frequentist solution is usually not determined by adamant belief in either philosophy, but by much more practical considerations. Perhaps example 2.4.3 illustrates this point most clearly: if one is concerned about correct classification for objects in the most difficult class, one should opt for the minimax decision rule. If, on the other hand, one wants to minimize the overall misclassification probability (disregarding misclassification per class), one should choose to adopt the conditional Bayes decision rule. In other words, depending on the risk to be minimized (minimax risk and Bayes risk are different!) one arrives at different classifiers. Some formulations are more natural in frequentist context and others belong in the Bayesian realm. Similarly, practicality may form an argument in favour of imposing a (possibly subjective) bias (see example 1.2.1). Bayesian

methods are a natural choice in such cases, due to the intrinsic bias priors express. For example, forensic statistics is usually performed using Bayesian methods, in order to leave room for common-sense bias. Another reason to use one or the other may be computational advantages or useful theoretical results that exist for one school but have no analog in the other.

Philosophical preference should not play a role in the choice for a statistical procedure, practicality should (and usually does).

2.5 Exercises

Exercise 2.1. CALIBRATION

A physicist prepares for repreated measurement of a physical quantity Z in his laboratory. To that end, he installs a measurement apparatus that will give him outcomes of the form $Y = Z + e$ where e is a measurement error due to the inaccuracy of the apparatus, assumed to be stochastically independent of Z. Note that if the expectation of e equals zero, long-run sample averages converge to the expectation of Z; if $Pe \neq 0$, on the other hand, averaging does not cancel out the resulting bias.

The manufacturer of the apparatus says that e is normally distributed with known variance $\sigma^2 > 0$. The mean θ of this normal distribution depends on the way the apparatus is installed and thus requires calibration. The following questions pertain to the calibration procedure.

The physicist decides to conduct the following steps to calibrate his measurement: if he makes certain that the apparatus receives no input signal, $Z = 0$. A sample of n independent measurements of Y then amounts to an i.i.d. sample from the distribution of e, which can be used to estimate the unknown mean θ. The physicist expects that Ee lies close to zero.

a. *Explain why, from a subjectivist point of view, the choice $\theta \sim N(0, \tau^2)$ forms a suitable prior in this situation. Explain the role of the parameter $\tau^2 > 0$.*

b. *With the choice of prior as in part* a.*, calculate the posterior density for θ.*

c. *Interpret the influence of τ^2 on the posterior, taking into account your answer under part* a. *(Hint: take limits $\tau^2 \downarrow 0$ and $\tau^2 \uparrow \infty$ in the expression you have found under* b.*)*

d. *What is the influence of the samplesize n? Show that the particular choice of the constant τ^2 becomes irrelevant in the large-sample limit $n \to \infty$.*

Exercise 2.2. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(0, \sigma^2)$, with unknown variance $\sigma^2 > 0$. As a prior for σ^2, let $1/\sigma^2 \sim \Gamma(1, 2)$. Calculate the posterior distribution for σ^2 with respect to the Lebesgue measure on $(0, \infty)$.*

Exercise 2.3. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution* Poisson(λ)*, with unknown parameter $\lambda > 0$. As a prior for λ, let $\lambda \sim \Gamma(2, 1)$. Calculate the posterior density for λ with respect to the Lebesgue measure on $(0, \infty)$.*

Exercise 2.4. *Let the measurement $Y \sim P_0$ be given. Assume that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated but possibly misspecified. Let Π denote a prior distribution on Θ. Show that the posterior distribution is P_0-almost-surely equal to the prior distribution iff the likelihood is $\Pi \times P_0$-almost-surely constant (as a function of $(\theta, y) \in \Theta \times \mathscr{Y}$). Explain the result of example 2.1.1 in this context.*

Exercise 2.5. *Consider the following questions in the context of exercise 2.3.*

a. *Calculate the maximum-likelihood estimator and the maximum-a-posteriori estimator for $\lambda \in (0, \infty)$.*

b. *Let $n \to \infty$ both in the MLE and MAP estimator and conclude that the difference vanishes in the limit.*

c. *Following remark 2.2.7, explain the difference between ML and MAP estimators exclusively in terms of the prior.*

d. *Consider and discuss the choice of prior $\lambda \sim \Gamma(2, 1)$ twice, once in a qualitative, subjectivist Bayesian fashion, and once following the frequentist interpretation of the log-prior-density.*

Exercise 2.6. *Let $Y \sim P_0$ denote the data. The following questions pertain to the small-ball estimator defined in remark 2.2.5 for certain, fixed $p \in (1/2, 1)$, which we shall denote by $\hat{P}(Y)$. Assume that the model $\mathscr{P}$ is compact in the topology induced by the metric d.*

a. *Show that for any two measurable model subsets $A, B \subset \mathscr{P}$,*

$$\big|\Pi(\, A \,|\, Y\,) - \Pi(\, B \,|\, Y\,)\big| \leq \Pi(\, A \cup B \,|\, Y\,) - \Pi(\, A \cap B \,|\, Y\,),$$

P_0-almost-surely.

b. *Prove that the map $(\epsilon, P) \mapsto \Pi(\, B_d(P, \epsilon) \,|\, Y\,)$ is continuous, P_0-almost-surely.*

c. *Show that $\hat{P}(Y)$ exists, P_0-almost-surely.*

d. *Suppose that $\epsilon > 0$ denotes the smallest radius for which there exists a ball $B_d(P, \epsilon) \subset \mathscr{P}$ of posterior probability greater than or equal to p. Show that, if both $\hat{P}_1(Y)$ and $\hat{P}_2(Y)$ are centre points of such balls, then $d(\hat{P}_1(Y), \hat{P}_2(Y)) < 2\epsilon$, P_0-almost-surely.*

Exercise 2.7. *Complete the proof of lemma 2.1.2. (Hint: Denote $S = \operatorname{supp}(\Pi)$; assume that $\Pi(S) = \pi < 1$; show that $\Pi(S^c \cap C) = 1 - \pi$ for any closed C such that $\Pi(C) = 1$; then use that intersections of closed sets are closed.*

Exercise 2.8. *Let Y be normally distributed with known variance $\sigma^2 > 0$ and unknown location θ. As a prior for θ, choose $\Pi = N(0, \tau^2)$. Let $\alpha \in (0, 1)$ be given. Using the posterior density with respect to the Lebesgue measure, express the level-α HPD-credible set in terms of Y, σ^2, τ^2 and quantiles of the standard normal distribution. Consider the limit $\tau^2 \to \infty$ and compare with level-α confidence intervals centred on the ML estimate for θ.*

Exercise 2.9. *Let $Y \sim \mathrm{Bin}(n;p)$ for known $n \geq 1$ and unknown $p \in (0,1)$. As a prior for p, choose $\Pi = \mathrm{Beta}(\frac{1}{2},\frac{1}{2})$. Calculate the posterior distribution for the parameter p. Using the Lebesgue measure on $(0,1)$ to define the posterior density, give the level-α HPD-credible interval for p in terms of Y, n and the quantiles of beta-distributions.*

Exercise 2.10. *Consider a dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for data Y, where $\Theta \subset \mathbb{R}$ is an interval. For certain $\theta_0 \in \Theta$, consider the simple null-hypothesis and alternative:*

$$H_0 : \quad \theta = \theta_0, \qquad H_1 \quad : \theta \neq \theta_0.$$

Show that if the prior Π is absolutely continuous with respect to the Lebesgue measure on Θ, then the Bayes factor B for the hypotheses H_0 versus H_1 satisfies $B = 0$.
Interpret this fact as follows: calculation of Bayes factors (and posterior/prior odds ratios) makes sense only *if both hypotheses receive non-zero prior mass. Otherwise, the statistical question we ask is rendered invalid* ex ante *by our beliefs concerning θ, as formulated through the choice of the prior.*

Exercise 2.11. PRISONER'S DILEMMA
Two men have been arrested on the suspicion of burglary and are held in separate cells awaiting interrogation. The prisoners have been told that burglary carries a maximum sentence of x years. However, if they confess, their prison terms are reduced to y years (where $0 < y < x$). If one of them confesses and the other does not, the first receives a sentence of y years while the other is sentenced to x years.

Guilty of the crime he is accused of, our prisoner contemplates whether to confess to receive a lower sentence, or to deny involvement in the hope of escaping justice altogether. He cannot confess without implicating the other prisoner. If he keeps his mouth shut and so does his partner in crime, they will both walk away free. If he keeps his mouth shut but his partner talks, he gets the maximum sentence. If he talks, he will always receive a sentence of y years and the other prisoner receives y or x years depending on whether he confessed or not himself. To talk or not to talk, that is the question.

There is no data in this problem, so we set θ equal to 1 or 0, depending on whether the other prisoner talks or not. Our prisoner can decide to talk ($t = 1$) or not ($t = 0$). The loss function $L(\theta, t)$ equals the prison term for our prisoner. In the absence of data, risk and loss are equal.

a. *Calculate the minimax risk for both $t = 0$ and $t = 1$. Argue that the minimax-optimal decision for our prisoner is to confess.*

As argued in section 2.4, the minimax decision can be overly pessimistic. In the above, it assumes that the other prisoner will talk and chooses t accordingly.

The Bayesian perspective balances matters depending on the chance that the other prisoner will confess when interrogated. This chance finds its way into the formalism as a prior for the trustworthiness of the other prisoner. Let $p \in [0,1]$ be the probability that the other prisoner confesses, i.e. $\Pi(\theta = 1) = p$ and $\Pi(\theta = 0) = 1 - p$.

b. *Calculate the Bayes risks for $t = 0$ and $t = 1$ in terms of x, y and p. Argue that the Bayes decision rule for our prisoner is as follows: if $y/x > p$ then our prisoner does not confess, if $y/x < p$, the prisoner confesses. If $y/x = p$, the Bayes decision criterion does not have a preference.*

So, depending on the degree to which our prisoner trusts his associate and the ratio of prison terms, the Bayesian draws his conclusion. The latter is certainly more sophisticated and perhaps more realistic, but it requires that our prisoner quantifies his trust in his partner in the form of a prior Bernoulli(p) distribution.

Chapter 3

Choice of the prior

Bayesian procedures have been the object of much criticism, often focusing on the choice of the prior as an undesirable source of ambiguity. The answer of the subjectivist that the prior represents the "belief" of the statistician or "expert knowledge" pertaining to the measurement elevates this ambiguity to a matter of principle, thus setting the stage for a heated debate between "pure" Bayesians and "pure" frequentists concerning the philosophical merits of either school within statistics. As said, the issue is complicated further by the fact that the Bayesian procedure does not refer to the "true" distribution P_0 for the observation (see section 2.1), providing another point of fundamental philosophical disagreement for the fanatically pure to lock horns over. Leaving the philosophical argumentation to others, we shall try to discuss the choice of a prior at a more conventional, practical level.

In this chapter, we look at the choice of the prior from various points of view: in section 3.1, we consider the priors that emphasize the subjectivist's prior "belief". In section 3.2 we construct priors with the express purpose *not* to emphasize any part of the model, as advocated by objectivist Bayesians. Because it is often desirable to control properties of the posterior distribution and be able to compare it to the prior, conjugate priors are considered in section 3.3. As will become clear in the course of the chapter, the choice of a "good" prior is also highly dependent on the model under consideration.

Since the Bayesian school has taken up an interest in non-parametric statistics only relatively recently, most (if not all) of the material presented in the first three sections of this chapter applies only to parametric models. To find a suitable prior for a non-parametric model can be surprisingly complicated. Not only does the formulation involve topological aspects that do not play a role in parametric models, but also the properties of the posterior may be surprisingly different from those encountered in parametric models! Priors on infinite-dimensional models are considered in section 3.4.

3.1 Subjective and objective priors

As was explained in chapters 1 and 2, all statistical procedures require the statistician to make certain choices, *e.g.* for model and method of inference. The subjectivist chooses the model as a collection of stochastic explanations of the data that he finds "reasonable", based on criteria no different from those frequentists and objectivist Bayesians would use.

Bayesians then proceed to choose a prior, preferably such that the support of this prior is not essentially smaller than the model itself. But even when the support of the prior is fixed, there is a large collection of possible priors left to be considered, each leading to a different posterior distribution. The objectivist Bayesian will choose from those possibilities a prior that is "homogeneous" (in a suitable sense), in the hope of achieving *unbiased* inference. The subjectivist, however, chooses his prior such as to emphasize parts of the model that he believes in stronger than others, thereby introducing a bias in his inferential procedure explicitly. Such a prior is called a subjective prior, or informative prior. The reason for this approach is best explained by examples like 1.2.1, which demonstrate that intuitive statistical reasoning is not free of bias either.

Subjectivity finds its mathematical expression when high prior "belief" is translated into "relatively large" amounts of assigned prior mass to certain regions of the model. However, there is no clear rule directing the exact fashion in which prior mass is to be distributed. From a mathematical perspective, this is a rather serious shortcoming, because it leaves us without a precise definition of the subjectivist approach. Often, the subjectivist will have a reasonably precise idea about his "beliefs" at the roughest level (*e.g.* concerning partitions of the model into a few subsets), but none at more detailed levels. When the parameter space Θ is unbounded this lack of detail becomes acute, given that the tail of the prior is hard to fix by subjective reasoning, yet highly influential for the inferential conclusions based on it. In practice, a subjectivist will often choose his prior without mathematical precision. He considers the problem, interprets the parameters in his model and chooses a prior to reflect all the (background) information at his disposition, ultimately filling in remaining details in an ad-hoc manner. It is worthwhile to mention that studies have been conducted focused on the ability of people to make a realistic guess at a probability distribution: they have shown that without specific training or practice, people tend to be overconfident in their assessment, assigning too much mass to possibilities they deem most likely and too little to others [1]. A tentative conclusion might be, that people tend to formulate their "beliefs" on a deterministic basis and deviate from that point of view only slightly (or, too little) when asked to give a realistic assessment of the probabilistic perspective. (For more concerning the intricacies of chosing subjective prior distributions, see Berger (1985) [8].)

Remark 3.1.1. *For this reason, it is imperative that a subjectivist prior is* always *reported alongside inferential conclusions based upon it! Reporting methods is important in any statistical setting, but if chosen methods lead to express bias, explanation is even more important. Indeed, not only the prior but also the reasoning leading to its choice should be reported, be-*

cause in a subjectivist setting, the motivation for the choice of a certain prior (and not any other) is part of the analysis *rather than an external consideration.*

If the model Θ is one-dimensional and the parameter θ has a clear interpretation, it is often not exceedingly difficult to find a reasonable prior Π expressing the subjectivist's "belief" concerning the value of θ.

Example 3.1.1. *If one measures the speed of light* in vacuo *c (a physical constant, approximately equal to* 299792458 *m/s), the experiment will be subject to random perturbations outside the control of the experimenter. For example, imperfection of the vacuum in the experimental equipment, small errors in timing devices, electronic noise and countless other factors may influence the resulting measured speed Y. We model the perturbations collectively as a normally distributed error $e \sim N(0, \sigma^2)$ where σ is known as a characteristic of the experimental setup. The measured speed is modelled as $Y = c + e$, i.e. the model $\mathscr{P} = \{N(c, \sigma^2) : c > 0\}$ is used to infer on c. Based on experiments in the past (most famous is the Michelson-Morley experiment (1887)), the experimenter knows that c has a value close to $3 \cdot 10^8$ m/s, so he chooses his prior to reflect this: a normal distribution located at* 300000000 *m/s with a standard deviation of (say)* 1000000 *m/s will do. The latter choice is arbitrary, just like the choice for a* normal *location model over other families.*

The situation changes when the parameter has a higher dimension, $\Theta \subset \mathbb{R}^d$: first of all, interpretability of each of the d components of $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ can be from straightforward, so that concepts like prior "belief" or "expert knowledge" become inadequate guidelines for the choice of a prior. Additionally, the choice for a prior in higher-dimensional models also involves choices concerning the dependence structure between parameters!

Remark 3.1.2. *Often, subjectivist inference employs exceedingly simple, parametric models for the sake of interpretability of the parameter (and to be able to choose a prior accordingly). Most frequentists would object to such choices for their obvious lack of realism, since they view the data as being generated by a "true, underlying distribution", usually assumed to be an element of the model. However, the subjectivist philosophy does not involve the ambition to be strictly realistic and calls for interpretability instead: to the subjectivist, inference is a personal rather than a universal matter. As such, the preference for simple parametric models is a matter of subjective interpretation rather than an assumption concerning reality or realistic distributions for the data.*

When confronted with the question which subjective prior to use on a higher-dimensional model, it is often of help to define the prior in several steps based on a choice for the dependence structure between various components of the parameter. Suppose that the subjectivist can imagine a reasonable distribution F for the first component θ_1, if he has definite values for all other components $\theta_2, \ldots, \theta_d$. This F is then none other than the (subjectivist prior)

distribution of θ_1, *given* $\theta_2, \ldots, \theta_d$,

$$F = \Pi_{\theta_1|\theta_2,\ldots,\theta_d}.$$

Suppose, furthermore, that a reasonable subjective prior G for the second component may be found, independent of θ_1, but given $\theta_3, \ldots, \theta_d$. Then,

$$G = \Pi_{\theta_2|\theta_3,\ldots,\theta_d}.$$

If we continue like this, eventually defining the marginal prior for the last component θ_d, we have found a prior for the full parameter θ, because for all $A_1, \ldots, A_d \in \mathscr{B}$,

$$\begin{aligned}\Pi(\theta_1 \in A_1, \ldots, \theta_d \in A_d) = \Pi(\theta_1 \in A_1 | \theta_2 \in A_2, \ldots, \theta_d \in A_d)\, \Pi(\theta_2 \in A_2 | \theta_3 \in A_3, \ldots, \theta_d \in A_d) \\ \times \ldots \times \Pi(\theta_{d-1} \in A_{d-1} | \theta_d \in A_d)\, \Pi(\theta_d \in A_d).\end{aligned}$$

Because prior beliefs may be more easily expressed when imagining a situation where other parameters have fixed values, one eventually succeeds in defining the prior for the high-dimensional model. The construction indicated here is that of a so-called hyperprior, which we shall revisit section 3.3. Note that when doing this, it is important to choose the parametrization of the model such that one may assume (with some plausibility), that θ_i is independent of $(\theta_1, \ldots, \theta_{i-1})$, *given* $(\theta_{i+1}, \ldots, \theta_d)$, for all $i \geq 1$.

