Lecture Notes on Bayesian Estimation and Classification

Mário A. T. Figueiredo,
Instituto de Telecomunicações, and
Instituto Superior Técnico
1049-001 Lisboa
Portugal
(Email address: mario.figueiredo@lx.it.pt)

October 2004

Contents

1 Introduction to Bayesian Decision Theory

1.1 Introduction

Statistical decision theory deals with situations where decisions have to be made under a state of uncertainty, and its goal is to provide a rational framework for dealing with such situations. The Bayesian approach, the main theme of this chapter, is a particular way of formulating and dealing with statistical decision problems. More specifically, it offers a method of formalizing *a priori* beliefs and of combining them with the available observations, with the goal of allowing a rational (formal) derivation of optimal (in some sense) decision criteria.

As can be inferred from the previous paragraph, this book's introduction to Bayesian theory adopts a decision theoretic perspective. An important reason behind this choice is that inference problems (e.g., how to estimate an unknown quantity) can be naturally viewed as special cases of decision problems; this way, all the conceptual tools of Bayesian decision theory (*a priori* information and loss functions) are incorporated into inference criteria.

The literature on Bayesian theory is vast and anyone interested in further reading is referred to the many excellent textbooks available on the subject; at the risk of unfairly but unintentionally leaving out important works, we mention here some books that somehow influenced the authors: Berger [8], Bernardo and Smith [14], Gelman, Carlin, Stern, and Rubin [46], Lee [69], and Robert [93]; a not recent, but still useful and insightful review is the one by Lindley [72]; a short and easily readable summary of the main arguments in favor of the Bayesian perspective can be found in a paper by Berger whose title, *"Bayesian Salesmanship,"* clearly reveals the nature of its contents [9]. Also highly recommended by its conceptual depth and the breadth of its coverage is Jaynes' (still unfinished but partially available) book [58]. Recent advances are reported in workshops and conferences (special emphasis should be be given to [13], [11], and [12]) and in several scientific journals (for example, the *Journal of the American Statistical Association* and the *Journal of the Royal Statistical Society*).

Bayesian frameworks have been used to deal with a wide variety of problems in many scientific and engineering areas. Whenever a quantity is to be inferred, or some conclusion is to be drawn, from observed data, Bayesian principles and tools can be used. Examples, and this is by no means an exhaustive list of mutually exclusive areas, include: statistics, signal processing, speech analysis, image processing, computer vision, astronomy, telecommunications, neural networks, pattern recognition, machine learning, artificial intelligence, psychology, sociology, medical decision making, econometrics, and biostatistics. Focusing more closely on the topic of interest to this book, we mention that, in addition to playing a major role in the design of machine (computer) vision techniques, the Bayesian framework has also been found very useful in understanding natural (e.g., human) perception [66]; this fact is a strong testimony in favor of the Bayesian paradigm.

Finally, it is worth pointing out that the Bayesian perspective is not only important at a practical application level, but also at deeper conceptual levels, touching foundational and philosophical aspects of scientific inference, as the title of Rozenkrantz's book [95] so clearly shows: *"Inference, Method, and Decision: Towards a Bayesian Philosophy of Science"*. On this issue, the book by Jaynes is a fundamental more recent reference [58].

1.2 Statistical Decision Theory

1.2.1 Basic Elements

The fundamental conceptual elements supporting the (formal) theory of statistical decision making are the following:

- **Formalization of the underlying unknown reality**. This is done by considering that all that is unknown but relevant for the decision maker, the so-called *state of nature*, can be represented by an entity s taking values on a state space $\mathcal{S}$. Often, this will be a single unknown numerical quantity (a parameter), or an ordered set of numerical parameters (a vector). In other problems, the elements of $\mathcal{S}$ may not be of numerical nature. Throughout most of this chapter, we will implicitly assume that s is a single quantity.

- **Formal model of the observations**. The observations, based on which decisions are to be made, are possibly random and depend on the state of nature s. In formal probabilistic terms, this dependence is expressed by assuming that the observations are a sample $\mathbf{x}$ of a random variable (or process, or vector, or field) $\mathbf{X}$, taking values on a *sample space* $\mathcal{X}$, whose probability (density or mass) function, for $\mathbf{x} \in \mathcal{X}$, is conditioned on the true state of nature s, i.e., we write $f_{\mathbf{X}}(\mathbf{x}|s)$. This probability function appears in the literature under several different names: *class-conditional* probability function (usually in pattern recognition problems, where the observations $\mathbf{x}$ are called *features*); *observation model* (typically in signal/image processing applications, where $\mathbf{x}$ is usually referred to as the *observed signal* or *observed image*); *parametric statistical model*, or *likelihood function* (terms from the statistics literature but also adopted by other communities).

- **Formal decision rules**. These are the goal of decision theory in the following sense: based on the observations, a decision rule has to choose an action amongst a set $\mathcal{A}$ of allowed *decisions* or *actions*. Formally, a decision rule is a function[1] $\delta(\mathbf{x})$ from $\mathcal{X}$ into $\mathcal{A}$, specifying how actions/decisions are chosen, given observation(s) $\mathbf{x}$. A set or class $\mathcal{D}$ of allowed decision rules may be specified.

- **Quantification of the consequences of the decisions.** This is formally expressed via a *loss* function $L(s,a) : \mathcal{S} \times \mathcal{A} \longrightarrow I\!R$, specifying the *"cost"* that is incurred when the true state of nature is s and the chosen decision is a. It is usually required that $L(s,a) \geq L_{\min} > -\infty$, often (but not necessarily) with $L_{\min} = 0$. Although $L(s,a)$ is required to be a real valued function, its range does not necessarily have to be $I\!R$; it can be some subset of $I\!R$, with typical examples being $I\!R_0^+$ and $\{0,1\}$. Sometimes, the consequences are viewed optimistically (for example, in the economics and business literature) and, rather than losses, one talks about an *utility* function $U(s,a) : \mathcal{S} \times \mathcal{A} \longrightarrow I\!R$, specifying the *"gain"* that is obtained when the state of nature is s, and a is the chosen action. Writing $L(s,a) = -U(s,a)$ makes it clear that these are two equivalent concepts.

A statistical decision problem is then formalized by specifying this set of elements $\{\mathcal{S}, \mathcal{A}, \mathcal{X}, L(s,a), \mathcal{D}, f_{\mathbf{X}}(\mathbf{x}|s)\}$. It will be considered solved when a decision rule $\delta(\mathbf{x})$ (from $\mathcal{D}$, the set of allowed rules) is chosen such that it achieves some sort of optimality criterion (associated with the loss function). Below, we will look more in detail at how this is done, both under the (so-called) *classical* (or frequentist) and Bayesian frameworks.

