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:

$$d\Pi_2(\theta|X_{n+1}) = \frac{p_\theta(X_{n+1})\,d\Pi_2(\theta|X_1=x_1,\ldots,X_n=x_n)}{\int_\Theta p_\theta(X_{n+1})\,d\Pi_2(\theta|X_1=x_1,\ldots,X_n=x_n)}$$

$$= \frac{p_\theta(X_{n+1})\prod_{i=1}^{n} p_\theta(x_i) d\Pi_1(\theta)}{\int_\Theta p_\theta(X_{n+1})\prod_{j=1}^{n} p_\theta(x_j)\,d\Pi_1(\theta)} \tag{3.1}$$

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:

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

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 $C \in \mathscr{B}_2$,

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

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,

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

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,

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

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 $\theta \in \Theta$,*

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

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*

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

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. $\overset{\circ}{H} \neq \emptyset$.

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*

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

(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),

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

(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:*

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

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:*

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

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,*

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

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 2.1.4. 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,

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

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:

$$\pi(p) = \frac{\Gamma\bigl(\sum_{i=1}^{k} \alpha_i\bigr)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \ldots p_{k-1}^{\alpha_{k-1} - 1} \Bigl(1 - \sum_{i=1}^{k-1} p_i\Bigr)^{\alpha_k - 1}$$

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:

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

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,

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

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:

$$\Pi\bigl(P(\emptyset)=0\bigr)=1.$$

(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:

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

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 $A\cup B=C$,

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

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:

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

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:

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

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 $M(\mathscr{X},\mathscr{B})$,*

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

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,

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

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 $B_1 = B$, $B_2 = \mathbb{R} \backslash B$. 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

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

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 $C \subset M(\mathbb{R}, \mathscr{B})$,*

$$\Pi\bigl(P \in C \bigm| X_1, \ldots, X_n\bigr) = D_{\alpha + n\mathbb{P}_n}(C),$$

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:

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

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

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

d. $X|\theta \sim \mathrm{Bin}(n;\theta)$ *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* $\lambda > 0$,

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

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* $\mathrm{Poisson}(\theta)$ *with unknown* $\theta > 0$*. As a family of possible priors for the Bayesian analysis of this data, consider exponential distributions* $\theta \sim \Pi_\lambda = \mathrm{Exp}(\lambda)$*, 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.