In certain situations, the subjectivist has more factual information at his disposal when defining the prior for his analysis. In particular, if a probability distribution on the model reflecting the subjectivist's "beliefs" can be found by other statistical means, it can be used as a prior. Suppose the statistician is planning to measure a quantity Y and infer on a model $\mathscr{P}$; suppose also that this experiment repeats or extends an earlier analysis. From the earlier analysis, the statistician may have obtained a posterior distribution on $\mathscr{P}$. For the new experiment, this posterior may serve as a prior.

Example 3.1.2. *Let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ be a parametrized model for an i.i.d. sample $X_1, X_2, \ldots, X_n$ with prior measure $\Pi_1 : \mathscr{G} \to [0,1]$. Let the model be dominated (see definition 1.1.3), so that the posterior $\Pi_1(\,\cdot\,|X_1, \ldots, X_n)$ satisfies (2.8). Suppose that this experiment has been conducted, with the sample realised as $(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)$. Next, consider a new, independent experiment in which a quantity X_{n+1} is measured (with the same model). As a prior Π_2 for the new experiment, we use the (realised) posterior of the earlier experiment, i.e. for all $G \in \mathscr{G}$,*

$$\Pi_2(G) = \Pi_1(\,G\,|X_1 = x_1, \ldots, X_n = x_n).$$

The posterior for the second experiment then satisfies:

$$\begin{aligned} d\Pi_2(\theta|X_{n+1}) &= \frac{p_\theta(X_{n+1})\, d\Pi_2(\theta|X_1 = x_1, \ldots, X_n = x_n)}{\displaystyle\int_\Theta p_\theta(X_{n+1})\, d\Pi_2(\theta|X_1 = x_1, \ldots, X_n = x_n)} \\ &= \frac{p_\theta(X_{n+1}) \displaystyle\prod_{i=1}^{n} p_\theta(x_i) d\Pi_1(\theta)}{\displaystyle\int_\Theta p_\theta(X_{n+1}) \prod_{j=1}^{n} p_\theta(x_j)\, d\Pi_1(\theta)} \end{aligned} \tag{3.1}$$

The latter form is comparable to the posterior that would have been obtained if we had conducted a single experiment with an i.i.d. sample $X_1, X_2, \ldots, X_{n+1}$ of size $n+1$ and prior Π_1. In that case, the posterior would have been of the form:

$$\Pi(\,\cdot\,|X_1, \ldots, X_{n+1}) = \frac{\displaystyle\prod_{i=1}^{n+1} p_\theta(X_i)\, d\Pi_1(\theta)}{\displaystyle\int_\Theta \prod_{j=1}^{n+1} p_\theta(X_j)\, d\Pi_1(\theta)}, \tag{3.2}$$

i.e. the only difference is the fact that the posterior $\Pi_1(\,\cdot\,|X_1 = x_1, \ldots, X_n = x_n)$ is realised. As such, we may interpret independent consecutive experiments as a single, interrupted experiment and the posterior $\Pi_1(\,\cdot\,|X_1, \ldots, X_n)$ can be viewed as an intermediate result.

Remark 3.1.3. *Note that it is necessary to assume that the second experiment is stochastically independent of the first, in order to enable comparison between (3.1) and (3.2).*

Clearly, there are other ways to obtain a distribution on the model that can be used as an informative prior. One example is the distribution that is obtained when a previously obtained frequentist estimator $\hat{\theta}$ for θ is subject to a procedure called the *bootstrap*. Although the bootstrap gives rise to a distribution that is interpreted (in the frequentist sense) as the distribution of the estimator $\hat{\theta}$ rather than θ itself, a subjectivist may reason that the estimator provides him with the "expert knowledge" on θ that he needs to define a prior on Θ. (For more on bootstrap methods, see Efron and Tibshirani (1993) [32].)

3.2 Non-informative priors

Objectivist Bayesians argee with frequentists that the "beliefs" of the statistician analyzing a given measurement should play a minimal role in the methodology. Obviously, the model choice already introduces a bias, but rather than embrace this necessity and expand upon it like subjectivists do, they seek to keep the remainder of the procedure unbiased. In particular, they aim to use priors that do not introduce additional information (in the form of prior "belief") in the procedure. Subjectivists introduce their "belief" by concentrating prior mass in certain regions of the model; correspondingly, objectivists prefer priors that are "homogeneous" in an appropriate sense.

At first glance, one may be inclined to argue that a prior is objective (or non-informative) if it is uniform over the parameter space: if we are inferring on parameter $\theta \in [-1, 1]$ and we do not want to favour any part of the model over any other, we would choose a prior of the form, $(A \in \mathscr{B})$,

$$\Pi(A) = \tfrac{1}{2}\mu(A), \tag{3.3}$$

where μ denotes the Lebesgue measure on $[-1, 1]$. Attempts to minimize the amount of subjectivity introduced by the prior therefore focus on uniformity (argumentation that departs from the Shannon entropy in discrete probability spaces reaches the same conclusion (see, for example, Ghosh and Ramamoorthi (2003) [42], p. 47)). The original references on Bayesian methods (*e.g.* Bayes (1763) [4], Laplace (1774) [57]) use uniform priors as well. But there are several problems with this approach: first of all, one must wonder how to extend such reasoning when $\theta \in \mathbb{R}$ (or any other unbounded subset of $\mathbb{R}$). In that case, $\mu(\Theta) = \infty$ and we can not normalize Π to be a probability measure! Any attempt to extend Π to such unbounded models as a probability measure (or even as a finite measure) would eventually lead to inhomogeneity, *i.e.* go at the expense of the unbiasedness of the procedure.

The compromise some objectivists are willing to make, is to relinquish the interpretation that subjectivists give to the prior: they do not express any prior "degree of belief" in $A \in \mathscr{G}$ through the subjectivist statement that the (prior) probability of finding $\vartheta \in A$ equals $\Pi(A)$. Although they maintain the Bayesian interpretation of the posterior, they view the prior as a mathematical definition rather than a philosophical concept. Then, the following definition can be made without further reservations.

Definition 3.2.1. *Given a model $(\Theta, \mathscr{G})$, a prior measure $\Pi : \mathscr{G} \to \bar{\mathbb{R}}$ such that $\Pi(\Theta) = \infty$ is called an improper prior.*

Note that the normalization factor $\frac{1}{2}$ in (3.3) cancels in the expression for the posterior, *c.f.* (2.4): any finite multiple of a (finite) prior is equivalent to the original prior as far as the posterior is concerned. However, this argument does not extend to the improper case: integrability problems or other infinities may ruin the procedure, even to the point where the posterior measure becomes infinite or ill-defined. So not just the philosophical foundation of the Bayesian approach is lost, mathematical integrity of the procedure can no longer be guaranteed either! When confronted with an improper prior, the entire procedure must be checked for potential problems. In particular, one must verify that the posterior is a well-defined *probability* measure.

Remark 3.2.1. *Throughout these notes, whenever we refer to a prior measure, it is implied that this measure is a probability measure unless stated otherwise.*

But even if one is willing to accept that objectivity of the prior requires that we restrict attention to models on which "uniform" probability measures exist (*e.g.* with Θ a bounded subset of $\mathbb{R}^d$), a more fundamental problem exists: the very notion of uniformity is dependent on the parametrization of the model! To see this we look at a model that can be parametrized

in two ways and we consider the way in which uniformity as seen in one parametrization manifests itself in the other parametrization. Suppose that we have a d-dimensional parametric model $\mathscr{P}$ with two different parametrizations, on $\Theta_1 \subset \mathbb{R}^d$ and $\Theta_2 \subset \mathbb{R}^d$ respectively,

$$\phi_1 : \Theta_1 \to \mathscr{P}, \qquad \phi_2 : \Theta_2 \to \mathscr{P} \tag{3.4}$$

both of which are bijective. Assume that $\mathscr{P}$ has a topology and is endowed with the corresponding Borel σ-algebra $\mathscr{G}$; let ϕ_1 and ϕ_2 be continuous and assume that their inverses ϕ_1^{-1} and ϕ_2^{-1} are continuous as well. Assuming that Θ_1 is bounded, we consider the uniform prior Π_1 on Θ_1, *i.e.* the normalized Lebesgue measure on Θ_1, *i.e.* for all $A \in \mathscr{B}_1$,

$$\Pi_1(A) = \mu(\Theta_1)^{-1}\mu(A),$$

This induces a prior Π_1' on $\mathscr{P}$: for all $B \in \mathscr{G}$,

$$\Pi_1'(B) = (\Pi_1 \circ \phi_1^{-1})(B). \tag{3.5}$$

In turn, this induces a prior Π_1'' on Θ_2: for all $C \in \mathscr{B}_2$,

$$\Pi_1''(C) = (\Pi_1' \circ (\phi_2^{-1})^{-1})(C) = (\Pi_1' \circ \phi_2)(C) = \big(\Pi_1 \circ (\phi_1^{-1} \circ \phi_2)\big)(C).$$

Even though Π_1 is uniform, generically Π_1'' is *not*, because, effectively, we are mapping (a subset of) $\mathbb{R}^d$ to $\mathbb{R}^d$ by $\phi_2^{-1} \circ \phi_1 : \Theta_1 \to \Theta_2$. (Such re-coordinatizations are used extensively in differential geometry, where a manifold can be parametrized in various ways by sets of maps called *charts*.)

Example 3.2.1. *Consider the model $\mathscr{P}$ of all normal distributions centred on the origin with unknown variance between 0 and 1. We may parametrize this model in many different ways, but we consider only the following two:*

$$\phi_1 : (0,1) \to \mathscr{P} : \tau \mapsto N(0,\tau), \qquad \phi_2 : (0,1) \to \mathscr{P} : \sigma \mapsto N(0,\sigma^2). \tag{3.6}$$

Although used more commonly than ϕ_1, parametrization ϕ_2 is not special in any sense: both parametrizations describe exactly the same model. Now, suppose that we choose to endow the first parametrization with a uniform prior Π_1, equal to the Lebesgue measure μ on $(0,1)$. By (3.5), this induces a prior on $\mathscr{P}$. Let us now see what this prior looks like if we consider $\mathscr{P}$ parametrized by σ: for any constant $C \in (0,1)$ the point $N(0,C)$ in $\mathscr{P}$ is the image of $\tau = C$ and $\sigma = \sqrt{C}$, so the relation between τ and corresponding σ is given by

$$\tau(\sigma) = (\phi_2^{-1} \circ \phi_1)(\sigma) = \sigma^2.$$

Since Π_1 equals the Lebesgue measure, we find that the density of Π_1'' with respect to the Lebesgue measure equals:

$$\pi_1''(\sigma) = \pi_1(\tau(\sigma))\frac{d\tau}{d\sigma}(\sigma) = 2\sigma.$$

This density is non-constant and we see that Π_1'' is non-uniform. In a subjectivist sense, the prior Π_1'' places higher prior "belief" on values of σ close to 1 than on values close to 0.

From the above argument and example 3.2.1, we see that uniformity of the prior is entirely dependent on the parametrization: what we call "uniform" in one parametrization, may be highly non-uniform in another. Consequently, what is deemed "objective" in one parametrization may turn out to be highly subjective in another.

What matters is the model $\mathscr{P}$, not its parametrization in terms of one parameter or another! The parametrization is a mere choice made by the statistician analyzing the problem. Therefore, any statistical concept that depends on the parametrization is flawed from the outset. Through $\mathscr{P}$ and *only* through $\mathscr{P}$ do the parameters σ and τ have any bearing on (the law of) the observation in example 3.2.1. If we could define what is meant by uniformity on the model $\mathscr{P}$ itself, instead of on its parametrizing spaces, one would obtain a viable way to formalize objectivity. But spaces of probability measures do not have an intrinsic notion of uniformity (like translation-invariance of Lebesgue measure on $\mathbb{R}^d$, or more generally, left-invariance of the Haar measure on locally compact topological groups).

Once it is clear that uniformity on any parametrizing space does not have intrinsic meaning in the model $\mathscr{P}$, the very definition of objectivity in terms of uniformity of the prior is void. A subjectivist can use any parametrization to formulate his prejudice (note that the subjectivist uses *relative* prior weights rather than deviations from uniformity to express his prior "belief"), but an objectivist has to define his notion of "objectivity" regardless of the parametrization used. Therefore, the emphasis is shifted: instead of looking for uniform priors, we look for priors that are well-defined on $\mathscr{P}$ and declare them objective. For differentiable parametric models, a construction from Riemannian geometry can be used to define a parameterisation-independent prior (see Jeffreys (1946), (1961) [46, 47]) if we interpret the Fisher information as a Riemannian metric on the model (as first proposed by Rao (1945) [71] and extended by Efron (1975) [31]; for an overview, see Amari (1990) [2]) and use the square-root of its determinant as a density with respect to the Lebesgue measure.

Definition 3.2.2. *Let $\Theta \subset \mathbb{R}$ be open and let $\Theta \to \mathscr{P}$ define a differentiable, parametric, dominated model. Assume that for every $\theta \in \Theta$, the score-function $\dot{\ell}_\theta$ is twice integrable with respect to P_θ. Then Jeffreys prior Π has the square root of the determinant of the Fisher information $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ as its density with respect to the Lebesgue measure on Θ:*

$$d\Pi(\theta) = \sqrt{\det(I_\theta)}\, d\theta. \tag{3.7}$$

Although the expression for Jeffreys prior has the appearance of being parametrization-dependent, the form (3.7) of this prior is the *same* in *any* parametrization (a property referred to sometimes as (coordinate-)covariance). In other words, no matter which parametrization we use to calculate Π in (*c.f.* (3.7)), the induced measure Π' on $\mathscr{P}$ is always the same one. As such, Jeffreys prior is a measure defined on $\mathscr{P}$ rather than a parametrization-dependent measure.

Example 3.2.2. *We calculate the density of Jeffreys prior in the normal model of example 3.2.1. The score-function with respect to the parameter σ in parametrization ϕ_2 of $\mathscr{P}$ is*

given by:

$$\dot{\ell}_\sigma(X) = \frac{1}{\sigma}\Big(\frac{X^2}{\sigma^2} - 1\Big).$$

The Fisher information (which is a 1×1-maxtrix in this case), is then given by:

$$I_\sigma = P_\sigma \dot{\ell}_\sigma \dot{\ell}_\sigma = \frac{1}{\sigma^2} P_\sigma \Big(\frac{X^2}{\sigma^2} - 1\Big)^2 = \frac{2}{\sigma^2}$$

Therefore, the density for Jeffries prior Π takes the form

$$d\Pi(\sigma) = \frac{\sqrt{2}}{\sigma}\, d\sigma,$$

for all $\sigma \in \Theta_2 = (0, 1)$. A similar calculation using the parametrization ϕ_1 shows that, in terms of the parameter τ, Jeffries prior takes the form:

$$d\Pi(\tau) = \frac{1}{\sqrt{2}\tau}\, d\tau,$$

for all $\tau \in \Theta_1 = (0, 1)$. That both densities give rise to the same measure on $\mathscr{P}$ is the assertion of the following lemma.

Lemma 3.2.1. *(Parameterization-independence of Jeffreys prior)*
Consider the situation of (3.4) and assume that the parametrizations ϕ_1 and ϕ_2 satisfy the conditions of definition 3.2.2. In addition, we require that the map $\phi_1^{-1} \circ \phi_2 : \Theta_2 \to \Theta_1$ is differentiable. Then the densities (3.7), calculated in coordinates ϕ_1 and ϕ_2 induce the same measure on $\mathscr{P}$, Jeffreys prior.