[1] This notation should not be confused with the Dirac delta function.

Parameter estimation problems (also called *point estimation* problems), that is, problems in which some unknown scalar *quantity* (real valued) is to be estimated, can be viewed from a statistical decision perspective: simply let the unknown quantity be the state of nature $s \in \mathcal{S} \subseteq I\!R$; take $\mathcal{A} = \mathcal{S}$, meaning that the decision rule will output estimates (guesses) of the true s; design a loss function $L(s, a)$ expressing how much wrong estimates are to be penalized (usually verifying $L(s, s) = 0$, and $L(s, a) > 0$ for any $a \neq s$). With this setup, the decision rule will output estimates, denoted $\widehat{s} = \delta(\mathbf{x})$, of the state of nature s; This approach can, of course, be extended to multidimensional states of nature (when s is a set of unknowns). This class of problems then contains all types of signal and image estimation problems, including restoration and reconstruction.

In addition to estimation problems, many other problems in image analysis and pattern recognition can naturally be cast into a statistical decision framework. For example, pattern classification is clearly a decision making problem, where $\mathcal{S} = \mathcal{A}$ is the set of possible classes; here, $\mathbf{x} \in \mathcal{X}$ is the observed *feature vector* which is input to the decision rule $\delta(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{A}$, and $f_{\mathbf{X}}(\mathbf{x}|s)$ is the *class conditional* probability function. For example, a system may have to decide what is the dominant type of vegetation in some satellite image, denoted $\mathbf{x}$, and then, as an illustration, $\mathcal{S} = \mathcal{A} = \{\text{pine}, \text{eucalyptus}, \text{oak}\}$. Notice here, the non-numerical, often called *categorical*, nature of the set $\mathcal{S} = \mathcal{A}$.

Signal detection, so widely studied in the statistical communications and signal processing literature, can also be interpreted as a classification problem; see classical references such as [76], [77], [96], [104], and [109], or more recent accounts in [53], [64], [68], [84], or [98]. The goal here is to design *receivers* which are able to optimally decide which of a possible set of symbols (for example, 0 or 1, in binary digital communication) was sent by an emitter, from a possibly noisy and somewhat corrupted received signal. Here again, $\mathcal{S} = \mathcal{A}$ is the set of possible symbols (e.g., $\mathcal{S} = \mathcal{A} = \{\text{"bit 0"}, \text{"bit 1"}\}$).

Both classification and estimation scenarios, i.e., those where the goal is to "guess" the state of nature (thus $\mathcal{A} = \mathcal{S}$), may be commonly referred to as *inference* problems (although this is a non-standard use if the term *inference*). The distinguishing feature is the discrete (classification) or continuous (estimation) nature of sets $\mathcal{S}$ and $\mathcal{A}$. Nevertheless, the naming convention is not rigid; e.g., many situations where $\mathcal{S} = \mathcal{A}$ is a discrete set with a large number of values are referred to as estimation problems.

1.2.2 Frequentist Risk Function and Decision Rules

In the *frequentist* perspective on decision problems, it is assumed that the (unknown) state of nature s is always the same (it is usually said that s is a deterministic, albeit unknown, parameter) and that the possible observations are generated according to the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$. A decision rule is then evaluated by how well it is expected to perform when applied repeatedly; this is formalized by computing, for each of loss function $L(s,a)$ and decision rule $\delta(\cdot)$, the *frequentist risk function*, naturally defined as the average loss, given the state of nature s,

$$R(s,\delta(\cdot)) = R_\delta(s) = E_{\mathbf{X}}\left[L(s,\delta(\mathbf{x}))|s\right], \tag{1.1}$$

where

$$E_{\mathbf{X}}\left[L(s,\delta(\mathbf{x}))|s\right] = \begin{cases} \displaystyle\int_{\mathcal{X}} L(s,\delta(\mathbf{x}))\, f_{\mathbf{X}}(\mathbf{x}|s)\, d\mathbf{x} & \Leftarrow \quad \mathcal{X} \text{ is continuous} \\ \displaystyle\sum_{\mathbf{x}_i \in \mathcal{X}} L(s,\delta(\mathbf{x}_i))\, f_{\mathbf{X}}(\mathbf{x}_i|s) & \Leftarrow \quad \mathcal{X} \text{ is discrete;} \end{cases} \tag{1.2}$$

(the notation $E_{\mathbf{X}}\left[\cdot|s\right]$ stands for the expected value, with respect to the random variable $\mathbf{X}$, given s; see Appendix A).

The central goal of decision theory is to answer the following question: what is the optimal decision rule for a certain problem? However, the *frequentist* risk function depends on the true state of nature s (as is clear from Eq. (1.2)), and thus is unable to provide a simple answer to this question; choosing an optimal decision rule that minimizes the *frequentist risk* would require knowledge of the unknown s. Moreover, each state of nature may lead to a different *optimal* decision rule. In more formal terms, it can be said that the *frequentist risk function* does not induce a *total ordering* in the set of all decision rules: it does not allow a direct comparison of the performance of two rules, independently of the (unknown) state of nature.

Nevertheless, the *frequentist risk function* does induce a partial ordering which is expressed by the concept of *admissibility.* This concept, in turn, is supported by the property of *domination*: a decision rule $\delta_1(\cdot)$ is *dominated* by another rule $\delta_0(\cdot)$ if:

(a) for any $s \in \mathcal{S}$, $R_{\delta_0}(s) \leq R_{\delta_1}(s)$, and

(b) there exists at least one $s_0 \in \mathcal{S}$ such that $R_{\delta_0}(s_0) < R_{\delta_1}(s_0)$.

A decision rule is said to be *admissible* if there exists no other rule that *dominates* it. It is clear that a decision rule which is inadmissible should not even be considered.

Example 1.2.1

This simple example (adapted from [8]) illustrates the strength and weakness of this concept: consider an estimation problem where $\mathcal{S} = \mathcal{A} = I\!R$ with a quadratic loss function $L(s,a) = (s-a)^2$; let the observation model be univariate Gaussian; more specifically, each observation consists of a single sample from a Gaussian random variable with mean s and unit variance $f_X(x|s) = \mathcal{N}(x|s,1)$ (where $\mathcal{N}(x|\mu,\sigma^2)$ denotes a Gaussian probability density function (p.d.f.) with mean μ and variance σ^2; see Appendix

A). Consider that the set of allowed decision rules we are interested in is $\mathcal{D} = \{\delta_k(x) = k\,x,\ k \in I\!R^+\}$. The risk function is

$$\begin{aligned} R_{\delta_k}(s) &= E_X\left[(s - k\,x)^2 | s\right] \\ &= s^2(1-k)^2 + k^2. \end{aligned} \tag{1.3}$$

What conclusions can be drawn from Eq. (1.3)? First, $\delta_1(x)$ dominates any $\delta_k(x)$, for $k > 1$; notice that $R_{\delta_1}(s) = 1$ and that $R_{\delta_k}(s) \geq 1$, for any $k \geq 1$. This fact makes all rules $\delta_k(x)$, with $k > 1$ inadmissible, and only $\delta_1(x)$ admissible. But then, on the negative side, observe that for $0 \leq k \leq 1$, no rule $\delta_k(x)$ dominates all the others (see Figure 1.1), thus all are admissible; in particular, notice that the rather unreasonable rule $\delta_0(x) = 0$ is also admissible, because $R_{\delta_0}(0) = 0$.

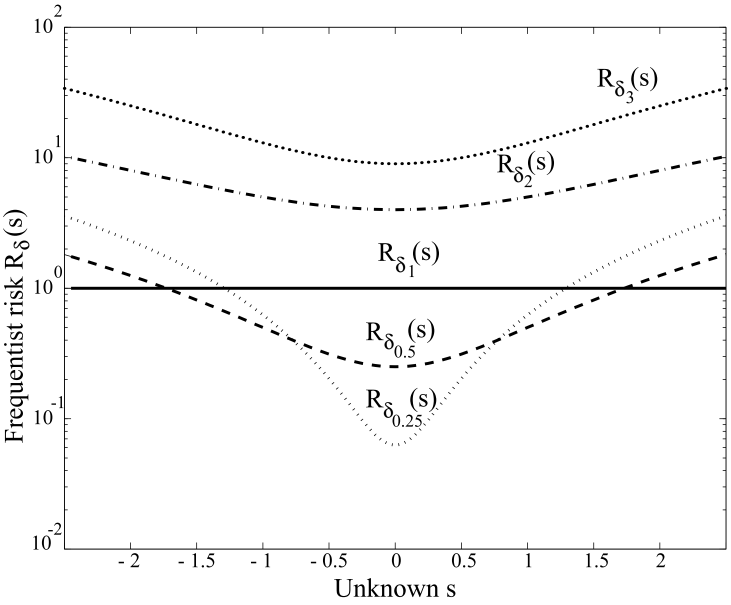

FIGURE 1.1. Frequentist risk functions for several rules of the type $\delta_k(x) = k\,x$, with a Gaussian observation model and under a quadratic loss function (see Example 1.2.1).

End of Example 1.2.1

There exists a more powerful frequentist concept which sometimes allows deriving decision rules: the *minimax risk* (associated with a given loss function $L(s, a)$) is defined as

$$\inf_{\delta \in \mathcal{D}} \sup_{s \in \mathcal{S}} R_\delta(s), \tag{1.4}$$

that is, the infimum over all allowed decision rules, of the supremum over all possible states of nature of the risk function. Notice that the minimax risk, if it exists, is a constant independent of s. A decision rule $\delta(\cdot)$ is said to be a *minimax* rule if, in the worst possible case, it achieves the minimax risk, i.e., if

$$\sup_{s\in\mathcal{S}} R_{\delta_0}(s) = \inf_{\delta\in\mathcal{D}} \sup_{s\in\mathcal{S}} R_\delta(s) \tag{1.5}$$

The reader should not worry too much about the technicalities involved in these concepts, but rather understand its simple meaning; if only closed sets are considered for $\mathcal{D}$ and $\mathcal{S}$, "inf" and "sup" can be replaced by "min" and "max", revealing the origin of the term "minimax". The concepts of *minimax risk* and *minimax rule* are well understood under a game theoretic view of statistical decision. Simply imagine "nature" and the "decision maker" as being involved in a two-person game: "nature" tries to choose its state s in such a way that it causes the maximum possible average loss, i.e., maximum risk (thus the "$\sup R_\delta(s)$"), to its opponent; knowing this, the "decision maker" tries to devise a decision rule tailored to minimizing the risk in that worst possible situation (thus the "$\inf_{\delta\in\mathcal{D}}$").

If a unique minimax decision rule exists, then this rule will clearly be admissible; if it were inadmissible, that would mean that it would be dominated by some other rule, and this would be in contradiction with the definition of minimaxity in Eq. (1.5). Of course, the converse is not true. However, it is important to keep in mind that a minimax decision rule may not exist. Finally, before presenting an example of a minimax rule, we refer the reader interested in a more complete treatment of *minimax analysis* to [8] (particularly, Chapter 5).

Example 1.2.2

Let us reconsider Example 1.2.1, recalling the expression for $R_{\delta_k}(s)$ in Eq. (1.3). Since

$$\begin{aligned}\sup_{s\in\mathbb{R}} R_{\delta_k}(s) &= \sup_{s\in\mathbb{R}} \left\{ s^2(1-k)^2 + k^2 \right\} \\ &= \begin{cases} 1 & \Leftarrow \quad k=1 \\ \infty & \Leftarrow \quad k\neq 1 \end{cases}\end{aligned} \tag{1.6}$$

(see also Figure 1.1 to verify that this is true) the minimax risk is

$$\inf_{k\in\mathbb{R}^+} \sup_{s\in\mathbb{R}} R_{\delta_k}(s) = 1 \tag{1.7}$$

and the corresponding *minimax rule* is $\delta_1(x) = x$; this is an intuitively acceptable rule.

End of Example 1.2.2

1.3 Bayesian Decision Theory

1.3.1 Subjective Probabilities and Degrees of Belief

The Bayesian approach to decision theory brings into play another element: *a priori* knowledge concerning the state of nature s, in the form of a probability function, usually referred to as "the prior." Mainly in statistics texts, priors are often denoted as $\pi(\cdot)$, but we will not adopt that convention here. Instead, we will use the conventional notation for probability functions $p_S(s)$.

From a conceptual point of view, the Bayesian approach to decision theory implies viewing probabilities as measures of knowledge or belief, the so-called *personal*, or *subjective*, probabilities [24], [30], [31], [58], [97]. Only by accepting such a view, is it possible to use probabilities to formalize *a priori* knowledge. The classical frequency-based interpretation of probability is clearly inadequate in many situations: for example, suppose that the unknown state of nature under which a decision has to be made (e.g., whether or not to perform surgery) is the presence or absence (a binary variable) of some disease in a certain patient. Clearly, there is nothing random about it: either the patient does or does not have the disease. Any (probabilistic type) statement such as *"there is a* 75% *chance that the patient has the disease"* has no frequency interpretation; there is no way we can have "a sequence of outcomes" of that same patient. Another example is any statement involving *"the probability of there being extra-terrestrial intelligent life in the known universe"*. Since there is only one known universe, there can be no frequency interpretation of such a probability; it simply expresses a degree of belief or a state of knowledge. Nevertheless, the statement is perfectly valid under the perpective of *subjective probability*; it expresses quantitatively a degree of personal belief. Several authors ([24], [30], [31], [48], [58], [97]) undertook the task of building formal theories for dealing with "degrees of belief"; these theories have to be consistent with standard (Boolean) logic but involve additional aspects allowing them to deal quantitatively with degrees of belief. It turns out that the "belief measures" are subject to the classical rules of probability theory, which itself does not depend on any frequency interpretation. This fact legitimates the use of probabilistic tools to formally deal with degrees of belief. The reader interested in these foundational aspects is encouraged to consult [58].