Proof Since the Fisher information can be written as:

$$I_{\theta_1} = P_{\theta_1}(\dot{\ell}_{\theta_1} \dot{\ell}_{\theta_1}^T),$$

and the score $\dot{\ell}_{\theta_1}(X)$ is defined as the derivative of $\theta_1 \mapsto \log p_{\theta_1}(X)$ with respect to θ_1, a change of parametrization $\theta_1(\theta_2) = (\phi_1^{-1} \circ \phi_2)(\theta_2)$ induces a transformation of the form

$$I_{\theta_2} = S_{1,2}(\theta_2)\, I_{\theta_1(\theta_2)}\, S_{1,2}(\theta_2)^T,$$

on the Fisher information matrix, where $S_{1,2}(\theta_2)$ is the total derivative matrix of $\theta_2 \mapsto \theta_1(\theta_2)$ in the point θ_2 of the model. Therefore,

$$\begin{aligned}\sqrt{\det I_{\theta_2}}\, d\theta_2 &= \sqrt{\det(S_{1,2}(\theta_2)\, I_{\theta_1(\theta_2)}\, S_{1,2}(\theta_2)^T)}\, d\theta_2 = \sqrt{\det(S_{1,2}(\theta_2))^2}\sqrt{\det(I_{\theta_1(\theta_2)})}\, d\theta_2 \\ &= \sqrt{\det(I_{\theta_1(\theta_2)})}\, \big|\det(S_{1,2}(\theta_2))\big|\, d\theta_2 = \sqrt{\det(I_{\theta_1})}\, d\theta_1\end{aligned}$$

i.e. the form of the density is such that reparametrization leads exactly to the Jacobian for the transformation of $d\theta_2$ to $d\theta_1$. □

Ultimately, the above construction derives from the fact that the Fisher information I_θ (or in fact, any other positive-definite symmetric matrix-valued function on the model, *e.g.* the Hessian of a twice-differentiable, convex function) can be viewed as a Riemann metric on the "manifold" $\mathscr{P}$. The construction of a measure with Lebesgue density (3.7) is then a standard construction in differential geometry.

Example 3.2.3. *To continue with the normal model of examples 3.2.1 and 3.2.2, we note that* $\sigma(\tau) = \sqrt{\tau}$*, so that* $d\sigma/d\tau(\tau) = 1/2\sqrt{\tau}$*. As a result,*

$$\sqrt{\det I_{\theta_2}}\, d\theta_2 = \frac{\sqrt{2}}{\sigma}\, d\sigma = \frac{\sqrt{2}}{\sigma(\tau)} \frac{d\sigma}{d\tau}(\tau)\, d\tau = \frac{\sqrt{2}}{\sqrt{\tau}} \frac{1}{2\sqrt{\tau}}\, d\tau = \frac{1}{\sqrt{2}\tau}\, d\tau = \sqrt{\det(I_{\theta_1})}\, d\theta_1,$$

which verifies the assertion of lemma 3.2.1 explicitly.

Other constructions and criteria for the construction of non-informative priors exist: currently very popular is the use of so-called reference priors, as introduced in Lindley (1956) [65] and rediscovered in Bernardo (1979) [12] (see also Berger and Bernardo (1992) [9]). By defining principle, a reference prior is required to maximize the Kullback-Leibler divergence between prior and posterior. To motivate this condition, we have to look at information theory, from which the Kullback-Leibler divergence has emerged as one (popular but by no means unique) way to quantify the notion of the "amount of information" contained in a probability distribution. Sometimes called the Shannon entropy, the Kullback-Leibler divergence of a distribution P with respect to the counting measure in discrete probability spaces,

$$S(P) = \sum_{\omega \in \Omega} p(\omega)\, \log(p(\omega)),$$

can be presented as such convincingly (see Bolzmann (1895, 1898) [22], Shannon (1948) [78]). For lack of a default dominating measure, the argument does not extend formally to continuous probability spaces but is generalized nevertheless. A reference prior Π on a dominated, parametrized model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for an observation Y is to be chosen such that the Lindley entropy,

$$S_L = \int \int \log\Big(\frac{\pi(\theta|Y=y)}{\pi(\theta)}\Big) d\Pi(\theta|Y=y)\, dP^{\Pi}(y),$$

is maximized. Note that this definition does not depend on the specific parametrization, since the defining property is parametrization independent. Usually, the derivation of a reference prior [12] is performed in the limit where the posterior becomes asymptotically normal, *c.f.* theorem 4.4.1. Jeffreys prior emerges as a special case of a reference prior.

For an overview of various objective methods of constructing priors, the reader is referred to Kass and Wasserman (1995) [49]. When using non-informative priors, however, the following general warning should be heeded

Remark 3.2.2. *In many models, non-informative priors, including Jeffries prior and reference priors, are improper.*

3.3 Conjugate families, hierarchical and empirical Bayes

Consider again the problem of estimating the mean of a single, normally distributed observation Y with known variance. The model consists of all normal distributions $P_\theta = N(\theta, \sigma^2)$, where $\theta \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is known. Imposing a normal prior on the parameter θ,

$\Pi = N(0, \tau^2)$, for some choice of $\tau^2 > 0$, we easily calculate that posterior distribution is a normal distribution,

$$\Pi(\theta \in A|Y) = N\left(\frac{\tau^2}{\sigma^2+\tau^2}Y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)(A),$$

for every $A \in \mathscr{B}$. The posterior mean, a point-estimator for θ, is then given by,

$$\hat{\theta}(Y) = \frac{\tau^2}{\sigma^2+\tau^2}Y.$$

The frequentist's critisism of Bayesian statistics focusses on the parameter τ^2: the choice that a subjectivist makes for τ^2 may be motivated by expert knownledge or belief, but remains the statistician's personal touch in a context where the frequentist would prefer an answer of a more universal nature. As long as some form of expert knowledge is available, the subjectivist's argument constitutes a tenable point of view (or may even be compelling, see examples 1.2.1 and 2.1.2). However, in situations where no prior belief or information on the parameter θ is available, or if the parameter itself does not have a clear interpretation, the subjectivist has no answer. Yet a choice for τ^2 is required! Enter the objectivist's approach: if we have no prior information on θ, why not express our prior ignorance by choosing a "uniform" prior for θ? As we have seen in section 3.2, uniformity is parametrization dependent (and, as such, still dependent on the statistician's personal choice for one parametrization and not another). Moreover, uniform priors are improper if Θ is unbounded in $\mathbb{R}^k$. In the above example of estimation of a normal mean, where $\theta \in \mathbb{R}$ is unbounded, insistance on uniformity leads to an improper prior as well. Perhaps more true to the original interpretation of the prior, we might express ignorance about τ^2 (and eliminate τ^2 from the point-estimator $\hat{\theta}(Y)$) by considering more and more homogeneous (but still normal) priors by means of the limit $\tau \to \infty$, in which case we recover the maximum-likelihood estimate: $\lim_{\tau^2\to\infty} \hat{\theta}(Y) = Y$.

Remark 3.3.1. *From a statistical perspective, however, there exists a better answer to the question regarding τ^2: if τ is not known, why not estimate its value from the data!*

In this section, we consider this solution both from the Bayesian and from the frequentist's perspective, giving rise to procedures known as hierarchical Bayesian modelling and empirical Bayesian estimation respectively.

Beforehand, we consider another type of choice of prior, which is motivated primarily by mathematical convenience. Taking another look at the normal example with which we began this section, we note that both the prior and the posterior are normal distributions. Since the calculation of the posterior is tractable, any choice for the location and variance of the normal prior can immediately be updated to values for location and variance of the normal posterior upon observation of $Y = y$. Not only does this signify ease of manipulation in calculations with the posterior, it also reduces the computational burden dramatically since simulation of (or, sampling from) the posterior is no longer necessary.

Definition 3.3.1. *Let $(\mathscr{P}, \mathscr{A})$ be a measurable model for an observation $Y \in \mathscr{Y}$. Let M denote a collection of probability distributions on $(\mathscr{P}, \mathscr{A})$. The set M is called a conjugate family for the model $\mathscr{P}$, if the posterior based on a prior from M again lies in M:*

$$\Pi \in M \quad \Rightarrow \quad \Pi(\,\cdot\,|Y = y) \in M, \tag{3.8}$$

for all $y \in \mathscr{Y}$.

This structure was first proposed by Raiffa and Schlaifer (1961) [70]. Their method for the prior choice is usually classified as objectivist because it does not rely on subjectivist notions and is motivated without reference to outside factors.

Remark 3.3.2. *Often in the literature, a prior is refered to as a conjugate prior if the posterior is of the same form. This practice is somewhat misleading, since it is the family M that is closed under conditioning on the data Y, a property that depends on the model and M, but* not *on the particular $\Pi \in M$.*

Example 3.3.1. *Consider an experiment in which we observe n independent Bernoulli trials and consider the total number of successes, $Y \sim \mathrm{Bin}(n, p)$ with unknown parameter $p \in [0, 1]$,*

$$P_p(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

For the parameter p we choose a prior $p \sim Beta(\alpha, \beta)$ from the Beta-family, for some $\alpha, \beta > 0$,

$$d\Pi(p) = B(\alpha, \beta)\, p^{\alpha-1}(1-p)^{\beta-1}\, dp,$$

where $B(\alpha, \beta) = \Gamma(\alpha+\beta)/(\Gamma(\alpha)\,\Gamma(\beta))$ normalizes Π. Then the posterior density with respect to the Lebesgue measure on $[0, 1]$ is proportional to:

$$d\Pi(p|Y) \propto p^Y (1-p)^{n-Y}\, p^{\alpha-1}(1-p)^{\beta-1}\, dp = p^{\alpha+Y-1}(1-p)^{\beta+n-Y-1}\, dp,$$

We conclude that the posterior again lies in the Beta-family, with parameters equal to a data-amended version of those of the prior, as follows:

$$\Pi(\,\cdot\,|Y) = \mathrm{Beta}(\alpha + Y, \beta + n - Y).$$

So the family of Beta-distributions is a conjugate family for the binomial model. Depending on the available amount of prior information on θ, the prior's parameters may be chosen on subjective grounds (see figure 2.1 for graphs of the densities of Beta-distributions for various parameter values). However, in the absence thereof, the parameters α, β suffer from the same ambiguity that plagues the parameter τ^2 featuring in the example with which we opened this section.

Example 3.3.1 indicates a strategy to find conjugate families for a given parametrized, dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$. We view densities $y \mapsto p_\theta(y)$ as functions of the outcome

$Y = y$ foremost, but they are functions of the parameter θ as well and their dependence $\theta \mapsto p_\theta(y)$ determines which prior densities $\theta \mapsto \pi(\theta)$ preserve their functional form when multiplied by the likelihood $p_\theta(Y)$ to yield the posterior density.

Although we shall encounter an example of a conjugate family for a non-parametric model in the next section, conjugate families are, by and large, part of parametric statistics. Many of these families are so-called exponential families, for which conjugate families of priors can be found readily.

Definition 3.3.2. *A dominated collection of probability measures $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is called a k-parameter exponential family, if there exists a $k \geq 1$ such that for all $\theta \in \Theta$,*

$$p_\theta(x) = \exp\Big(\sum_{i=1}^{k} \eta_i(\theta)\, T_i(x) - B(\theta)\Big) h(x), \tag{3.9}$$

where h and T_i, $i = 1, \ldots, k$, are statistics and B, η_i, $i = 1, \ldots, k$ are real-valued functions on Θ.

Any exponential family can be parametrized such that the exponent in (3.9) is linear in the parameter. By the mapping $\Theta \to H : \eta_i = \eta_i(\theta)$ (a bijection if the original parametrization is identifiable), taking Θ into $H = \eta(\Theta)$ and B into $A(\eta) = B(\theta(\eta))$, any exponential family can be rewritten in its so-called canonical form.

Definition 3.3.3. *An exponential family $\mathscr{P} = \{P_\eta : \eta \in H\}$, $H \subset \mathbb{R}^k$ is said to be in its canonical representation, if*

$$p_\eta(x) = \exp\Big(\sum_{i=1}^{k} \eta_i\, T_i(x) - A(\eta)\Big)\, h(x). \tag{3.10}$$

In addition, $\mathscr{P}$ is said to be of full rank if the interior of $H \subset \mathbb{R}^k$ is non-void, i.e. $\mathring{H} \neq \emptyset$.

Although parametric, exponential families are both versatile modelling tools and mathematically tractable; many common models, like the Bernoulli-, normal-, binomial-, Gamma-, Poisson-models, *etcetera*, can be rewritten in the form (3.9). One class of models that can immediately be disqualified as possible exponential families is that of all models in which the support depends on the parameter, like the family of all uniform distributions on $\mathbb{R}$, or the Pareto-model. Their statistical practicality stems primarily from the fact that for an exponential family of full rank, the statistics T_i, $i = 1, \ldots, k$ are sufficient and complete, enabling the use of the Lehmann-Scheffé theorem for minimal-variance unbiased estimation (see, for instance, Lehmann and Casella (1998) [59]). Their versatility can be understood in many ways, *e.g.* by the Pitman-Koopman-Darmois theorem (see, Jeffreys (1961) [47]), which says that a family of distributions whose support does not depend on the parameter, is exponential, if and only if in the models describing its *i.i.d.* samples, there exist sufficient statistics whose dimension remains bounded asymptotically (*i.e.* as we let the sample size diverge to infinity).

Presently, however, our interest lies in the following theorem, which says that if a model $\mathscr{P}$ constitutes an exponential family, there exists a conjugate family of priors for $\mathscr{P}$.

Theorem 3.3.1. *Let $\mathscr{P}$ be a model that can be written as an exponential family, c.f. definition 3.3.2. Then there exists a parametrization of $\mathscr{P}$ of the form (3.10) and the family of distributions $\Pi_{\mu,\lambda}$, defined by Lebesgue probability densities*

$$\pi_{\mu,\lambda}(\eta) = K(\mu,\lambda)\, \exp\Big(\sum_{i=1}^{k} \eta_i \mu_i - \lambda\, A(\eta)\Big), \tag{3.11}$$

(where $\mu \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}$ are such that $0 < K(\mu,\lambda) < \infty$), is a conjugate family for $\mathscr{P}$.

Proof It follows from the argument preceding definition 3.3.3 that $\mathscr{P}$ can be parametrized as in (3.10). Choosing a prior on H of the form (3.11), we find that the posterior again takes the form (3.11),

$$\pi(\eta|X) \propto \exp\Big(\sum_{i=1}^{k} \eta_i(\mu_i + T_i(X)) - (\lambda+1)\, A(\eta)\Big)$$

(the factor $h(X)$ arises both in numerator and denominator of (2.4) and is η-independent, so that it cancels). The data-amended versions of the parameters μ and λ that emerge from the posterior are therefore given by:

$$(\mu + T(X), \lambda + 1),$$

and we conclude that the distributions $\Pi_{\mu,\lambda}$ form a conjugate family for $\mathscr{P}$. □

Remark 3.3.3. *From a frequentist perspective, it is worth noting the import of the factorization theorem, which says that the parameter-dependent factor in the likelihood is a function of the data only through the sufficient statistic. Since the posterior is a function of the likelihood, in which data-dependent factors that do not depend on the parameter can be cancelled between numerator and denominator, the posterior is a function of the data X only through the sufficient statistic $T(X)$. Therefore, if the exponential family $\mathscr{P}$ is of full rank (so that $T(X)$ is also complete for $\mathscr{P}$), any point-estimator we derive from this posterior (e.g. the posterior mean, see definition 2.2.1) that is unbiased and quadratically integrable, is optimal in the sense of Rao-Blackwell, c.f. the theorem of Lehmann-Scheffé (see Lehmann and Casella (1998) [59], for explanation of the Rao-Blackwell and Lehmann-Scheffé theorems).*

Next, we turn to the Bayesian answer to remark 3.3.2 which said that parameters of the prior (*e.g.* τ^2) are to be estimated themselves. Recall that the Bayesian views a parameter to be estimated as just another random variable in the probability model. In case we want to estimate the parameter for a family of priors, then that parameter is to be included in the probability space from the start. Going back to the example with which we started this section, this means that we still use normal distributions $P_\theta = N(\theta, \sigma^2)$ to model the uncertainty in the data Y, supply $\theta \in \mathbb{R}$ with a prior $\Pi_1 = N(0, \tau^2)$ and then proceed to choose a another prior Π_2 for $\tau^2 \in (0, \infty)$:

$$Y|\theta,\tau^2 = Y|\theta \sim P_\theta = N(\theta,\sigma^2), \quad \theta|\tau^2 \sim \Pi_1 = N(0,\tau^2), \quad \tau^2 \sim \Pi_2,$$

Note that the parameter τ^2 has no direct bearing on the model distributions: conditional on θ, Y is independent of τ^2. In a sense, the hierarchical Bayesian approach to prior choice combines subjective and objective philosophies: whereas the subjectivist will make a definite, informed choice for τ^2 and the objectivist will keep himself as uncommitted as possible by striving for uniformity, the choice for a hierarchical prior expresses uncertainty about the value of τ^2 to be used in the form of a probability distribution Π_2. As such, the hierarchical Bayesian approach allows for intermediate prior choices: if Π_2 is chosen highly concentrated around one point in the model, resembling a degenerate measure, the procedure will be close to subjective; if Π_2 is spread widely and is far from degenerate, the procedure will be less biased and closer to objective. Besides interpolating between objective and subjective prior choices, the flexibility gained through introduction of Π_2 offers a much wider freedom of modelling. In particular, we may add several levels of modelled parameter uncertainty to build up a hierarchy of priors for parameters of priors. Such structures are used to express detailed subjectivist beliefs, much in the way graphical models are used to build intricate dependency structures for observed data (for a recent text on graphical models, see chapter 8 of Bishop (2006) [20]). The origins of the hierarchical approach go back, at least, to Lindley and Smith (1972) [66].