Let us then assume that the available knowledge (set of beliefs) about the unknown state of nature can be formalized by considering it as a random variable S, characterized by its probability function $p_S(s)$, for $s \in \mathcal{S}$; this will be a probability density or mass function, depending on $\mathcal{S}$ being a continuous or discrete set, respectively. In rare cases it can also happen that $\mathcal{S}$ contains both isolated points and continuous subsets, and thus $p_S(s)$ will have to be a mixed probability function including both point masses and densities. The terms *prior* and *a priori* are meant to stress that the knowledge they refer to is not provided by the observations, but is somehow possessed (*a priori*) by the decision maker (although possibly resulting from previous observations or experiments). A Bayesian decision problem is thus defined by the set of elements $\{\mathcal{S}, \mathcal{A}, \mathcal{X}, L(s,a), f_{\mathbf{X}}(\mathbf{x}|s), p_S(s)\}$. Again, the task of the (now Bayesian) *decision maker* is to derive a decision rule $\delta(\mathbf{x}) : \mathcal{X} \to \mathcal{A}$ under some optimality criterion.

1.3.2 A Posteriori Expected Loss and Bayesian Decisions

Let us recall that the frequentist evaluation of the performance of a decision rule is given by the *frequentist risk* in Eq. (1.2); this *risk* is obtained by averaging the loss function over all possible observations, thus ignoring that a specific observation is available, whenever a decision has to be made. A fundamental disadvantage of that measure, which we have already addressed, is that it depends on the unknown state of nature and therefore does not induce a total ordering on the set of allowed decision rules; in other words, it can not be used to derive an optimal decision rule.

The Bayesian approach, on the other hand, proposes a very different course of action. First of all, it adopts a *conditional* perspective under which all the importance is given to the performance of the decision rule $\delta(\mathbf{x})$ for the actual observed data $\mathbf{x}$, not for other possible observations that might have occurred but did not (see Section 2.14, below, for the implications of this choice). Secondly, it states that the loss function should be averaged over the state space $\mathcal{S}$, according to $p_S(s)$, since it is the state of nature s (not the observation $\mathbf{x}$) that is unknown. This rationale naturally leads to the *a posteriori expected loss*, conditioned on observation $\mathbf{x}$, as the fundamental criterion; it is defined as

$$\begin{aligned}\rho\left(p_S(s), \delta(\mathbf{x})|\mathbf{x}\right) &= E_S\left[L(s,\delta(\mathbf{x}))|\mathbf{x}\right] \\ &= \begin{cases} \displaystyle\int_{\mathcal{S}} L(s,\delta(\mathbf{x}))\, p_S(s|\mathbf{x})\, ds & \Leftarrow \quad \mathcal{S} \text{ is continuous} \\ \displaystyle\sum_{s\in\mathcal{S}} L(s,\delta(\mathbf{x}))\, p_S(s|\mathbf{x}) & \Leftarrow \quad \mathcal{S} \text{ is discrete.} \end{cases} \end{aligned} \tag{1.8}$$

In Eq. (1.8), $p_S(s|\mathbf{x})$ denotes the *a posteriori* probability (density or mass) function (also called the *posterior*); it is obtained via Bayes theorem (or law) which states (both in the discrete and continuous case) that

$$p_S(s|\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|s)\, p_S(s)}{f_{\mathbf{X}}(\mathbf{x})}, \tag{1.9}$$

where

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \displaystyle\int_{\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}|s)\, p_S(s)\, ds, & \text{if} \quad \mathcal{S} \text{ is continuous} \\ \displaystyle\sum_{s\in\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}|s)\, p_S(s), & \text{if} \quad \mathcal{S} \text{ is discrete,} \end{cases} \tag{1.10}$$

is the marginal (or unconditional) probability (mass or density, depending on whether $\mathcal{X}$ is continuous or discrete) function. In some contexts, the marginal $f_{\mathbf{X}}(\mathbf{x})$ is called the *predictive distribution* because this would be the probability (mass or density) assigned (predicted) to the particular observation $\mathbf{x}$ given the information carried by the prior.

The probability functions involved in Eqs. (1.9) and (1.10) should, to be rigorous, be written as $p_S(s|H)$, $f_{\mathbf{X}}(\mathbf{x}|s,H)$, $f_{\mathbf{X}}(\mathbf{x}|H)$, and $p_S(s|\mathbf{x},H)$, where H is the set of modeling hypotheses under which the prior $p_S(s|H)$ and the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s,H)$ were built. This would also stress that, in fact, there are no such thing as unconditional probabilities (or probability densities); all probability functions are (at least implicitly) conditioned by the (modeling) assumptions and hypotheses under which they were built. Having said this, we will still not include any reference to H in our notation unless we want to focus explicitly on the model assumptions.

Bayes' theorem plays the central role in Bayesian inference: it provides a formal tool to invert the roles of the unknown s and the observations $\mathbf{x}$, with respect to how they appear in the observation model (likelihood) $f(\mathbf{x}|s)$. In a sense, it is the general solution to probabilistic *inverse problems*, through which *a priori* probabilities $p(s)$ are *updated* into *a posteriori* ones, once the observations $\mathbf{x}$ have been obtained.

The *a posteriori* expected loss $\rho\,(p_S(s), \delta(\mathbf{x})|\mathbf{x})$ is *the* basic Bayesian criterion for the evaluation of decision rules; the fact that it depends on $\mathbf{x}$ is not a problem, because every time a decision is to be made, an observation $\mathbf{x}$ is in fact available. We stress again that this is one of the fundamental differences between frequentist and Bayesian approaches; the frequentist risk function is obtained by averaging over all possible observations, while the (Bayesian) *a posteriori* expected loss is a function of the particular observation at hand.

Finally, the optimal *Bayes' decision* is naturally obtained by looking for the action that minimizes the *a posteriori* expected loss

$$\delta(\mathbf{x}) = \arg\min_{d\in\mathcal{A}} \rho\,(p(s), d|\mathbf{x})\,, \tag{1.11}$$

for each the particular observation $\mathbf{x} \in \mathcal{X}$.

1.3.3 Bayes Risk

A common way of evaluating a decision rule is by computing the so called *Bayes' risk* or *integrated risk*, $r(p(s), \delta(\cdot))$; this is simply the frequentist risk averaged with respect to the unknown state of nature, i.e., the loss function averaged over $\mathcal{S}$ and $\mathcal{X}$, according to the respective probability functions,

$$\begin{aligned} r(p_S(s), \delta(\cdot)) &= \int_{\mathcal{S}} R_\delta(s)\, p_S(s)\, ds && (1.12) \\ &= \int_{\mathcal{S}} \int_{\mathcal{X}} L(s, \delta(\mathbf{x}))\, f_{\mathbf{X}}(\mathbf{x}|s)\, p_S(s)\, d\mathbf{x}\, ds \\ &= \int_{\mathcal{S}} \int_{\mathcal{X}} L(s, \delta(\mathbf{x}))\, p_{\mathbf{x},S}(\mathbf{x}, s)\, d\mathbf{x}\, ds, && (1.13) \end{aligned}$$

where

$$\begin{aligned} p_{\mathbf{x},S}(\mathbf{x}, s) &= f_{\mathbf{X}}(\mathbf{x}|s)\, p_S(s) \\ &= p_S(s|\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x}) && (1.14) \end{aligned}$$

is the joint probability (density or mass) function of the random variables S and $\mathbf{X}$. One or both integrations have to be replaced by summations in the case where $\mathcal{S}$ or/and $\mathcal{X}$ are discrete sets. Exact conditions for the validity of interchanging the order of integration are given by Fubini's theorem and will not be considered here; see, e.g., [3].

The Bayes risk function has the following properties.

- It yields a real number (not a function of s or $\mathbf{x}$) for each decision rule, thus inducing a total ordering in the set of decision rules; i.e. they can be compared directly.

- More importantly, the *a posteriori* expected loss and Bayes risk are absolutely equivalent, i.e., they lead to the same decision rule. Considering (without loss of generality) that $L(s, a) \geq L_{\min} = 0$ (and recalling Eq. (1.13)),

$$\begin{aligned} \min_{\delta(\mathbf{x})} r(p_S(s), \delta(\cdot)) &= \min_{\delta(\mathbf{x})} \int_{\mathcal{S}} \int_{\mathcal{X}} L(s, \delta(\mathbf{x}))\, p_{\mathbf{x},S}(\mathbf{x}, s)\, d\mathbf{x}\, ds \\ &= \min_{\delta(\mathbf{x})} \int_{\mathcal{X}} \left(\int_{\mathcal{S}} L(s, \delta(\mathbf{x}))\, p_S(s|\mathbf{x})\, ds \right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\ &= \min_{\delta(\mathbf{x})} \int_{\mathcal{X}} \rho\left(p_S(s), \delta(\mathbf{x})|\mathbf{x}\right)\, f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\ &= \int_{\mathcal{X}} \left(\min_{\delta(\mathbf{x})} \rho\left(p_S(s), \delta(\mathbf{x})|\mathbf{x}\right) \right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} && (1.15) \end{aligned}$$

  because if $L(s, a) \geq 0$ then $\rho\left(p(s), \delta(\mathbf{x})\right) \geq 0$, and minimizing the integral of a non-negative function is equivalent to minimizing the function at each point. Comparing the function inside parentheses in Eq. (1.15) with Eq. (1.11) makes it clear that both specify the same decision rule. If $L(s, a) \geq L_{\min} \neq 0$, we use $L'(s, a) = L(s, a) - L_{\min} \geq 0$ since two loss functions differing by a constant obviously lead to the same decision.

It should be kept in mind that the truly Bayesian criterion is the posterior expected loss, and not the integrated risk (or Bayes risk); only the posterior expected loss respects the *conditionality principle* (see Section 2.14 below), that is, only relies on the observed data. Widespread use of the integrated risk as a criterion for deriving Bayesian decision rules is only ratified by the equivalence property just described.

1.3.4 Admissibility of Bayesian Rules

It is interesting to see how Bayesian decision rules fare with respect to frequentist performance measures. In fact, it turns out that Bayesian decision rules can be shown to be *admissible* (see Section 1.2.2), that being a reassuring fact.

To see why this is so, consider that $\mathcal{S}$ is a discrete and finite set $\mathcal{S} = \{s_1, s_2 ..., s_M\}$ and that the prior assigns strictly positive probability to each possible state of nature in $\mathcal{S}$. Let $\delta(\mathbf{x})$ be a Bayes rule, i.e., one that minimizes the *a posteriori* expected loss, according to Eq. (1.11). Now assume that this rule was *inadmissible* (see Section 1.2.2). This would mean that $\delta(\mathbf{x})$ would be dominated by another rule $\delta'(\mathbf{x})$; their frequentist risks would follow $R_{\delta'}(s_i) \leq R_\delta(s_i)$, for all $s_i \in \mathcal{S}$, and there would be at least one state of nature, say s_j, such that $R_{\delta'}(s_j) < R_\delta(s_j)$. Now, recall from Section 1.3.3, that a Bayes' rule also minimizes the Bayes risk which, according to Eq. (1.12), is simply the average of the frequentist risk over all possible states of nature. Writing the Bayes' risk for $\delta'(\mathbf{x})$,

$$\begin{aligned} r(p_S(s), \delta'(\cdot)) &= \sum_{s_i \in \mathcal{S}} R_{\delta'}(s_i) p_S(s_i) \\ &< \sum_{s_i \in \mathcal{S}} R_\delta(s_i) p_S(s_i) \;=\; r(p_S(s), \delta(\cdot)), \end{aligned} \tag{1.16}$$

where the strict inequality results from the existence of one s_j for which $R_{\delta'}(s_j) < R_\delta(s_j)$ and the fact that all $p_S(s_i) > 0$. But this last expression clearly contradicts the fact that $\delta(\mathbf{x})$ is a Bayes' rule, our starting hypothesis. This shows, by contradiction, that $\delta(\mathbf{x})$ can not be inadmissible.

A similar proof can be used in the case where $\mathcal{S}$ is a continuous set. It suffices to assume the condition that the prior does not give zero probability to any open subset of $\mathcal{S}$ (for the same reason that we assumed that no single element of a discrete $\mathcal{S}$ could have zero probability). That condition allows performing the same steps as in Eq. (1.16), with the summation replaced by an integral over $\mathcal{S}$. In this case there is another (technical) condition which is the continuity of the loss function; the reader interested in further details is referred to [8], [93].

This relation between the concept of admissibility and the Bayes risk could be naturally expected. Admissibility characterizes a decision rule based on how it performs for all possible states of nature, and so does the Bayes' risk although in a different way.