Definition 3.3.4. *Let the data Y be random in $(\mathscr{Y}, \mathscr{B})$. A hierarchical Bayesian model for Y consists of a collection of probability measures $\mathscr{P} = \{P_\theta : \theta \in \Theta_0\}$, with $(\Theta_0, \mathscr{G}_0)$ measurable and endowed with a prior $\Pi : \mathscr{G}_0 \to [0,1]$ built up in the following way: for some $k \geq 1$, we introduce measurable spaces $(\Theta_i, \mathscr{G}_i)$, $i = 1, 2, \ldots, k$ and conditional priors*

$$\mathscr{G}_i \times \Theta_{i+1} \to [0,1] : (G, \theta_{i+1}) \mapsto \Pi_i(G|\theta_{i+1}),$$

for $i = 1, \ldots, k-1$ and a marginal $\Pi_k : \mathscr{G}_k \to [0,1]$ on Θ_k. The prior for the original parameter θ is then defined by,

$$\Pi(\theta \in G) = \int_{\Theta_1 \times \ldots \times \Theta_k} \Pi_0(\theta \in G|\theta_1)\, d\Pi(\theta_1|\theta_2) \ldots d\Pi(\theta_{k-1}|\theta_k)\, d\Pi_k(\theta_k), \tag{3.12}$$

for all $G \in \mathscr{G}_0$. The parameters $\theta_1, \ldots \theta_k$ and the priors $\Pi_1, \ldots, \Pi_2$ are called hyperparameters and their hyperpriors.

This definition elicits several remarks immediately.

Remark 3.3.4. *Definition 3.3.4 of a hierarchical Bayesian model does* not *constitute a generalization of the Bayesian procedure in any formal sense: after specification of the hyperpriors, one may proceed to calculate the prior Π, c.f. (3.12), and use it to infer on θ in the ways indicated in chapter 2 without ever having to revisit the hierachical background of Π. As such, the significance of the definition lies entirely in its conceptual, subjective interpretation.*

Remark 3.3.5. *Definition 3.3.4 is very close to the general Bayesian model that incorporates all parameters $(\theta, \theta_1, \ldots, \theta_k)$ as modelling parameters. What distinguishes hierarchical modelling from the general situation is the dependence structure imposed on the parameters. The*

parameter θ is distinct from the hyperparameters by the fact that conditional on θ, the data Y is independent of $\theta_1, \ldots, \theta_k$. This distinction is repreated at higher levels in the hierarchy, i.e. levels are separate from one another through the conditional independence of $\theta_i|\theta_{i+1}$ from $\theta_{i+2}, \ldots, , \theta_k$.

Remark 3.3.6. *The hierarchy indicated in definition 3.3.4 inherently loses interpretability as we ascend in level. One may be able to give a viable interpretation to the parameter θ and to the hyperparameter θ_1, but higher-level parameters $\theta_2, \theta_3, \ldots$ become harder and harder to understand heuristically. Since the interpretation of the hierarchy requires a subjective motivation of the hyperpriors, interpretability of each level is imperative, or left as a non-informative choice. In practice, Bayesian hierarchical models are rarely more than two levels deep ($k = 2$) and the last hyperprior Π_k is often chosen by objective criteria.*

Example 3.3.2. *We observe the number of surviving offspring from a bird's litter and aim to estimate the number of eggs the bird laid: the bird lays $N \geq 0$ eggs, distributed according to a Poisson distribution with parameter $\lambda > 0$. For the particular species of bird in question, the Poisson rate λ is not known exactly: the uncertainty in λ can be modelled in many ways; here we choose to model it by a Gamma-distribution $\Gamma(\alpha, \beta)$, where α and β are chosen to reflect our imprecise knowledge of λ as well as possible. Each of the eggs then comes out, producing a viable chick with known probability $p \in [0, 1]$, independently. Hence, the total number Y of surviving chicks from the litter is distributed according to a binomial distribution, conditional on N,*

$$Y|N \sim \mathrm{Bin}(N, p), \quad N|\lambda \sim \mathrm{Poisson}(\lambda), \quad \lambda \sim \Gamma(\alpha, \beta).$$

The posterior distribution is now obtained as follows: conditional on $N = n$, the probability of finding $Y = k$ is binomial,

$$P(Y = k|N = n) = \binom{n}{k} p^k (1-p)^{n-k},$$

so Bayes' rule tells us that the posterior is given by:

$$P(N = n|Y = k) = \frac{P(N = n)}{P(Y = k)} \binom{n}{k} p^k (1-p)^{n-k}.$$

Since $\sum_{n \geq 0} P(N = n|Y = k) = 1$ for every k, the marginal $P(Y = k)$ (viz. the denominator or normalization factor for the posterior given $Y = k$) can be read off once we have the expression for the numerator. We therefore concentrate on the marginal for $N = n$, ($n \geq 0$):

$$P(N = n) = \int_{\mathbb{R}} P(N = n|\lambda)\, p_{\alpha,\beta}(\lambda)\, d\lambda = \frac{1}{\Gamma(\alpha)\, \beta^\alpha} \int_0^\infty \frac{e^{-\lambda}\, \lambda^n}{n!} \lambda^{\alpha-1}\, e^{-\lambda/\beta}\, d\lambda.$$

The integral is solved using the normalization constant of the $\Gamma((\alpha+n), (\beta/\beta+1))$-distribution:

$$\int_0^\infty e^{-\lambda \frac{\beta+1}{\beta}} \lambda^{\alpha+n-1}\, d\lambda = \Gamma(\alpha + n) \Big(\frac{\beta}{\beta+1} \Big)^{\alpha+n}.$$

Substituting and using the identity $\Gamma(\alpha+1)=\alpha\,\Gamma(\alpha)$*, we find:*

$$P(N=n)=\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\frac{1}{n!}\frac{1}{\beta^{\alpha}}\Big(\frac{\beta}{\beta+1}\Big)^{\alpha+n}=\frac{1}{n!}\Big(\frac{\beta}{\beta+1}\Big)^{n}\frac{1}{(\beta+1)^{\alpha}}\prod_{l=1}^{n}(\alpha+l-1) \tag{3.13}$$

Although not in keeping with the subjective argumentation we insist on in the introduction to this example, for simplicity we consider $\alpha=\beta=1$ *and find that in that case,*

$$P(N=n)=(1/2)^n.$$

The posterior for $N=n$ *given* $Y=k$ *then takes the form:*

$$P(N=n|Y=k)=\frac{1}{2^n}\binom{n}{k}p^k(1-p)^{n-k}\Big/\sum_{m\geq 0}\frac{1}{2^m}\binom{m}{k}p^k(1-p)^{m-k}.$$

The eventual form of the posterior illustrates remark 3.3.4: in case we choose $\alpha=\beta=1$*, the posterior we find from the hierarchical Bayesian model does not differ from the posterior that we would have found if we had have started from the non-hierarchical model with a geometric prior,*

$$Y|N\sim\text{Bin}(N,p),\quad N\sim\text{Geo}(1/2).$$

Indeed, even if we leave α *and* β *free, the marginal distribution for* N *we found in (3.13) is none other than the prior (3.12) for this problem.*

The conclusion one should draw from remark 3.3.4 and example 3.3.2, is that the hierarchical Bayesian approach adds nothing new to the *formal* Bayesian procedure: eventually, it amounts a choice for the prior just like in chapter 2. However, in a subjectivist sense, the hierarchical approach allows for greater freedom and a more solid foundation to *motivate* the choice for certain prior over other possibilities. This point is all the more significant in light of remark 3.1.1: the motivation of a subjectivist choice for the prior is part of the statistical analysis rather than an external aspect of the procedure. Hierarchical Bayesian modelling helps to refine and justify motivations for subjectivist priors.

But the subjectivist answer is not the only one relevant to the statistical perspective of remark 3.3.2 on the initial question of this section. The objectivist Bayesian may argue that any hyperprior should be chosen in a non-informative fashion, either as a matter of principle, or to reflect lack of interpretability or prior information on the parameter τ^2. Such a strategy amounts to the hierarchical Bayesian approach with one or more levels of objective hyperpriors, a point of view that retains only the modelling freedom gained through the hierarchical approach.

More unexpected is the frequentist perspective on remark 3.3.2: if τ^2 is an unknown, point-estimate it first and then perform the Bayesian analysis with this point-estimate as a "plug-in" for the unknown τ^2. Critical notes can be placed with the philosophical foundations for this practice, since it appears to combine the methods of two contradictory schools of statistics. Be that as it may, the method is used routinely based on its practicality: eventually, the

justification comes from the subjectivist who does not reject frequentist methods to obtain expert knowledge on his parameters, as required in his own paradigm.

Remark 3.3.7. *Good statistical practice dictates that one may not "peek" at the data to decide which statistical method to use for the analysis of the same data. The rationale behind this dictum is that pre-knowledge of the data could bias the analysis. If we take this point strictly, the choice for a prior (read, the point-estimate for τ^2) should not be made on the basis of the same data Y that is to be used later to derive the posterior for θ. If one has two independent realisations of the data, one can be used to choose the prior, (here, by a point-estimate for τ^2) and the other to condition on, in the posterior.*

Yet the above "rule" cannot be taken too strictly. Any statistician (and common sense) will tell you that it is crucial for the statistical analysis that one first obtains a certain feeling for the statistical problem by inspection of the data, before making decisions on how to analyse it (to see this point driven to the extreme, read, e.g. Tukey (1977) [82]). Ideally, one would make those decisions based on a sample of the data that is independent of the data used in the analysis proper. This precaution is often omitted, however: for example, it is common practice to use "plug-in" parameters based on the sample Y whenever the need arises, possibly leading to a bias in the subsequent analysis of the same data Y (unless the "plug-in" estimator is independent of all other estimators used, of course).

There are many different ways in which the idea of a prior chosen by frequentist methods is applied, all of which go under the name *empirical Bayes*. Following Berger [8], we note two types of statistical questions that are especially well suited for application. When we analyse data pertaining to an individual from a larger population and it is reasonable to assume that the prior can be inferred from the population, then one may estimate parameters like τ^2 above from population data and use the estimates in the prior for the individual.

Another situation where empirical Bayes is often used, is in *model selection*: suppose that there are several models $\mathscr{P}_1, \mathscr{P}_2, \ldots$ with priors $\Pi_1, \Pi_2, \ldots$, each of which may serve as a reasonable explanation of the data, depending on an unknown parameter $K \in \{1, 2, \ldots\}$. The choice to use model-prior pair k in the determination of the posterior can only be made after observation (or estimation) of K. If K is estimated by freqentist methods, the resulting procedure belongs to the realm of the empirical Bayes methods.

Example 3.3.3. *Consider the situation where we are provided with a specimen from a population that is divided into an unknown number of classes. Assume that all we know about the classes is that they occur with equal probabilities in the population. The particular class of our specimen remains unobserved. We perform a real-valued measurement Y on the specimen, which is normally distributed with known variance σ^2 and an unknown mean $\mu_k \in \mathbb{R}$ that depends on the class k. Then Y is distributed according to a discrete mixture of normal distributions of the form*

$$Y \sim P_{K;\mu_1,\ldots,\mu_K} = \frac{1}{K} \sum_{k=1}^{K} N(\mu_k, 1)$$

where $\mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K$ *are unknown. For every* $K \geq 1$*, we have a model of the form,*

$$\mathscr{P}_K = \{P_{K;\mu_1,\ldots,\mu_K} : \mu_1, \ldots, \mu_K \in \mathbb{R}\}$$

Each of these models can be endowed with a prior Π_K *on* $\mathbb{R}^K$*, for example, by declaring* $\mu_1, \ldots, \mu_K$ *independent and marginally distributed standard normal:*

$$\mu \sim \Pi_K = N(0, I_K).$$

At this point, a Bayesian would choose a hyperprior Π_2 *for the discrete hyperparameter* $K \geq 1$ *and proceed to calculate the posterior on* all models $\mathscr{P}_K$*, weighed by the prior masses* $\Pi_2(K = k)$ *for all* $k \geq 1$*. Alternatively, the Bayesian can use Bayes' factors to make a decision as to which value of* K *to use, reducing the analysis to a selected, or estimated value for* K*.*

Here, we concentrate on the frequentist approach. The frequentist also aims to select one of the models $\mathscr{P}_K$*: in the empirical Bayes approach, we "point-estimate" which model-prior combination we shall be using to analyse the data, from the choices* $(\mathscr{P}_K, \Pi_K)$*,* $K \geq 1$*. In such a case, inspection of the data may reveal which number of classes is most appropriate, if one observes clearly separated peaks in the observations, in accordance with the second point made in remark 3.3.7. Otherwise, frequentist methods exist to estimate* K*, for instance from a larger population of specimens. After we have an estimate* $\hat{K}$ *for* K*, we are in a position to calculate the posterior for* μ *based on* $(\mathscr{P}_{\hat{K}}, \Pi_{\hat{K}})$*.*

There are two remarks to be made with regard to the estimation of K *from a larger population of specimens: first of all, maximization of the likelihood will always lead to a number of classes in the order of the samplesize, simply because the largest number of classes offers the most freedom and hence always provides the best fit to the data. A similar phenomenon arises in regression, where it is called over-fitting, if we allow regression polynomials of arbitrary degree: the MLE will fit the data perfectly by choosing a polynomial of degree in the order of the samplesize. Therefore in such questions of model selection, penalized likelihood criteria are employed which favour low-dimensional models over high-dimensional ones, i.e. smaller choices for* K *over larger ones. Note that it is not clear, neither intuitively nor mathematically, how the penalty should depend on* K*, nor which proportionality between penalty and likelihood is appropriate (see, however, the AIC and BIC criteria for model selection [77]). The Bayesian faces the same problem when he chooses a prior for* K*: if he assigns too much prior weight to the higher-dimensional models, his estimators (or, equivalently, the bulk of the resulting posterior's mass) will get the chance to "run off" to infinity with growing samplesize, indicating inconsistency from over-fitting. Indeed, the correspondence between the frequentist's necessity for a penalty in maximum-likelihood methods on the one hand, and the Bayesian's need for a prior expressing sufficient bias for the lower-dimensional model choices on the other, is explained in remark 2.2.7.*

On another sidenote: it is crucial in the example above that all classes are represented in equal proportions. Otherwise identifiability and testability problems arise and persist even after we decide to exclude from the model the vectors μ *which have* $\mu_i = \mu_j$ *for some* $i \neq j$*.*

If one imagines the situation where the number of observations is of the same order as the number of classes, this should come as no surprise.

A less ambitious application of empirical Bayesian methods is the estimation of hyperparameters by maximum-likelihood estimation through the prior predictive distribution (see definition 2.1.4). Recall that the marginal distribution of the data in the subjectivist Bayesian formulation (*c.f.* section 2.1) predicts how the data is distributed. This prediction may be reversed to decide which value for the hyperparameter leads to the best explanation of the observed data, where our notion of "best" is based on the likelihood principle.

Denote the data by Y and assume that it takes its values in a measurable space $(\mathscr{Y}, \mathscr{B})$. Denote the model by $\mathscr{P} = \{P_\theta : \theta \in \Theta_0\}$. Consider a family of priors parametrized by a hyperparameter $\eta \in H$, $\{\Pi_\eta : \eta \in H\}$. For every η, the prior predictive distribution P_η is given by:

$$P_\eta(A) = \int_\Theta P_\theta(A)\, d\Pi_\eta(\theta),$$

for all $A \in \mathscr{B}$, *i.e.* we obtain a new model for the observation Y, given by $\mathscr{P}' = \{P_\eta : \eta \in H\}$, contained in the convex hull of the original model $\mathrm{co}(\mathscr{P})$. Note that this new model is parametrized by the hyperparameter; hence if we close our eyes to the rest of the problem and we follow the maximum-likelihood procedure for estimation of η in this new model, we find the value of the hyperparameter that best explains the observation Y. Assuming that the model $\mathscr{P}'$ is dominated, with densities $\{p_\eta : \eta \in H\}$, the maximum-likelihood estimate is found as the point $\hat{\eta}(Y) \in H$ such that

$$p_{\hat{\eta}}(Y) = \sup_{\eta \in H} p_\eta(Y).$$

by the usual methods, analytically or numerically.

Definition 3.3.5. *The estimator $\hat{\eta}(Y)$ is called the ML-II estimator, provided it exists and is unique.*

Remark 3.3.8. *There is one caveat that applies to the ML-II approach: in case the data Y consists of an i.i.d.-distributed sample, the prior predictive distribution describes the sample as exchangeable, but not i.i.d.! Hence, comparison of prior predictive distributions with the data suffer from the objection raised in remark 2.1.3. The frequentist who assumes that the true, underlying distribution P_0^n of the sample is i.i.d., therefore has to keep in mind that the ML-II model is misspecified. By the law of large numbers, the maximum-likelihood estimator $\hat{\eta}_n(X_1, \ldots, X_n)$ will converge asymptotically to the set of points S in H that minimize the Kullback-Leibler divergence, i.e. those $\eta^* \in H$ such that:*

$$-P_0 \log \frac{p_{\eta^*}}{p_0} = \inf_{\eta \in H} -P_0 \log \frac{p_\eta}{p_0},$$

provided that such points exist. (What happens otherwise is left as an exercise to the reader.)