1.3.5 Predictive Problems

In practical problems it is often the case that the consequences of a decision do not depend directly on the unknown state of nature s but rather on the outcome of another observable $\mathbf{y}$ which depends on s. To be concrete, let us consider a prior $p_S(s)$, and an observation model $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}|s)$. Then, from some observed outcome $\mathbf{x}$, a decision $\delta(\mathbf{x})$ is chosen whose loss is measured by $L(\mathbf{y},\delta(\mathbf{x}))$, where $\mathbf{y}$ is the next outcome (unobserved at the time when the decision was chosen) of the random variable $\mathbf{Y}$.

In a typical instance of this scenario, $\mathbf{X}$ and $\mathbf{Y}$ correspond to two (possibly consecutive) time instants of some discrete-time random process $\mathbf{Z}(t)$, $t = 1, 2, ...$; i.e., $\mathbf{X} = \mathbf{Z}(t_1)$ and $\mathbf{Y} = \mathbf{Z}(t_2)$ with $t_2 > t_1$, and having observed a particular outcome $\mathbf{x} = \mathbf{z}(t_1)$, the objective is to *predict* the outcome of $\mathbf{Y} = \mathbf{Z}(t_2)$. Notice that this formulation also covers the case where $\mathbf{x}$ is a sequence of, say n, past (dependent or independent and not necessarily consecutive) observations and $\mathbf{y}$ the next still unobserved one; it suffices to write $\mathbf{X} = [\mathbf{Z}(t_1 - k_1), ..., \mathbf{Z}(t_1 - k_n)]$ and $\mathbf{Y} = \mathbf{Z}(t_1)$. Another common setting is to assume that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent (given s) and identically distributed, and thus the problem is one of estimating the next outcome of a random variable, given the present observed one. All problems of this type are called *predictive problems*, and can easily be addressed by the standard tools of Bayesian decision theory.

To deal with *predictive problems* it is necessary to have a predictive version of the *a posteriori* expected loss; this is easily obtained by following the Bayesian course of action: conditioning on the observed, in this case $\mathbf{x}$, and averaging over the unknowns, s and $\mathbf{y}$ (see Section 1.3.2). According to these principles, the *predictive a posteriori expected loss* is given by

$$\begin{aligned}
\rho_{\mathrm{p}}\left(p_S(s), \delta(\mathbf{x})|\mathbf{x}\right) &= E_{S,\mathbf{Y}}\left[L(\mathbf{y},\delta(\mathbf{x}))|\mathbf{x}\right] \\
&= \int_{\mathcal{Y}} \int_{\mathcal{S}} L(\mathbf{y},\delta(\mathbf{x}))\, p_{S,\mathbf{Y}}(s,\mathbf{y}|\mathbf{x})\, ds\, d\mathbf{y} \\
&= \int_{\mathcal{Y}} L(\mathbf{y},\delta(\mathbf{x})) \int_{\mathcal{S}} p_{S,\mathbf{Y}}(s,\mathbf{y}|\mathbf{x})\, ds\, d\mathbf{y} \\
&= \int_{\mathcal{Y}} L(\mathbf{y},\delta(\mathbf{x})) p_{\mathbf{Y}}(\mathbf{y}|\mathbf{x})\, d\mathbf{y} \qquad (1.17)
\end{aligned}$$

where

$$p_{\mathbf{Y}}(\mathbf{y}|\mathbf{x}) = \int_{\mathcal{S}} p_{S,\mathbf{Y}}(s,\mathbf{y}|\mathbf{x})\, ds = \int_{\mathcal{S}} \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}|s)\, p_S(s)}{f_{\mathbf{X}}(\mathbf{x})}\, ds \qquad (1.18)$$

is the *a posteriori predictive density*, and

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{\mathcal{Y}} \int_{\mathcal{S}} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}|s)\, p_S(s)\, ds\, d\mathbf{y} \tag{1.19}$$

is the marginal density of $\mathbf{X}$. Equivalent expressions, with summations replacing the integrals, may be written for discrete problems.

Observe the meaning of Eqs. (1.17) and (1.19): since the loss function does not involve the state of nature directly, it is simply integrated out, leaving only explicit reference to the relevant quantities: the observed $\mathbf{x}$ and the unknown $\mathbf{y}$. Then, the *a posteriori predictive density* $p_{\mathbf{Y}}(\mathbf{y}|\mathbf{x})$ becomes the basis of any Bayesian inference. This is recurrent feature of Bayesian analysis; anything that is unknown but about which we are not interested in making inferences (does not appear in the loss function) is simply removed by marginalization.

A particular important case is the one where $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent (given s) and identically distributed according to some common probability function, say $f_{\mathbf{Z}}(\mathbf{z}|s)$; this allows writing

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}|s) = f_{\mathbf{Z}}(\mathbf{x}|s)\, f_{\mathbf{Z}}(\mathbf{y}|s); \tag{1.20}$$

as a consequence, the marginal $f_{\mathbf{X}}(\mathbf{x})$ becomes simpler to obtain

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{\mathcal{S}} f_{\mathbf{Z}}(\mathbf{x}|s)\, p_S(s) \underbrace{\int_{\mathcal{Z}} f_{\mathbf{Z}}(\mathbf{y}|s) d\mathbf{y}}_{=1}\, ds = \int_{\mathcal{S}} f_{\mathbf{Z}}(\mathbf{x}|s)\, p_S(s)\, ds. \tag{1.21}$$

Notice that this includes the case where we have only one observation model, say $f_{\mathbf{Z}}(\mathbf{z}|s)$, and $\mathbf{x} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$ is a sequence of n independent samples from that observation model, while the goal is to make some inference involving the next (still unobserved) observation. This scenario is captured by the independence assumption in Eq. (1.20), together with

$$f_{\mathbf{X}}(\mathbf{x}|s) = \prod_{j=1}^{n} f_{\mathbf{Z}}(\mathbf{z}_j|s)$$

which can, in turn, be reinserted in Eqs. (1.20) and (1.21).

1.3.6 Inference versus Decision

Although we have introduced the Bayesian framework from a decision theoretic point of view, this may be seen by fundamentalist Bayesians as somewhat heretic. It is often defended that the decision making step is outside the scope of Bayesian approach, whose role would be complete once it produced an *a posteriori* probability function; this is often called *inference*, in Bayesian parlance, as opposed to *decision.* From that perspective, the

Bayesian statistician should just report the *a posteriori* probability distribution, and refrain from producing any decisions which should be left to the final user. In fact, choosing a loss function is a highly problem-dependent issue which turns out, more often than not, to be dominated by computational tractability consideration (even if this is rarely acknowledged).

Once the *a posteriori* probability function is obtained, it can be exploited in many ways; obtaining an estimate or decision that minimizes some loss function is just one possibility. Another choice is to somehow report a summary of its main features. The Bayesian viewpoint advocates that, just as the *a priori* knowledge is contained in the *a priori* probability function, all the knowledge available about the unknown state of nature after observing data is expressed in the *a posteriori* probability function. Any reduction of this function to a single value (e.g., an estimate) causes an irrecoverable loss of information.

Having said this, we will proceed into the presentation of specific loss functions; we stress again that only the full *a posteriori* probability function contains all the *a posteriori* knowledge available. Reporting one decision may be seen as a way of summarizing this posterior by means of a single point optimally chosen with respect to the loss function being adopted.

1.4 Bayesian Classification

As was mentioned above, many image analysis and most pattern recognition problems can be classified as statistical classification problems. The fundamental characteristic here is the non-numerical, that is, the *categorical*, nature of the unknown.

1.4.1 Introduction

In classification problems, it is assumed that nature takes values on a discrete set $\mathcal{S}$ and that the goal is to *decide* which is the true state of nature (i.e., we take $\mathcal{A} = \mathcal{S}$), given the observation $\mathbf{x}$. There is also knowledge of $f_{\mathbf{X}}(\mathbf{x}|s)$, for $s \in \mathcal{S}$, which in this context is usually referred to as the class-conditional observation model. The actual observations are a sample of the true one, i.e., of $f_{\mathbf{X}}(\mathbf{x}|s_{\text{true}})$. The prior here is a probability mass function (since $\mathcal{S}$ is discrete), $p_S(s)$, for $s \in \mathcal{S}$, and the denominator of Bayes' theorem (Eq. (1.9)) then appears in its discrete version

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{s \in \mathcal{S}} f_{\mathbf{X}}(\mathbf{x}|s)p(s);$$

notice that this is true, regardless of $f_{\mathbf{X}}(\mathbf{x})$ and $f_{\mathbf{X}}(\mathbf{x}|s)$ being probability densities (continuous $\mathcal{X}$) or mass functions (discrete $\mathcal{X}$).

1.4.2 Classification Under The "0/1" Loss Function

The *0/1 loss function*, for classification problems, assigns zero cost to any correct decision, and unit cost to any wrong decision,

$$L(s,a) = \begin{cases} 1 & \Leftarrow \quad s \neq a \\ 0 & \Leftarrow \quad s = a. \end{cases} \tag{1.22}$$

Inserting this definition into the general expression for the *a posteriori expected loss*, Eq. (1.8), and then looking for the Bayes' optimal decision as given by Eq. (1.11), leads to (recall that $\mathcal{S} = \mathcal{A}$)

$$\begin{aligned} \delta(\mathbf{x}) &= \arg\min_{d\in\mathcal{A}} \sum_{s\in\mathcal{S}} L(s,d) p_S(s|\mathbf{x}) && (1.23) \\ &= \arg\min_{d\in\mathcal{S}} \left(\left[\sum_{s\in\mathcal{S}} p_S(s|\mathbf{x}) \right] - p_S(d|\mathbf{x}) \right) \\ &= \arg\min_{s\in\mathcal{S}} \left(1 - p_S(s|\mathbf{x})\right) \\ &= \arg\max_{s\in\mathcal{S}} p_S(s|\mathbf{x}) \equiv \delta_{\text{MAP}}(\mathbf{x}), && (1.24) \end{aligned}$$

which is referred to as the *maximum a posteriori* (MAP) classifier. In other words, the optimal Bayes decision rule is to choose the class presenting the maximum posterior probability, given the particular observation at hand.

Since, for a given observation $\mathbf{x}$, the marginal $f_{\mathbf{X}}(\mathbf{x})$ in the denominator of Bayes' theorem (Eq. (1.9)) is a constant, the MAP criterion can be further simplified to

$$\delta_{\text{MAP}}(\mathbf{x}) = \arg\max_{s\in\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}|s)\, p_S(s), \tag{1.25}$$

which is equivalent to the maximizer of the joint probability function

$$\delta_{\text{MAP}}(\mathbf{x}) = \arg\max_{s\in\mathcal{S}} f_{\mathbf{X},S}(\mathbf{x},s). \tag{1.26}$$

It is also common to see the logarithmic version of the MAP criterion

$$\delta_{\text{MAP}}(\mathbf{x}) = \arg\max_{s\in\mathcal{S}} \left\{\log f_{\mathbf{X}}(\mathbf{x}|s) + \log p_S(s)\right\}; \tag{1.27}$$

this form, probably even more than Eq. (1.25), makes clear that the MAP decision rule tries to reach a compromise between the *a priori* expectations carried by $p_S(s)$ and the evidence provided by the data via the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$.

In binary (two-class) problems, when $\mathcal{A} = \mathcal{S} = \{s_1, s_2\}$, it is possible to go a little further beyond the general form of the MAP classifier and write

$$\delta_{\text{MAP}}(\mathbf{x}) = \begin{cases} s_1 & \Leftarrow \quad l(\mathbf{x}) \geq t \\ s_2 & \Leftarrow \quad l(\mathbf{x}) < t \end{cases} \tag{1.28}$$

with

$$l(\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|s_1)}{f_{\mathbf{X}}(\mathbf{x}|s_2)} \quad \text{and} \quad t = \frac{p_S(s_2)}{p_S(s_1)} \tag{1.29}$$

where $l(\mathbf{x})$ is called a *likelihood ratio* and t is a decision threshold. Notice the competition between the data evidence (provided by the likelihood ratio) and the *a priori* probability ratio.

Example 1.4.1

A classical example of binary decision, in which the "0/1" loss function is used (for a reason to be made clear below) is the binary *detection* problem, as known in the communications literature. In its simplest form, an emitter outputs one of two constant values, say s_0 and s_1, representing, respectively, the "0" and "1" binary digits; their (known) *a priori* probabilities are p_0 and $p_1 = 1 - p_0$, respectively. Each digit is transmitted through a *noisy channel* that adds to it a sample of a Gaussian random variable of zero mean and variance σ^2; this is the simplest, but by far the most common, model for channel noise in digital communications. The corresponding class-conditionals of the received value x, are then $f_X(x|s_0) = \mathcal{N}(x|s_0, \sigma^2)$ and $f(x|s_1) = \mathcal{N}(x|s_1, \sigma^2)$. Due to the exponential nature of the Gaussian p.d.f. it is more convenient to work here with the logarithmic version of Eq. (1.29), which, after simple manipulations yields

$$\delta_{\text{MAP}}(x) = \begin{cases} s_0 & \Leftarrow & l(x) \geq t \\ s_1 & \Leftarrow & l(x) < t \end{cases} \tag{1.30}$$

with

$$l(x) = (x - s_1)^2 - (x - s_0)^2 \tag{1.31}$$

and

$$t = 2\,\sigma^2 \log\left(\frac{p(s_1)}{p(s_0)}\right). \tag{1.32}$$

Notice that $l(x)$ measures the difference between the squared distances from the observed value to s_1 and s_2. We will examine this problem in more detail in Example 1.4.10 below.