Example 3.3.4. *Consider the example with which we began this section: the data Y is normally distributed with unknown mean θ and known variance σ^2. The prior for θ is chosen normal with mean 0 and variance τ^2.*

3.4 Dirichlet process priors

The construction of priors on non-parametric models is far from trivial. Broadly, there are two mathematical reasons for this: whereas the usual norm topology on $\mathbb{R}^k$ is unique (in the sense that all other norm topologies are equivalent, see [67]), infinite-dimensional vector spaces support many different norm topologies and various other topologies besides. Similarly, whereas on $\mathbb{R}^k$ the (unique shift-invariant) Lebesgue measure provides a solid foundation for the definition of models in terms of densities, no such default uniform dominating measure exists in infinite-dimensional spaces.

Nevertheless, there are constructions of probability measures on infinite-dimensional spaces, for example so-called Gaussian measures on Banach and Hilbert spaces. Some of these constructions and the properties of the measures they result in, are discussed in great detail in Ghosh and Ramamoorthi (2003) [42]. In this section, we look at a class of priors first proposed by Ferguson (1973) [34], which have become known as Dirichlet process priors.

The Dirichlet process prior arises as the non-parametric analog of the Dirichlet distribution on finite-dimensional spaces of probability distributions, which we consider in some detail first. Let $\mathscr{X} = \{1, 2, \ldots, k\}$ (with its powerset $2^{\mathscr{X}}$ as a σ-algebra) and consider the collection $M(\mathscr{X})$ of all probability measures on $\mathscr{X}$. Every $P \in M(\mathscr{X})$ has a density $p : \mathscr{X} \to [0, 1]$ (with respect to the counting measure on $\mathscr{X}$) and we denote $p_i = p(i) = P(\{i\})$, so that for every $A \in 2^{\mathscr{X}}$,

$$P(A) = \sum_{l \in A} p_l.$$

Therefore, the space $M(\mathscr{X})$ can be parametrized as follows,

$$M(\mathscr{X}) = \Big\{P : 2^{\mathscr{X}} \to [0, 1] \,:\, \sum_{i=1}^{k} p_i = 1,\, p_i \geq 0,\, (1 \leq i \leq k)\Big\},$$

and is in bijective correspondence with the simplex in $\mathbb{R}^k$. For reasons to be discussed shortly, we consider the following family of distributions on $M(\mathscr{X})$.

Definition 3.4.1. *(Finite-dimensional Dirichlet distribution)*
Let $\alpha = (\alpha_1, \ldots, \alpha_k)$ with $\alpha_i > 0$ for all $1 \leq i \leq k$. A stochastic vector $p = (p_1, \ldots, p_k)$ is said to have Dirichlet distribution D_α with parameter α, if the density π for p satisfies:

$$\pi(p) = \frac{\Gamma\big(\sum_{i=1}^{k} \alpha_i\big)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1}\, p_2^{\alpha_2 - 1} \ldots p_{k-1}^{\alpha_{k-1} - 1} \Big(1 - \sum_{i=1}^{k-1} p_i\Big)^{\alpha_k - 1}$$

If $\alpha_i = 0$ for some i, $1 \leq \alpha_i \leq k$, then we set $D_\alpha(p_i = 0) = 1$ marginally and we treat the remaining components of p as $(k-1)$-dimensional.

As an example, consider the case where $k = 2$ (so that $p_2 = 1 - p_1$): in that case, the density of the Dirichlet distribution takes the form:

$$\pi(p_1, p_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\, \Gamma(\alpha_2)} p_1^{\alpha_1 - 1} (1 - p_1)^{\alpha_2 - 1},$$

i.e. p_1 has a Beta distribution $B(\alpha_1, \alpha_2)$. Examples of graphs of Beta densities with $\alpha_1 = k+1$, $\alpha_2 = n - k + 1$ for various integer values of k are depicted in figure 2.1). We also note the following two well-known facts on the Dirichlet distribution (proofs can be found in [42]).

Lemma 3.4.1. *(Gamma-representation of D_α)*
If $Z_1, \ldots, Z_k$ are independent and each marginally distributed according to a Γ-distribution with parameter α_i, i.e.

$$Z_i \sim \Gamma(\alpha_i),$$

for all $1 \leq i \leq k$, then the normalized vector

$$\Big(\frac{Z_1}{S}, \ldots, \frac{Z_k}{S}\Big) \sim D_\alpha, \tag{3.14}$$

with $S = \sum_{i=1}^k Z_i$.

Lemma 3.4.1 shows that we may think of a D_α-distributed vector as being composed of k independent, Γ-distributed components, normalized to form a probability distribution, through division by S in (3.14). This division should be viewed as an L_1-projection from the positive cone in $\mathbb{R}^k$ onto the $k-1$-dimensional simplex. The following property can also be viewed as a statement on the effect of a projection on a distribution, this time from the simplex in $\mathbb{R}^k$ to lower-dimensional simplices. It is this property (related to a property called *infinite divisibility* of the Dirichlet distribution) that motivates the choice for the Dirichlet distribution made by definition 3.4.1.

Lemma 3.4.2. *Let $\mathscr{X}$ be a finite pointset. If the density $p : \mathscr{X} \to [0,1]$ of a distribution P is itself distributed according to a Dirichlet distribution with parameter α, $p \sim D_\alpha$, then for any partition $\{A_1, \ldots, A_m\}$ of $\mathscr{X}$, the vector of probabilities $(P(A_1), P(A_2), \ldots, P(A_m))$ has a Dirichlet distribution again,*

$$\big(P(A_1), P(A_2), \ldots, P(A_m)\big) \sim D_{\alpha'},$$

where the parameter α' is given by:

$$(\alpha'_1, \ldots, \alpha'_m) = \left(\sum_{l \in A_1} \alpha_l, \ldots, \sum_{l \in A_m} \alpha_l \right). \tag{3.15}$$

The identification (3.15) in lemma 3.4.2 suggests that we adopt a slightly different perspective on the definition of the Dirichlet distribution: we view α as a *finite measure* on $\mathscr{X}$, so that $P \sim D_\alpha$, if and only if, for every partition $(A_1, \ldots, A_m)$,

$$\big(P(A_1), \ldots, P(A_m)\big) \sim D_{(\alpha(A_1), \ldots, \alpha(A_m))}. \tag{3.16}$$

Property (3.16) serves as the point of departure of the generalization to the non-parametric model, because it does not depend on the finite nature of $\mathscr{X}$.

Definition 3.4.2. *Let $\mathscr{X}$ be a finite pointset; denote the collection of all probability measures on $\mathscr{X}$ by $M(\mathscr{X})$. The Dirichlet family $\mathscr{D}(\mathscr{X})$ is defined to be the collection of all Dirichlet distributions on $M(\mathscr{X})$, i.e. $\mathscr{D}(\mathscr{X})$ consists of all D_α with α a finite measure on $\mathscr{X}$.*

The following property of the Dirichlet distribution describes two independent Dirichlet-distributed quantities in convex combination, which form a new Dirichlet-distributed quantity if mixed by means of an (independent) Beta-distributed parameter.

Lemma 3.4.3. *Let $\mathscr{X}$ be a finite pointset and let α_1, α_2 be two measures on $(\mathscr{X}, 2^{\mathscr{X}})$. Let (P_1, P_2) be independent and marginally distributed as*

$$P_1 \sim D_{\alpha_1}, \quad P_2 \sim D_{\alpha_2}.$$

Furthermore, let λ be independent of P_1, P_2 and marginally distributed according to $\lambda \sim B(\alpha_1(\mathscr{X}), \alpha_2(\mathscr{X}))$. Then the convex combination $\lambda P_1 + (1-\lambda) P_2$ again has a Dirichlet distribution with base measure $\alpha_1 + \alpha_2$:

$$\lambda P_1 + (1-\lambda) P_2 \sim D_{\alpha_1+\alpha_2}.$$

Many other properties of the Dirichlet distribution could be considered here, most notably the so-called *tail-free property* and *neurality to the right* (see [42]). We do not provide details because both are rather technical and we do not use them in following chapters, but the reader should be aware of their existence because some authors use them extensively.

A most important property of the family of Dirichlet distributions is its conjugacy for the full non-parametric model.

Theorem 3.4.1. *Let $\mathscr{X}$ be a finite pointset; let $X_1, \ldots, X_n$ denote an i.i.d. sample of observations taking values in $\mathscr{X}$. The Dirichlet family $\mathscr{D}(\mathscr{X})$ is a conjugate family: if the prior equals D_α, the posterior equals $D_{\alpha+n\mathbb{P}_n}$.*

Proof Since $\mathscr{X}$ is finite ($\#(\mathscr{X}) = k$), $M(\mathscr{X})$ is dominated (by the counting measure), so the posterior can be written as in (2.8). The likelihood takes the form:

$$P \mapsto \prod_{i=1}^{n} p(X_i) = \prod_{l=1}^{k} p_l^{n_l},$$

where $n_l = \#\{X_i = l : 1 \le i \le n\}$. Multiplying by the prior density for $\Pi = D_\alpha$, we find that the posterior density is proportional to,

$$\begin{aligned}\pi(p_1, \ldots, p_k | X_1, \ldots, X_n) &\propto \pi(p_1, \ldots, p_k) \prod_{i=1}^{n} p_{X_i} \\ &\propto \prod_{l=1}^{k} p_l^{n_l} \prod_{l=1}^{k-1} p_l^{\alpha_l - 1} \Big(1 - \sum_{i=1}^{k-1} p_i\Big)^{\alpha_k - 1} = \prod_{l=1}^{k-1} p_l^{\alpha_l + n_l - 1} \Big(1 - \sum_{i=1}^{k-1} p_i\Big)^{\alpha_k + n_k - 1},\end{aligned}$$

which is again a Dirichlet density, but with changed base measure α. Since the posterior is a probability distribution, we know that the normalization factor follows suit. Noting that we may view n_l as the density of the measure $n\mathbb{P}_n$ since

$$n_l = \sum_{i=1}^{n} 1\{X_i = l\} = n\mathbb{P}_n 1\{X = l\},$$

we complete the argument. □

Next we consider the Dirichlet process prior, a probability measure on the full non-parametric model for a measurable space $(\mathscr{X}, \mathscr{B})$. For the sake of simplicity, we assume that $\mathscr{X} = \mathbb{R}$ and $\mathscr{B}$ is the Borel σ-algebra on $\mathbb{R}$. We denote the collection of all probability measures on $(\mathbb{R}, \mathscr{B})$ by $M(\mathbb{R}, \mathscr{B})$. We consider the collection of random quantities $\{P(A) : A \in \mathscr{B}\}$ and impose two straightforward conditions on its finite-dimensional marginals. The Kolmogorov existence theorem (see theorem A.5.1) then guarantees existence of a stochastic process with finitely additive sample path $P : \mathscr{B} \to [0, 1]$. Said straightforward conditions are satisfied if we choose the finite-dimensional marginal distributions to be (finite-dimensional) Dirichlet distributions (3.16). Also by this choice, σ-additivity of P can be guaranteed. The resulting process on the space of all probability measures on $(\mathscr{X}, \mathscr{B})$ is called the called the *Dirichlet process* and the associated probability measure Π is called the Dirichlet process prior.

Theorem 3.4.2. *(Existence of the Dirichlet process)*
Given a finite measure α on $(\mathbb{R}, \mathscr{B})$, there exists a probability measure D_α on $M(\mathbb{R}, \mathscr{B})$ (called the Dirichlet process prior with parameter α) such that for $P \sim D_\alpha$ and every $\mathscr{B}$-measurable partition $(B_1, \ldots, B_k)$ of $\mathbb{R}$,

$$\big(P(B_1), \ldots, P(B_k)\big) \sim D_{(\alpha(B_1),\ldots,\alpha(B_k))}. \tag{3.17}$$

Proof Let $k \geq 1$ and $A_1, \ldots, A_k \in \mathscr{B}$ be given. Through the indicators 1_{A_i} for these sets, we define 2^k new sets

$$1_{B_{\nu_1 \ldots \nu_k}} = \prod_{i=1}^{k} 1_{A_i}^{\nu_i} (1 - 1_{A_i})^{1-\nu_i},$$

where $\nu_1, \ldots, \nu_k \in \{0, 1\}$. Then the collection $\{B_{\nu_1 \ldots \nu_k} : \nu_i \in \{0, 1\}, 1 \leq i \leq k\}$ forms a partition of $\mathbb{R}$. For the P-probabilities corresponding to this partition, we assume finite-dimensional marginals

$$\big(P(B_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0, 1\}, 1 \leq i \leq k\big) \sim \Pi_{B_{\nu_1 \ldots \nu_k} : \nu_i \in \{0,1\}, 1 \leq i \leq k},$$

The distribution of the vector $(P(A_1), \ldots, P(A_k))$ then follows from the definition:

$$P(A_i) = \sum_{\{i : \nu_i = 1\}} P(B_{\nu_1 \ldots \nu_k}),$$

for all $1 \leq i \leq k$. This defines marginal distributions for all finite subsets of $\mathscr{B}$, as needed in theorem A.5.1. To define the underlying probability space $(\Omega, \mathscr{F}, \Pi)$ we now impose two conditions.

(F1) With Π-probability one, the empty set has P-measure zero:

$$\Pi\big(P(\emptyset) = 0\big) = 1.$$

(F2) Let $k, k' \geq 1$ be given. If $(B_1, \ldots, B_k)$ is a partition and $(B'_1, \ldots, B'_{k'})$ a refinement thereof, with

$$B_1 = \bigcup_{i=1}^{r_1} B'_i, \quad \ldots \quad , \quad B_k = \bigcup_{i=r_{k-1}+1}^{k'} B'_i,$$

(for certain $r_1 < \ldots < r_{k-1}$), then we have the following equality in distribution:

$$\mathscr{L}\Big(\sum_{i=1}^{r_1} P(B'_i), \ldots, \sum_{i=r_{k-1}+1}^{k'} P(B'_i)\Big) = \mathscr{L}\big(P(B_1), \ldots, P(B_k)\big).$$

Condition (F1) ensures that if $(A_1, \ldots, A_k)$ is itself a partition of $\mathbb{R}$, the above construction does not lead to a contradiction. Condition (F2) ensures finite additivity of P with prior probability one, *i.e.* for any $A, B, C \in \mathscr{B}$ such that $A \cap B = \emptyset$ and $A \cup B = C$,

$$\Pi\big(\, P(A) + P(B) = P(C) \,\big) = 1. \tag{3.18}$$

Ferguson (1973,1974) [34, 35] has shown that conditions (F1) and (F2) imply that Kolmogorov's consistency conditions (K1) and (K2) (see section A.5) are satisfied. As we have seen in the first part of this section, if we impose the Dirichlet distribution:

$$\big(P(B_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0,1\}, 1 \leq i \leq k\big) \sim D_{\{\alpha(B_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0,1\}, 1 \leq i \leq k\}}. \tag{3.19}$$

and α is a measure on $\mathscr{B}$, condition (F2) is satisfied. Combining all of this, we conclude that there exists a probability space $(\Omega, \mathscr{F}, \Pi)$ on which the stochastic process $\{P(A) : A \in \mathscr{B}\}$ can be represented with finite dimensional marginals *c.f.* (3.19). Lemma 3.4.4 shows that $\Pi(P \in M(\mathbb{R}, \mathscr{B})) = 1$, completing the proof. □

The last line in the above proof may require some further explanation: P is merely the sample-path of our stochastic process. The notation $P(A)$ suggests that P is a probability measure, but all we have shown up to that point, is that (F1) and (F2) imply that P is a finitely additive set-function such that:

$$\Pi\big(P(B) \in [0,1]\big) = 1,$$

with Π-probability equal to one. What remains to be demonstrated is Π-almost-sure σ-additivity of P.

Lemma 3.4.4. *If Π is a Dirichlet process prior D_α on $M(\mathscr{X}, \mathscr{B})$,*

$$\Pi\big(\, P \textit{ is } \sigma\textit{-additive} \,\big) = 1.$$

Proof Let $(A_n)_{n\geq 1}$ be a sequence in $\mathscr{B}$ that decreases to $\emptyset$. Since α is σ-additive, $\alpha(A_n) \to \alpha(\emptyset) = 0$. Therefore, there exists a subsequence $(A_{n_j})_{j\geq 1}$ such that $\sum_j \alpha(A_{n_j}) < \infty$. For fixed $\epsilon > 0$, using Markov's inequality first,

$$\sum_{j\geq 1} \Pi\big(P(A_{n_j}) > \epsilon\big) \leq \sum_{j\geq 1} \frac{1}{\epsilon} \int P(A_{n_j})\, d\Pi(P) = \frac{1}{\epsilon} \sum_{j\geq 1} \frac{\alpha(A_{n_j})}{\alpha(\mathbb{R})} < \infty,$$

according to lemma 3.4.5. From the Borel-Cantelli lemma (see lemma A.2.1), we see that

$$\Pi\big(\limsup_{j\to\infty} \{P(A_{n_j}) > \epsilon\}\big) = \Pi\Big(\bigcap_{J\geq 1} \bigcup_{j\geq J} \{P(A_{n_j}) > \epsilon\}\Big) = 0,$$

which shows that $\lim_j P(A_{n_j}) = 0$, Π-almost-surely. Since, by Π-almost-sure finite additivity of P,

$$\Pi\big(\, P(A_n) \geq P(A_{n+1}) \geq \dots \big) = 1,$$

we conclude that $\lim_n P(A_n) = 0$, Π-almost-surely. By the continuity theorem for measures (see theorem A.2.1 and the proof in [52], theorem 3.2), P is σ-additive Π-almost-surely. □

The proof of lemma 3.4.4 makes use of the following lemma, which establishes the basic properties of the Dirichlet process prior.

Lemma 3.4.5. *Let α be a finite measure on $(\mathbb{R}, \mathscr{B})$ and let $\{P(A) : A \in \mathscr{B}\}$ be the associated Dirichlet process with distribution D_α. Let $B \in \mathscr{B}$ be given.*

(i) If $\alpha(B) = 0$, then $P(B) = 0$, $\Pi - a.s.$

(ii) If $\alpha(B) > 0$, then $P(B) > 0$, $\Pi - a.s.$

(iii) The expectation of P under D_α is given by

$$\int P(B)\, dD_\alpha(P) = \frac{\alpha(B)}{\alpha(\mathbb{R})}.$$

Proof Let $B \in \mathscr{B}$ be given. Consider the partition (B_1, B_2) of $\mathbb{R}$, where $B_1 = B$, $B_2 = \mathbb{R}\setminus B$. According to (3.17),

$$\big(P(B_1), P(B_2)\big) \sim D_{(\alpha(B), \alpha(\mathbb{R})-\alpha(B))},$$

so that $P(B) \sim B(\alpha(B), \alpha(\mathbb{R}) - \alpha(B))$. Stated properties then follow from the properties of the Beta-distribution. □

This concludes the proof for the existence of Dirichlet processes and the associated priors. One may then wonder what is the nature of the prior we have constructed. As it turns out, the Dirichlet process prior has some remarkable properties.

Lemma 3.4.6. *(Support of the Dirichlet process prior)*
Consider $M(\mathbb{R}, \mathscr{B})$, endowed with the topology of weak convergence. Let α be a finite measure on $(\mathbb{R}, \mathscr{B})$. The support of D_α is given by

$$M_\alpha(\mathbb{R}, \mathscr{B}) = \big\{P \in M(\mathbb{R}, \mathscr{B}) \,:\, \mathrm{supp}(P) \subset \mathrm{supp}(\alpha)\big\}.$$

In fact, we can be more precise, as shown in the following lemma.

Lemma 3.4.7. *Let α be a finite measure on $(\mathbb{R}, \mathscr{B})$ and let $\{P(A) : A \in \mathscr{B}\}$ be the associated Dirichlet process with distribution D_α. Let $Q \in M(\mathbb{R}, \mathscr{B})$ be such that $Q \ll \alpha$. Then, for any $m \geq 1$ and $A_1, \ldots, A_m \in \mathscr{B}$ and $\epsilon > 0$,*

$$D_\alpha\big\{P \in M(\mathbb{R}, \mathscr{B}) \,:\, |P(A_i) - Q(A_i)| < \epsilon, 1 \leq i \leq m\big\} > 0.$$

Proof The proof of this lemma can be found in [42], theorem 3.2.4. □

So if we endow $M(\mathbb{R}, \mathscr{B})$ with the (slightly stronger) topology of pointwise onvergence (see definition A.7.2), the support of D_α remains large, consisting of all $P \in M(\mathbb{R}, \mathscr{B})$ that are dominated by α.

The following property reveals a most remarkable characterization of Dirichlet process priors: the subset $D(\mathbb{R}, \mathscr{B})$ of all finite convex combinations of Dirac measures (see example A.2.2) receives prior mass equal to one.

Lemma 3.4.8. *Let α be a finite measure on $(\mathbb{R}, \mathscr{B})$ and let $\{P(A) : A \in \mathscr{B}\}$ be the associated Dirichlet process with distribution D_α. Then,*

$$D_\alpha\big\{P \in D(\mathbb{R}, \mathscr{B})\big\} = 1.$$

Proof The proof of this lemma can be found in [42], theorem 3.2.3. □

The above phenomenon leads to problems with support or convergence in stronger topologies (like total variation or Hellinger topologies) and with regard to the Kullback-Leibler criteria mentioned in the asymptotic theorems of chapter 4. Generalizing this statement somewhat, we may infer from the above that the Dirichlet process prior is not suited to (direct) estimation of densities. Although clearly dense enough in $M(\mathbb{R}, \mathscr{B})$ in the toplogy of weak convergence, the set $D(\mathbb{R}, \mathscr{B})$ may be rather sparse in stronger topologies! (Notwithstanding the fact that mixture models with a Dirichlet process prior for the mixing distribution can be (minimax) optimal for the estimation of mixture densities [41].)

Lemma 3.4.9. *Let α be a finite measure on $(\mathbb{R}, \mathscr{B})$ and let $\{P(A) : A \in \mathscr{B}\}$ be the associated Dirichlet process with distribution D_α. Let $g : \mathbb{R} \to \mathbb{R}$ be non-negative and Borel-measurable. Then,*

$$\int_{\mathbb{R}} g(x)\, d\alpha(x) < \infty \quad \Leftrightarrow \quad \int_{\mathbb{R}} g(x)\, dP(x) < \infty, \quad (D_\alpha - a.s.).$$

Proof Add proof! □

Perhaps the most important result of this section is the fact that the family of Dirichlet process priors on $M(\mathbb{R}, \mathscr{B})$ is a *conjugate family* for the full, non-parametric model on $(\mathbb{R}, \mathscr{B})$, as stated in the following theorem.

Theorem 3.4.3. *Let $X_1, X_2, \ldots$ be an i.i.d. sample of observations in $\mathbb{R}$. Let α be a finite measure on $(\mathbb{R}, \mathscr{B})$ with associated Dirichlet process prior $\Pi = D_\alpha$. For any measurable $C \subset M(\mathbb{R}, \mathscr{B})$,*

$$\Pi\big(P \in C \mid X_1, \ldots, X_n\big) = D_{\alpha + n\mathbb{P}_n}(C),$$

i.e. the posterior is again a Dirichlet process distribution, with base measure $\alpha + n\mathbb{P}_n$

Proof The proof is a direct consequence of theorem 3.4.1 and the fact that equality of two measures on a generating ring implies equality on the whole σ-algebra. (Cylindersets generate the relevant σ-algebra and for cylindersets, theorem 3.4.1 asserts equality.) □

Example 3.4.1. *Let $X_1, X_2, \ldots$ be an i.i.d. sample of observations in $\mathbb{R}$. Let α be a finite measure on $(\mathbb{R}, \mathscr{B})$ with associated Dirichlet process prior $\Pi = D_\alpha$. Let $B \in \mathscr{B}$ be given. The expectation of $P(B)$ under the prior distribution equals,*

$$\int P(B)\, dD_\alpha(P) = \frac{\alpha(B)}{\alpha(\mathbb{R})}, \tag{3.20}$$

the measure of B under α normalized to be a probability measure (which we denote by $P_\alpha(B)$). The posterior mean (see definition 2.2.1), is then given by:

$$\begin{aligned}\int P(B)\, d\Pi\big(P \mid X_1, \ldots, X_n\big) = \int P(B)\, dD_{\alpha + n\mathbb{P}_n}(P) &= \frac{(\alpha + n\mathbb{P}_n)(B)}{(\alpha + n\mathbb{P}_n)(B)} \\ &= \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + n} P_\alpha(B) + \frac{n}{\alpha(\mathbb{R}) + n}\mathbb{P}_n(B),\end{aligned}$$

P_0^n-almost-surely. Defining $\lambda_n = \alpha(\mathbb{R})/(\alpha(\mathbb{R}) + n)$, we see that the posterior mean $\hat{P}_n$ can be viewed as a convex combination of the prior mean distribution and the empirical distributions,

$$\hat{P}_n = \lambda_n\, P_\alpha + (1 - \lambda_n)\mathbb{P}_n,$$

P_0^n-almost-surely. As a result, we see that

$$\|\hat{P}_n - \mathbb{P}_n\|_{TV} = \lambda_n \|P_\alpha - \mathbb{P}_n\| \leq \lambda_n,$$

P_0^n-almost-surely. Since $\lambda_n \to 0$ as $n \to \infty$, the difference between the sequence of posterior means $(\hat{P}_n)_{n \geq 1}$ and the sequence of empirical measures $(\mathbb{P}_n)_{n \geq 1}$ converges to zero in total variation as we let the samplesize grow to infinity. Generalizing likelihood methods to non-dominated models, Dvoretzky, Kiefer and Wolfowitz (1956) [30] have shown that the empirical distribution $\mathbb{P}_n$ can be viewed as the non-parametric maximum-likelihood estimator (usually abbreviated NPMLE). This establishes (an almost-sure form of) consistency for the posterior mean, in the sense that it has the same point of convergence as the NPMLE. In chapter 4, convergence of the posterior distribution (and in particular its mean) to the MLE will manifest itself as a central connection between frequentist and Bayesian statistics.

Remark 3.4.1. *The above example provides the subjectivist with a guideline for the choice of the base measure α. More particularly, equality (3.20) says that the prior predictive distribution equals the (normalized) base measure α. In view of the fact that subjectivists should choose the prior to reflect their prior "beliefs", α should therefore be chosen such that it assigns relatively high mass to sets $B \in \mathscr{B}$ that are believed to be probable.*

3.5 Exercises

Exercise 3.1. A PROPER JEFFREYS PRIOR

Let X be a random variable, distributed $Bin(n;p)$ for known n and unknown $p \in (0,1)$. Calculate Jeffreys prior for this model, identify the standard family of probability distributions it belongs to and conclude that this Jeffreys prior is proper.

Exercise 3.2. JEFFREYS AND UNIFORM PRIORS

Let $\mathscr{P}$ be a model parametrized according to some mapping $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$. Assuming differentiability of this map, Jeffreys prior Π takes the form (3.7). In other parametrizations, the form *of this expression remains the same, but the actual dependence on the parameter changes. This makes it possible that there exists another parametrization of $\mathscr{P}$ such that Jeffreys prior is* equal *to the uniform prior. We shall explore this possibility in three exercises below.*

For each of the following models in their 'standard' parametrizations $\theta \mapsto P_\theta$, find a parameter $\eta \in H$, $\eta = \eta(\theta)$, such that the Fisher information I_η, expressed in terms of η, is constant.

a. *Find η for $\mathscr{P}$ the model of all Poission distributions.*

b. *In the cases $\alpha = 1, 2, 3$, find η for the model $\mathscr{P}$ consisting of all $\Gamma(\alpha, \theta)$-distributions, with $\theta \in (0, \infty)$.*

c. *Find η for the model $\mathscr{P}$ of all $Bin(n;\theta)$ distributions, where n is known and $\theta \in (0,1)$. Note that if the Fisher information I_η is constant, Jeffries prior is uniform. Therefore, if H is unbounded, Jeffries prior is improper.*

Exercise 3.3. OPTIMALITY OF UNBIASED BAYESIAN POINT ESTIMATORS

Let $\mathscr{P}$ be a dominated, parametric model, parametrized identifiably by $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, for some $\Theta \subset \mathbb{R}^k$. Assume that $(X_1, \ldots, X_n) \in \mathscr{X}^n$ form an i.i.d. sample from a distribution $P_0 = P_{\theta_0} \in \mathscr{P}$, for some $\theta_0 \in \Theta$. Let Π be a prior on Θ and denote the posterior by $\Pi(\cdot|X_1, \ldots, X_n)$. Assume that $T : \mathscr{X}^n \to \mathbb{R}^m$ is a sufficient statistic for the model $\mathscr{P}$.

a. *Use the factorization theorem to show that the posterior depends on the data only through the sufficient statistic $T(X_1, \ldots, X_n)$.*

b. *Let $\hat{\theta}_n : \mathscr{X}^n \to \Theta$ denote a point-estimator derived from the posterior. Use a. above to argue that there exists a function $\tilde{\theta}_n : \mathbb{R}^m \to \Theta$, such that,*

$$\hat{\theta}_n(X_1, \ldots, X_n) = \tilde{\theta}_n(T(X_1, \ldots, X_n)).$$

Bayesian point-estimators share this property with other point-estimators that are derived from the likelihood function, like the maximum-likelihood estimator and penalized versions thereof. Next, assume that $P_0^n(\hat{\theta}_n)^2 < \infty$ *and that* $\hat{\theta}_n$ *is unbiased, i.e.* $P_0^n\hat{\theta}_n = \theta_0$.

c. *Apply the Lehmann-Scheffé theorem to prove that, for any other unbiased estimator* $\hat{\theta}'_n : \mathscr{X}^n \mapsto \Theta$,

$$P_0^n(\hat{\theta}_n - \theta_0)^2 \leq P_0^n(\hat{\theta}'_n - \theta_0)^2.$$

The message of this exercise is, that Bayesian point-estimators that happen to be unbiased and quadratically integrable, are automatically L_2*-optimal in the class of all unbiased estimators for* θ. *They share this remarkable property with maximum-likelihood estimators.*

Exercise 3.4. CONJUGATE MODEL-PRIOR PAIRS
In this exercise, conjugate model-prior pairs $(\mathscr{P}, \Pi)$ *are provided. In each case, we denote the parameter we wish to estimate by* θ *and assume that other parameters have known values. Let* X *denote a single observation.*
In each case, derive the posterior distribution to prove conjugacy and identify the X*-dependent transformation of parameters that takes prior into posterior.*

a. $X|\theta \sim N(\theta, \sigma^2)$ *and* $\theta \sim N(\mu, \tau^2)$.

b. $X|\theta \sim \text{Poisson}(\theta)$ *and* $\theta \sim \Gamma(\alpha, \beta)$.

c. $X|\theta \sim \Gamma(\nu, \theta)$ *and* $\theta \sim \Gamma(\alpha, \beta)$.

d. $X|\theta \sim \text{Bin}(n; \theta)$ *and* $\theta \sim \beta(\alpha, \beta)$.

e. $X|\theta \sim N(\mu, \theta^{-1})$ *and* $\theta \sim \Gamma(\alpha, \beta)$.

f. $X|\theta_1, \ldots, \theta_k \sim M(n; \theta_1, \ldots, \theta_k)$ *and* $\theta \sim D_\alpha$, *where* M *denotes the multinomial distribution for* n *observations drawn from* k *classes with probabilities* $\theta_1, \ldots, \theta_k$ *and* D_α *is a Dirichlet distribution on the simplex in* $\mathbb{R}^k$ *(see definition 3.4.1).*

Exercise 3.5. *In this exercise, we generalize the setup of example 3.3.2 to multinomial rather than binomial context. Let* $k \geq 1$ *be known. Consider an observed random variable* Y *and an unobserved* $N = 1, 2, \ldots$, *such that, conditionally on* N, Y *is distributed multinomially over* k *classes, while* N *has a Poisson distribution with hyperparameter* $\lambda > 0$,

$$Y|N \sim M_k(N; p_1, p_2, \ldots, p_k), \qquad N \sim \text{Poisson}(\lambda).$$

Determine the prior predictive distribution of Y, *as a function of the hyperparameter* λ.

Exercise 3.6. *Let* $X_1, \ldots, X_n$ *form an i.i.d. sample from a Poisson distribution* $\text{Poisson}(\theta)$ *with unknown* $\theta > 0$. *As a family of possible priors for the Bayesian analysis of this data, consider exponential distributions* $\theta \sim \Pi_\lambda = \text{Exp}(\lambda)$, *where* $\lambda > 0$ *is a hyperparameter.*
Calculate the prior predictive distribution for X *and the ML-II estimate* $\hat{\lambda}$. *With this estimated hyperparameter, give the posterior distribution* $\theta|X_1, \ldots, X_n$. *Calculate the resulting posterior mean and comment on its data-dependence.*

Exercise 3.7. *Let $X_1, \ldots, X_n$ form an i.i.d. sample from a binomial distribution* $\mathrm{Bin}(n; p)$*, given $p \in [0, 1]$. For the parameter p we impose a prior $p \sim \beta(\alpha, \beta)$ with hyperparameters $\alpha, \beta > 0$.*

Show that the family of β-distributions is conjugate for binomial data. Using (standard expressions for) the expectation and variance of β-distributions, give the posterior mean and variance in terms of the original α and β chosen for the prior and the data. Calculate the prior predictive distribution and give frequentist estimates for α and β. Substitute the result in the posterior mean and comment on (asymptotic) data dependence of the eventual point estimator.

Appendix A

Measure theory

In this appendix we collect some important notions from measure theory. The goal is not to present a self-contained presentation, but rather to establish the basic definitions and theorems from the theory for reference in the main text. As such, the presentation omits certain existence theorems and many of the proofs of other theorems (although references are given). The focus is strongly on finite (*e.g.* probability-) measures, in places at the expense of generality. Some background in elementary set-theory and analysis is required. As a comprehensive reference, we note Kingman and Taylor (1966) [52], alternatives being Dudley (1989) [29] and Billingsley (1986) [15].

A.1 Sets and sigma-algebras

Rough setup: set operations, monotony of sequences of subsets, set-limits, sigma-algebra's, measurable spaces, set-functions, product spaces.

Definition A.1.1. *A measurable space $(\Omega, \mathscr{F})$ consists of a set Ω and a σ-algebra $\mathscr{F}$ of subsets of Ω.*

A.2 Measures

Rough setup: set-functions, (signed) measures, probability measures, sigma-additivity, sigma-finiteness

Theorem A.2.1. *Let $(\Omega, \mathscr{F})$ be a measurable space with measure $\mu : \mathscr{F} \to [0, \infty]$. Then,*

(i) for any monotone decreasing sequence $(F_n)_{n\geq 1}$ in $\mathscr{F}$ such that $\mu(F_n) < \infty$ for some n,

$$\lim_{n\to\infty} \mu(F_n) = \mu\Big(\bigcap_{n=1}^{\infty} F_n\Big), \tag{A.1}$$

(ii) for any monotone increasing sequence $(G_n)_{n\geq 1}$ *in* $\mathscr{F}$*,*

$$\lim_{n\to\infty} \mu(G_n) = \mu\Big(\bigcup_{n=1}^{\infty} G_n\Big), \tag{A.2}$$

Theorem A.2.1) is sometimes referred to as the continuity theorem for measures, because if we view $\cap_n F_n$ as the monotone limit $\lim F_n$, (A.1) can be read as $\lim_n \mu(F_n) = \mu(\lim_n F_n)$, expressing continuity from below. Similarly, (A.2) expresses continuity from above. Note that theorem A.2.1 does *not* guarantee continuity for arbitrary sequences in $\mathscr{F}$. It should also be noted that theorem A.2.1) is presented here in simplified form: the full theorem states that continuity from below is equivalent to σ-additivity of μ (for a more comprehensive formulation and a proof of theorem A.2.1, see [52], theorem 3.2).

Example A.2.1. *Let* Ω *be a discrete set and let* $\mathscr{F}$ *be the powerset* 2^Ω *of* Ω*, i.e.* $\mathscr{F}$ *is the collection of* all *subsets of* Ω*. The counting measure* $n : \mathscr{F} \to [0,\infty]$ *on* $(\Omega, \mathscr{F})$ *is defined simply to count the number* $n(F)$ *of points in* $F \subset \Omega$*. If* Ω *contains a finite number of points,* n *is a finite measure; if* Ω *contains a (countably) infinite number of points,* n *is* σ*-finite. The counting measure is* σ*-additive.*

Example A.2.2. *We consider* $\mathbb{R}$ *with any* σ*-algebra* $\mathscr{F}$*, let* $x \in \mathbb{R}$ *be given and define the measure* $\delta_x : \mathscr{F} \to [0,1]$ *by*

$$\delta_x(A) = 1\{x \in A\},$$

for any $A \in \mathscr{F}$*. The probability measure* δ_x *is called the Dirac measure (or delta measure, or atomic measure) degenerate at* x *and it concentrates all its mass in the point* x*. Clearly,* δ_x *is finite and* σ*-additive. Convex combinations of Dirac measures, i.e. measures of the form*

$$P = \sum_{j=1}^{m} \alpha_j \delta_{x_j},$$

for some $m \geq 1$ *with* $\alpha_1, \ldots, \alpha_m$ *such that* $\alpha_j \geq 0$ *and* $\sum_{j=1}^m \alpha_j = 1$*, can be used as a statistical model for an observation* X *that take values in a discrete (but unknown) subset* $\{x_1, \ldots, x_m\}$ *of* $\mathbb{R}$*. The resulting model (which we denote* $D(\mathbb{R}, \mathscr{B})$ *for reference) is not dominated.*

Often, one has a sequence of events $(A_n)_{n\geq 1}$ and one is interested in the probability of a limiting event A, for example the event that A_n occurs infinitely often. The following three lemmas pertain to this situation.

Lemma A.2.1. *(First Borel-Cantelli lemma)*
Let $(\Omega, \mathscr{F}, P)$ *be a probability space and let* $(A_n)_{n\geq 1} \subset \mathscr{F}$ *be given and denote* $A = \limsup A_n$*. If*

$$\sum_{n\geq 1} P(A_n) < \infty,$$

then $P(A) = 0$*.*

In the above lemma, the sequence $(A_n)_{n\geq 1}$ is general. To draw the converse conclusion, the sequence needs to exist of *independent* events.

Lemma A.2.2. *(Second Borel-Cantelli lemma)*
Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $(A_n)_{n\geq 1} \subset \mathscr{F}$ be independent and denote $A = \limsup A_n$. If

$$\sum_{n\geq 1} P(A_n) = \infty,$$

then $P(A) = 1$.

Together, the Borel-Cantelli lemmas assert that for a sequence of independent events $(A_n)_{n\geq 1}$, $P(A)$ equals zero or one, according as $\sum_n P(A_n)$ converges or diverges. As such, this corollary is known as a *zero-one law*, of which there are many in probability theory.

exchangability, De Finetti's theorem

Theorem A.2.2. *(De Finetti's theorem) State De Finetti's theorem.*

Theorem A.2.3. *(Ulam's theorem) State Ulam's theorem.*

Definition A.2.1. *Let $(\mathscr{Y}, \mathscr{B})$ be a measurable space. Given a set-function $\mu : \mathscr{B} \to [0, \infty]$, the total variation total-variation norm of μ is defined:*

$$\|\mu\|_{TV} = \sup_{B\in\mathscr{B}} |\mu(B)|. \tag{A.3}$$

Lemma A.2.3. *Let $(\mathscr{Y}, \mathscr{B})$ be a measurable space. The collection of all signed measures on $\mathscr{Y}$ forms a linear space and total variation is a norm on this space.*

A.3 Measurability and random variables

Rough setup: measurability, monotone class theorem, simple functions, random variables, approximating sequences.

A.4 Integration

Rough setup: the definition of the integral, its basic properties, limit-theorems (Fatou, dominated convergence) and L_p-spaces.

Definition A.4.1. *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. A real-valued measurable function $f : \Omega \to \mathbb{R}$ is said to be μ-integrable if*

$$\int_O mega|f|\, d\mu < \infty. \tag{A.4}$$

Remark A.4.1. *If f is a stochastic vector taking values in $\mathbb{R}^d$, the above definition of integrability is extended naturally by imposing (A.4) on each of the component functions. This extension is more problematic in infinite-dimensional spaces. However, various generalizations can be found in an approach motivated by functional analysis (see Megginson (1998) [67] for an introduction to functional analysis): suppose that $f : \Omega \to X$ takes its values in an infinite-dimensional space X. If $(X, \|\cdot\|)$ is a normed space, one can impose that*

$$int_\Omega \|f\| \, d\mu < \infty,$$

but this definition may be too strong, in the sense that too few functions f satisfy it. If X has a dual X^*, one may impose that for all $x^* \in X^*$,

$$\int_\Omega x^*(f) \, d\mu < \infty,$$

which is often a weaker condition than the one in the previous display. In case X is itself (a subset of) the dual of a space X', then $X' \subset X^*$ and we may impose that for all $x \in X'$,

$$\int_\Omega f(x) \, d\mu < \infty$$

which is weaker than both previous displays.

Example A.4.1. *Our primary interest here is in Bayesian statistics, where the prior and posterior can be measures on a non-parametric model, giving rise to a situation like that in remark A.4.1. Frequently, observations will lie in $\mathbb{R}^n$ and we consider the space of all bounded, measurable functions on $\mathbb{R}^n$, endowed with the supermum-norm. This space forms a Banach space X' and $\mathscr{P}$ is a subset of the unit-sphere of the dual X'^*, since $X \to \mathbb{R} : f \mapsto Pf$ satisfies $|Pf| \leq \|f\|$, for all $f \in X$. Arguably, P should be called integrable with respect to a measure Ξ on $\mathscr{P}$, if*

$$\left| \int_{\mathscr{P}} Pf \; d\Xi(P) \right| < \infty.$$

for all $f \in X$. Then, "suitable integrability" is not an issue in the definition of the posterior mean (2.2.1), since $P|f| \leq \sup_{\mathbb{R}^n} |f| = \|f\| < \infty$ for all $f \in X$ and the posterior is a probability measure.

Theorem A.4.1. *(Fubini's theorem) State Fubini's theorem.*

Theorem A.4.2. *(Radon-Nikodym theorem) Let $(\Omega, \mathscr{F})$ be a measurable space and let $\mu, \nu : \mathscr{F} \to [0, \infty]$ be two σ-finite measures on $(\Omega, \mathscr{F})$. There exists a unique decomposition*

$$\mu = \mu_\parallel + \mu_\perp,$$

such that $\nu_parallel \ll \nu$ and $\mu_\perp$ and ν are mutually singular. Furthermore, there exists a finite-valued, $\mathscr{F}$-measurable function $f : \Omega \to \mathbb{R}$ such that for all $F \in \mathscr{F}$,

$$\mu_\parallel(F) = \int_F f \, d\nu. \tag{A.5}$$

The function f is ν-almost-everywhere unique.

The function $f : \Omega \to \mathbb{R}$ in the above theorem is called the *Radon-Nikodym derivative* of μ with respect to ν. If μ is a probability distribution, then f is called the (probability) density for μ with respect to ν. The assertion that f is "ν-almost-everywhere unique" means that if there exists a measurable function $g : \Omega \to \mathbb{R}$ such that (A.5) holds with g replacing f, then $f = g$, $(\nu - a.e.)$, *i.e.* f and g may differ only on a set of ν-measure equal to zero. Through a construction involving increasing sequences of simple functions, we see that the Radon-Nikodym theorem has the following implication.

Corollary A.4.1. *Assume that the conditions for the Radon-Nikodym theorem are satisfied. Let $X : \Omega \to [0, \infty]$ be measurable and μ-integrable. Then the product Xf is ν-integrable and*

$$\int X \, d\mu = \int Xf \, d\nu.$$

Remark A.4.2. *Integrability is not a necessary condition here, but the statement of the corollary becomes rather less transparent if we indulge in generalization.*

A.5 Existence of stochastic processes

A stochastic processes have the following broad definition.

Definition A.5.1. *Let $(\Omega, \mathscr{F}, P)$ be a probability space, let T be an arbitrary set. A collection of $\mathscr{F}$-measurable random variables $\{ X_t : \Omega \to \mathbb{R} \, : \, t \in T\}$ is called a stochastic process indexed by T.*

The problem with the above definition is the requirement that there exists an underlying probability space: typically, one approaches a problem that requires the use of stochastic processes by proposing a collection of random quantities $\{X_t : t \in T\}$. The guarantee that an underlying probability space $(\Omega, \mathscr{F}, P)$ exists on which all X_t can be realised as random variables is then lacking so that we have not defined the stochastic process properly yet. Kolmogorov's existence theorem provides an explicit construction of $(\Omega, \mathscr{F}, P)$. Clearly, if the X_t take their values in a measurable space space $(\mathscr{X}, \mathscr{B})$, the obvious choice for Ω is the collection $\mathscr{X}^T$ in which the process takes its values. The question remains how to characterize P and its domain $\mathscr{F}$. Kolmogorov's solution here is to assume that for any *finite* subset $S = \{t_1, \ldots, t_k\} \subset T$, the distribution $P_{t_1 \ldots t_k}$ of the k-dimensional stochastic vector $(X_{t_1}, \ldots, X_{t_k})$ is given. Since the distributions $P_{t_1 \ldots t_k}$ are as yet unrelated and given for *all* finite subsets of T, consistency requirements are implicit if they are to serve as marginals to the probability distribution P: if two finite subsets $S_1, S_2 \subset T$ satisfy $S_1 \subset S_2$, then the distribution of $\{X_t : t \in S_1\}$ should be marginal to that of $\{X_t : t \in S_2\}$. Similarly, permutation of the components of the stochastic vector in the above display should be reflected in the respective distributions as well. The requirements for consistency are formulated in two requirements called Kolmogorov's *consistency conditions*:

(K1) Let $k \geq 1$ and $\{t_1, \ldots, t_{k+1}\} \subset T$ be given. For any $C \in \sigma(\mathscr{B}^k)$,

$$P_{t_1 \ldots t_k}(C) = P_{t_1 \ldots t_{k+1}}(C \times \mathscr{X}),$$

(K2) Let$k \geq 1$, $\{t_1, \ldots, t_k\} \subset T$ and a permutation π of k elements be given. For any $A_1, \ldots, A_k \in \mathscr{B}$,

$$P_{t_{\pi(1)} \ldots t_{\pi(k)}}(A_1 \times \ldots \times A_k) = P_{t_1 \ldots t_k}(A_{\pi^{-1}(1)} \times \ldots \times A_{\pi^{-1}(k)}).$$

Theorem A.5.1. *(Kolmogorov's existence theorem)*
Let a collection of random quantities $\{X_t : t \in T\}$ be given. Suppose that for any $k \geq 1$ and all $t_1, \ldots, t_k \in T$, the finite-dimensional marginal distributions

$$(X_{t_1}, \ldots, X_{t_k}) \sim P_{t_1 \ldots t_k}, \tag{A.6}$$

are defined and satisfy conditions (K1) and (K2). Then there exists a probability space $(\Omega, \mathscr{F}, P)$ and a stochastic process $\{ X_t : \Omega \to \mathscr{X} : t \in T \}$ such that all distributions of the form (A.6) are marginal to P.

Kolmogorov's approach to the definition and characterization of stochastic processes in terms of finite-dimensional marginals turns out to be of great practical value: it allows one to restrict attention to finite-dimensional marginal distributions when characterising the process. The drawback of the construction becomes apparent only upon closer inspection of the σ-algebra $\mathscr{F}$: $\mathscr{F}$ is the σ-algebra generated by the cylinder sets, which implies that measurability of events restricing an uncountable number of X_t's simultaneously can not be guaranteed! For instance, if $T = [0, \infty)$ and $\mathscr{X} = \mathbb{R}$, the probability that sample-paths of the process are coninuous,

$$P\big(\mathbb{R} \to \mathbb{R} : t \mapsto X_t \,\text{iscontinuous} \big),$$

may be ill-defined because it involves an uncountable number of t's. This is the ever-recurring trade-off between generality and strength of a mathematical result: Kolmogorov's existence theorem always works but it does not give rise to a comfortably 'large' domain for the resulting $P : \mathscr{F} \to [0, 1]$.

A.6 Conditional distributions

In this section, we consider conditioning of probability measures. In first instance, we consider straightforward conditioning on events and illustrate Bayes' rule, but we also cover conditioning on σ-algebras and random variables, to arrive at the posterior distribution and Bayes' rule for densities.

Definition A.6.1. *Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $B \in \mathscr{F}$ be such that $P(B) > 0$. For any $A \in \mathscr{F}$, the conditional probability of the event A given event B is defined:*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{A.7}$$

Conditional probability given B describes a set-function on $\mathscr{F}$ and one easily checks that this set-function is a measure. The conditional probability measure $P(\,\cdot\,|B) : \mathscr{F} \to [0,1]$ can be viewed as the restriction of P to $\mathscr{F}$-measurable subsets of B, normalized to be a probability measure. Definition (A.7) gives rise to a relation between $P(A|B)$ and $P(B|A)$ (in case both $P(A) > 0$ and $P(B) > 0$, of course), which is called Bayes' Rule.

Lemma A.6.1. (Bayes' Rule)
Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $A, B \in \mathscr{F}$ be such that $P(A) > 0$, $P(B) > 0$. Then

$$P(A|B)\,P(B) = P(B|A)\,P(A).$$

However, being able to condition on events B of non-zero probability only is too restrictive. Furthermore, B above is a definite event; it is desirable also to be able to discuss probabilities conditional on events that have not been measured yet, *i.e.* to condition on a σ-algebra.

Definition A.6.2. *Let $(\Omega, \mathscr{F}, P)$ be a probability space, let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{F}$ and let X be a P-integrable random variable. The conditional expectation of X given $\mathscr{C}$, denoted $E[X|\mathscr{C}]$, is a $\mathscr{C}$-measurable random variable such that for all $C \in \mathscr{C}$,*

$$\int_C X\,dP = \int_C E[X|\mathscr{C}]\,dP.$$

The condition that X be P-integrable is sufficient for the existence of $E[X|\mathscr{C}]$; $E[X|\mathscr{C}]$ is unique P-almost-surely (see theorem 10.1.1 in Dudley (1989)). Often, the σ-algebra $\mathscr{C}$ is the σ-algebra $\sigma(Z)$ generated by another random variable Z. In that case we denote the conditional expectation by $E[X|Z]$. Note that conditional expectations are random themselves: realisation occurs only when we impose $Z = z$.

Definition A.6.3. *Let $(\mathscr{Y}, \mathscr{B})$ be a measurable space, let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{F}$. Furthermore, let $Y : \Omega \to \mathscr{Y}$ be a random variable taking values in $\mathscr{Y}$. The conditional distribution of Y given $\mathscr{C}$ is P-almost-surely defined as follows:*

$$P_{Y|\mathscr{C}}(A, \omega) = E[1_A|\mathscr{C}](\omega). \tag{A.8}$$

Although seemingly innocuous, the fact that conditional expectations are defined only P-almost-surely poses a rather subtle problem: for every $A \in \mathscr{B}$ there exists an A-dependent null-set on which $P_{Y|\mathscr{C}}(A, \cdot)$ is not defined. This is not a problem if we are interested only in A (or in a countable number of sets). But usually, we wish to view $P_{Y|\mathscr{C}}$ as a probability measure, that is to say, it must be well-defined as a *map* on the σ-algebra $\mathscr{B}$ almost-surely. Since most σ-algebras are uncountable, there is no guarantee that the corresponding union of exceptional null-sets has measure zero as well. This means that definition (A.8) is not sufficient for our purposes: the property that the conditional distribution is well-defined as a map is called *regularity*.

Definition A.6.4. *Under the conditions of definition A.6.3, we say that the conditional distribution $\Pi_{Y|\mathscr{C}}$ is regular, if there exists a set $E \in \mathscr{F}$ such that $P(E) = 0$ and for all $\omega \in \Omega \setminus E$, $\Pi_{Y|\mathscr{C}}(\,\cdot\,, \omega)$ satisfies A.8 for all $A \in \mathscr{B}$.*

Definition A.6.5. *A topological space $(S, \mathscr{T})$ is said to be a Polish space if $\mathscr{T}$ is metrizable with metric d and (S, d) is complete and separable.*

Polish spaces appear in many subjects in probability theory, most notably in a theorem that guarantees that conditional distributions are regular.

Theorem A.6.1. *(regular conditional distributions) Let $\mathscr{Y}$ be a Polish space and denote its Borel σ-algebra by $\mathscr{B}$. Furthermore let $(\Omega, \mathscr{F}, P)$ be a probability space and $Y : \Omega \to \mathscr{Y}$ a random variable taking values in $\mathscr{Y}$. Let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{F}$. Then a conditional distribution [MORE MORE]*

Proof For a proof of this theorem, the reader is referred to Dudley (1989) [29], theorem 10.2.2). □

In Bayesian context we can be more specific regarding the sub-σ-algebra $\mathscr{C}$: since $\Omega = \mathscr{X} \times \Theta$ (so that $\omega = (x, \theta)$) and we condition on θ, we choose $\mathscr{C} = \{\mathscr{X} \times G : G \in \mathscr{G}\}$.

Note also that due to the special choice for $\mathscr{C}$, $\mathscr{C}$-measurability implies that $\Pi_{Y|\mathscr{C}}(\,\cdot\,, (y, \theta))$ depends on θ alone. Hence we denote it $\Pi_{Y|\vartheta} : \mathscr{B} \times \Theta \to [0, 1]$.

Lemma A.6.2. (Bayes' Rule for densities)
State Bayes' rule for densities.

A.7 Convergence in spaces of probability measures

Let $M(\mathbb{R}, \mathscr{B})$ denote the space of all probability measures on $\mathbb{R}$ with Borel σ-algebra $\mathscr{B}$.

Definition A.7.1. *(topology of weak convergence)*
Let $(Q_n)_{n\geq 1}$ and Q in $M(\mathbb{R}, scrB)$ be given. Denote the set of points in $\mathbb{R}$ where $\mathbb{R} \to [0, 1] : t \mapsto Q(-\infty, t]$ is continuous by C. We say that Q_n converges weakly to Q if, for all $t \in C$, $Q_n(-\infty, t] \to Q(-\infty, t]$.

Weak convergence has serveral equivalent definitions. The following lemma, known as the Portmanteau lemma (from the French word for coat-rack),

Lemma A.7.1. *Let $(Q_n)_{n\geq 1}$ and Q in $M(\mathbb{R}, scrB)$ be given. The following are equivalent:*

(i) Q_n converges weakly to Q.

(ii) For every bounded, continuous $f : \mathbb{R} \to \mathbb{R}$, $Q_n f \to Qf$.

(iii) For every bounded, Lipschitz $g : \mathbb{R} \to \mathbb{R}$, $Q_n g \to Qg$.

(iv) For all non-negative, continuous $h : \mathbb{R} \to \mathbb{R}$, $\liminf_{n\to\infty} Q_n f \geq Qf$.

(v) For every open set $F \subset \mathbb{R}$, $\liminf_{n\to\infty} Q_n(F) \geq Q(F)$.

(vi) For every closed set $G \subset \mathbb{R}$, $\limsup_{n\to\infty} Q_n(G) \leq Q(G)$.

(vii) For every Borel set B *such that* $Q(\delta B) = 0$, $Q_n(B) \to Q(B)$.

In *(vii)* above, δB denotes the boundary of B, which is defined as the closure of B minus the interior of B.

Lemma A.7.2. *When endowed with the topology of weak convergence, the space* $M(\mathbb{R}, \mathscr{B})$ *is Polish, i.e. complete, separable and metric.*

Definition A.7.2. *(topology of pointwise convergence)*
Let $(Q_n)_{n\geq 1}$ *and* Q *in* $M(\mathbb{R}, scrB)$ *be given. We say that* Q_n *converges pointwise to* Q *if, for all* $B \in \mathscr{B}$, $Q_n(B) \to Q(B)$.

Definition A.7.3. *(topology of total variation)*
Let $(Q_n)_{n\geq 1}$ *and* Q *in* $M(\mathbb{R}, scrB)$ *be given. We say that* Q_n *converges in total variation to* Q *if,*

$$\sup_{B\in\mathscr{B}} \big|Q_n(B) - Q(B)\big| \to 0.$$

Lemma A.7.3. *When endowed with the topology of total variation, the space* $M(\mathbb{R}, \mathscr{B})$ *is a Polish subspace of the Banach space of all signed measures on* $(\mathbb{R}, \mathscr{B})$.

Bibliography

[1] M. Alpert and H. Raiffa, *A progress report on the training of probability assessors,* In *Judgement under uncertainty: heuristics and biases*, eds. D. Kahneman, P. Slovic and A. Tversky, Cambridge University Press, Cambridge (1982).

[2] S. Amari, *Differential-geometrical methods in statistics,* Lecture Notes in Statistics No. 28, Springer Verlag, Berlin (1990).

[3] M. Bayarri and J. Berger, *The interplay of Bayesian and frequentist analysis,* Preprint (2004).

[4] T. Bayes, *An essay towards solving a problem in the doctrine of chances,* Phil. Trans. Roy. Soc. **53** (1763), 370–418.

[5] A. Barron, L. Birgé and P. Massart, *Risk bounds for model selection via penalization*, Probability Theory and Related Fields **113** (1999), pp. 301–413.

[6] S. Bernstein, *Theory of probability,* (in Russian), Moskow (1917).

[7] A. Barron, M. Schervish and L. Wasserman, *The consistency of posterior distributions in nonparametric problems,* Ann. Statist. **27** (1999), 536–561.

[8] J. Berger, *Statistical decision theory and Bayesian analysis,* Springer, New York (1985).

[9] J. Berger and J. Bernardo, *On the development of reference priors,* Bayesian Statististics **4** (1992), 35–60.

[10] R. Berk, *Consistency of a posteriori,* Ann. Math. Statist. **41** (1970), 894–906.

[11] R. Berk and I. Savage, *Dirichlet processes produce discrete measures: an elementary proof,* Contributions to statistics, Reidel, Dordrecht (1979), 25–31.

[12] J. Bernardo, *Reference posterior distributions for Bayesian inference,* J. Roy. Statist. Soc. **B41** (1979), 113–147.

[13] J. Bernardo and A. Smith, *Bayesian theory,* John Wiley & Sons, Chichester (1993).

[14] P. Bickel and J. Yahav, *Some contributions to the asymptotic theory of Bayes solutions,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **11** (1969), 257–276.

[15] P. Billingsley, *Probability and Measure, 2nd edition,* John Wiley & Sons, Chichester (1986).

[16] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65** (1983), 181–238.

[17] L. BIRGÉ, *Sur un théorème de minimax et son application aux tests,* Probability and Mathematical Statistics **3** (1984), 259–282.

[18] L. BIRGÉ and P. MASSART, *From model selection to adaptive estimation,* Festschrift for Lucien Le Cam, Springer, New York (1997), 55–87.

[19] L. BIRGÉ and P. MASSART, *Gaussian model selection*, J. Eur. Math. Soc. **3** (2001), 203–268.

[20] C. BISHOP, *Pattern Recognition and Machine Learning,* Springer, New York (2006).

[21] D. BLACKWELL and L. DUBINS, *Merging of opinions with increasing information,* Ann. Math. Statist. **33** (1962), 882–886.

[22] L. BOLTZMANN, *Vorlesungen über Gastheorie,* (2 Volumes), Leipzig (1895, 1898).

[23] H. CRAMÉR, *Mathematical methods of statistics,* Princeton University Press, Princeton (1946).

[24] A. DAWID, *On the limiting normality of posterior distribution,* Proc. Canad. Phil. Soc. **B67** (1970), 625-633.

[25] P. DIACONIS and D. FREEDMAN, *On the consistency of Bayes estimates,* Ann. Statist. **14** (1986), 1–26.

[26] P. DIACONIS and D. FREEDMAN, *On inconsistent Bayes estimates of location,* Ann. Statist. **14** (1986), 68–87.

[27] P. DIACONIS and D. FREEDMAN, *Consistency of Bayes estimates for nonparametric regression: Normal theory,* Bernoulli, **4** (1998), 411-444.

[28] J. DOOB, *Applications of the theory of martingales,* Le calcul des Probabilités et ses Applications, Colloques Internationales du CNRS, Paris (1948), 22–28.

[29] R. DUDLEY, *Real analysis and probability,* Wadsworth & Brooks-Cole, Belmont (1989).

[30] A. DVORETZKY, J. KIEFER, and J. WOLFOWITZ, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator,* Ann. Math. Statist. **27** (1956), 642–669.

[31] B. EFRON, *Defining curvature on a statistical model,* Ann. Statist. **3** (1975), 1189–1242.

[32] B. EFRON and R. Tibshirani, An introduction to the Bootstrap, Chapman and Hall, London (1993).

[33] M. ESCOBAR and M. WEST, *Bayesian density estimation and inference with mixtures,* Journal of the American Statistical Association **90** (1995), 577–588.

[34] T. FERGUSON, *A Bayesian analysis of some non-parametric problems,* Ann. Statist. **1** (1973), 209–230.

[35] T. FERGUSON, *Prior distribution on the spaces of probability measures,* Ann. Statist. **2** (1974), 615–629.

[36] D. FREEDMAN, *On the asymptotic behavior of Bayes estimates in the discrete case I,* Ann. Math. Statist. **34** (1963), 1386–1403.

[37] D. FREEDMAN, *On the asymptotic behavior of Bayes estimates in the discrete case II,* Ann. Math. Statist. **36** (1965), 454–456.

[38] D. Freedman, *On the Bernstein-von Mises theorem with infinite dimensional parameters,* Ann. Statist. **27** (1999), 1119–1140.

[39] S. van de Geer, Empirical Processes in M-Estimation. Cambridge University Press, Cambridge (2000).

[40] S. Ghosal, J. Ghosh and R. Ramamoorthi, *Non-informative priors via sieves and packing numbers,* Advances in Statistical Decision theory and Applications (eds. S. Panchapakeshan, N. Balakrishnan), Birkhäuser, Boston (1997).

[41] S. Ghosal, J. Ghosh and A. van der Vaart, *Convergence rates of posterior distributions,* Ann. Statist. **28** (2000), 500–531.

[42] J. Ghosh and R. Ramamoorthi, *Bayesian Nonparametrics,* Springer Verlag, Berlin (2003).

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

[44] T.-M. Huang, *Convergence rates for posterior distributions and adaptive estimation,* Carnegie Mellon University, preprint (accepted for publication in Ann. Statist.).

[45] I. Ibragimov and R. Has'minskii, *Statistical estimation: asymptotic theory,* Springer, New York (1981).

[46] H. Jeffreys, *An invariant form for the prior probability in estimation problems,* Proc. Roy. Soc. London **A186** (1946), 453–461.

[47] H. Jeffreys, *Theory of probability (3rd edition),* Oxford University Press, Oxford (1961).

[48] R. Kass and A. Raftery, *Bayes factors,* Journal of the American Statistical Association **90** (1995), 773–795.

[49] R. Kass and L. Wasserman, *A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,* Journal of the American Statistical Association **90** (1995), 928–934.

[50] Yongdai Kim and Jaeyong Lee, *The Bernstein-von Mises theorem of survival models,* (accepted for publication in Ann. Statist.)

[51] Yongdai Kim and Jaeyong Lee, *The Bernstein-von Mises theorem of semiparametric Bayesian models for survival data,* (accepted for publication in Ann. Statist.)

[52] J. Kingman and S. Taylor, *Introduction to measure and probability,* Cambridge University Press, Cambridge (1966).

[53] B. Kleijn and A. van der Vaart, *Misspecification in Infinite-Dimensional Bayesian Statistics,* Ann. Statist. **34** (2006), 837–877.

[54] B. Kleijn and A. van der Vaart, *The Bernstein-Von-Mises theorem under misspecification,* (submitted for publication in the Annals of Statistics).

[55] B. Kleijn and A. van der Vaart, *A Bayesian analysis of errors-in-variables regression,* (submitted for publication in the Annals of Statistics).

[56] A. Kolmogorov and V. Tikhomirov, *Epsilon-entropy and epsilon-capacity of sets in function spaces,* American Mathematical Society Translations (series 2), **17** (1961), 277–364.

[57] P. Laplace, *Mémoire sur la probabilité des causes par les évenements,* Mem. Acad. R. Sci. Presentés par Divers Savans **6** (1774), 621–656. (Translated in Statist. Sci. **1**, 359–378.)

[58] P. Laplace, *Théorie Analytique des Probabilités (3rd edition),* Courcier, Paris (1820).

[59] E. Lehmann and G. Casella, *Theory of point-estimation, (2nd ed.)* Springer, New York (1998).

[60] E. Lehmann and J. Romano, *Testing statistical hypothesis,* Pringer, New York (2005).

[61] L. Le Cam, *On some asymptotic properties of maximum-likelihood estimates and related Bayes estimates,* University of California Publications in Statistics, **1** (1953), 277–330.

[62] L. Le Cam, *On the assumptions used to prove asymptotic normality of maximum likelihood estimators,* Ann. Math. Statist. **41** (1970), 802–828.

[63] L. Le Cam, *Asymptotic methods in statistical decision theory,* Springer, New York (1986).

[64] L. Le Cam and G. Yang, *Asymptotics in Statistics: some basic concepts,* Springer, New York (1990).

[65] D. Lindley, *A measure of the information provided by an experiment,* Ann. Math. Statist. **27** (1956), 986–1005.

[66] D. Lindley and A. Smith, *Bayes estimates for the linear model,* J. Roy. Statist. Soc. **B43** (1972), 1–41.

[67] R. Megginson, *An introduction to Banach Space Theory,* Springer, New York (1998).

[68] R. Von Mises, *Wahrscheinlichkeitsrechnung,* Springer Verlag, Berlin (1931).

[69] J. Munkres, *Topology (2nd edition),* Prentice Hall, Upper Saddle River (2000).

[70] H. Raiffa, and R. Schlaifer, *Decision analysis: introductory lectures on choices under uncertainty,* Addison-Wesley, Reading (1961).

[71] C. Rao, *Information and the accuracy attainable in the estimation of statistical parameters,* Bull. Calcutta Math. Soc. **37** (1945), 81–91.

[72] C. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation,* Springer, New York (2001).

[73] B. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge (1996).

[74] L. Savage, *The subjective basis of statistical practice,* Technical report, Dept. Statistics, University of Michigan (1961).

[75] M. Schervish, *Theory of statistics,* Springer, New York (1995).

[76] L. Schwartz, *On Bayes procedures,* Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **4** (1965), 10–26.

[77] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (1978), pp. 461–464.

[78] C. Shannon, *A Mathematical Theory of Communication,* Bell System Technical Journal **27** (1948), 379–423, 623–656.

[79] X. Shen and L. Wasserman, *Rates of convergence of posterior distributions,* Ann. Statist. **29** (2001), 687–714.

[80] X. Shen, *Asymptotic normality of semiparametric and nonparametric posterior distributions,* Journal of the American Statistical Association **97** (2002), 222–235.

[81] H. Strasser, *Mathematical Theory of Statistics,* de Gruyter, Amsterdam (1985).

[82] J. Tukey, *Exploratory data analysis,* Addison-Wesley, Reading (1977).

[83] A. van der Vaart, *Asymptotic Statistics,* Cambridge University Press, Cambridge (1998).

[84] A. Walker, *On the asymptotic behaviour of posterior distributions,* J. Roy. Statist. Soc. **B31** (1969), 80–88.

[85] L. Wasserman, *Bayesian model selection and model averaging*, J. Math. Psych. **44** (2000), 92–107.

[86] Y. Yang and A. Barron, *An asymptotic property of model selection criteria*, IEEE Transactions on Information Theory **44** (1998), 95–116.

Index

Cover Illustration

The figure on the front cover originates from Bayes (1763), *An essay towards solving a problem in the doctrine of chances,* (see [4] in the bibliography), and depicts what is nowadays known as Bayes' Billiard. To demonstrate the uses of conditional probabilities and Bayes' Rule, Bayes came up with the following example: one white ball and n red balls are placed on a billiard table of length normalized to 1, at independent, uniformly distributed positions. Conditional on the distance X of the white ball to one end of the table, the probability of finding exactly k of the n red balls closer to that end, is easily seen to be:

$$P\big(\, k \;\big|\; X = x \,\big) = \frac{n!}{k!\,(n-k)!}\, x^k (1-x)^{n-k}.$$

One finds the probability that k red balls are closer than the white, by integrating with respect to the position of the white ball:

$$P(\, k \,) = \frac{1}{n+1}.$$

Application of Bayes' Rule then gives rise to a Beta-distribution $B(k+1, n-k+1)$ for the position of the white ball conditional on the number k of red balls that are closer. The density:

$$\beta_{k+1,n-k+1}(x) = \frac{(n+1)!}{k!\,(n-k)!} x^k (1-x)^{n-k},$$

for this Beta-distribution is the curve drawn at the bottom of the billiard in the illustration. (See example 2.1.2)