End of Example 1.4.1

1.4.3 A Special Case: Gaussian Observations

An exhaustively studied family of classification problems is that where the *class conditionals* (the *likelihoods*) are (possibly multivariate) Gaussian densities. From a pattern classification perspective, [37] is the classical reference; more recent texts (of varying technical depth) include [16], [34], [38], [75], and [87]. References [76], [77], [96], [104], and [109] are classical and often cited fundamental texts from the communications/signal detection point of view (for more recent texts, see also, e.g., [53], [64], [68], [84], or [98]).

Example 1.4.1 considered the simplest member of that family: univariate Gaussian observation with a common variance. For an $M-$class problem, with $\mathcal{S} = \{s_1, s_2, ...s_M\}$, if the observations are n-dimensional vectors (i.e., $\mathcal{X} = I\!R^n$) the Gaussian likelihoods are

$$f_{\mathbf{X}}(\mathbf{x}|s_i) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C}_i)}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)^T \mathbf{C}_i^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)\right\}. \quad (1.33)$$

In Eq. (1.33), $(\cdot)^T$ denotes vector transpose and $\det(\cdot)$ stands for the determinant of a matrix; each $\boldsymbol{\mu}_i$ is the $n-$dimensional mean vector corresponding to class s_i,

$$\boldsymbol{\mu}_i = E_{\mathbf{X}}\left[\mathbf{x}|s_i\right]$$

and each $\mathbf{C}_i$ is the covariance matrix associated with the observations from class s_i, defined as

$$\mathbf{C}_i = E_{\mathbf{X}}\left[(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T|s_i\right].$$

This particular form allows rewriting the MAP classifier from Eq. (1.27), using natural logarithms and dropping constants, as

$$\delta_{\text{MAP}}(\mathbf{x}) = \quad (1.34)$$
$$\arg\max_{s_i \in \mathcal{S}}\left\{2\log p_S(s_i) - \log\det(\mathbf{C}_i) - \left(\mathbf{x} - \boldsymbol{\mu}_i\right)^T \mathbf{C}_i^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)\right\}.$$

This form can be further simplified in the following two special cases:

Independent observations with common variance: A scenario where only the mean is assumed to differ from one class to another, and the observations are independent with equal variance σ^2, is described by taking $\mathbf{C}_i = \sigma^2\mathbf{I}$, for $i = 1, 2, .., M$ (where $\mathbf{I}$ denotes an identity matrix). This is a common model in signal processing and communications where it is known as *independent additive white Gaussian noise* (IAWGN) channel (here, *independent* refers to the fact that the noise characteristics are independent of the true class); notice that such a model results if we assume that, given some s_i, the observation model is $\mathbf{X} = \boldsymbol{\mu}_i + \mathbf{N}$, where $\mathbf{N}$ is a $n-$dimensional vector of independent zero-mean Gaussian random variables with variance σ^2. Clearly, the terms $\log\det(\mathbf{C}_i)$ are all equal and can be dropped from Eq. (1.34). Also, $\mathbf{C}_i^{-1} = \mathbf{I}/\sigma^2$ which results in the following simpler classifier:

$$\begin{aligned}\delta_{\text{MAP}}(\mathbf{x}) &= \arg\max_{s_i \in \mathcal{S}}\left\{2\,\sigma^2\,\log p_S(s_i) - \left(\mathbf{x} - \boldsymbol{\mu}_i\right)^T\left(\mathbf{x} - \boldsymbol{\mu}_i\right)\right\}\\ &= \arg\min_{s_i \in \mathcal{S}}\left\{-2\,\sigma^2\,\log p_S(s_i) + \|\mathbf{x} - \boldsymbol{\mu}_i\|^2\right\}, \quad (1.35)\end{aligned}$$

where $\|\mathbf{x}-\boldsymbol{\mu}_i\|^2$ denotes the squared Euclidean distance between $\mathbf{x}$ and $\boldsymbol{\mu}_i$. This form makes it clear that the classifier is finding a compromise between selecting the class whose mean $\boldsymbol{\mu}_i$ is closest to the observation $\mathbf{x}$ and the *a priori* probability of that class. Notice how the criterion derived in Example 1.4.1 is a simple instance of this result. Although it may not be apparent at first sight, there are still some constant terms (with respect to the classes) in Eq. (1.35) which can be dropped; to see this, notice that $\|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \|\mathbf{x}\|^2+\|\boldsymbol{\mu}_i\|^2-2\mathbf{x}^T\boldsymbol{\mu}_i$ and that, in the presence of an observation $\mathbf{x}$, the first term $\|\mathbf{x}\|^2$ is a constant. We can then reduce the MAP classifier to its simplest form

$$\delta_{\mathrm{MAP}}(\mathbf{x}) = \arg\max_{s_i\in\mathcal{S}}\{\mathbf{x}^T\boldsymbol{\mu}_i + \underbrace{\sigma^2\log p_S(s_i) - \|\boldsymbol{\mu}_i\|^2/2}_{\beta_i}\}; \qquad (1.36)$$

Notice that the classifier has to simply compute the inner product $\mathbf{x}^T\boldsymbol{\mu}_i$ of the observed vector $\mathbf{x}$ with each of the $\boldsymbol{\mu}_i$, add a class-dependent *bias* β_i, and choose the largest result. Classifiers of this form are called *linear classifiers* because the only operations involving the observed data (inner products) are linear ones.

Dependent observations with common covariance: In this next most simple configuration, it is assumed that although the observations are not independent of each other, all covariance matrices are the same, $\mathbf{C}_i = \mathbf{C}, i = 1,2,...M$. From a signal processing perspective, this models what is called *independent* (from the class) *additive colored (non-white) Gaussian noise* (IACGN) channel. This model is applicable if, given some s_i, the observation model is $\mathbf{X} = \boldsymbol{\mu}_i + \mathbf{N}$, where N is now a $n-$dimensional vector of mutually dependent (or non-white, or colored) zero-mean Gaussian random variables with covariance matrix $\mathbf{C}$. Since $\mathbf{C}_i = \mathbf{C}$, it is still possible to drop the (constant) $\log\det(\mathbf{C}_i)$ term and write

$$\begin{aligned}\delta_{\mathrm{MAP}}(\mathbf{x}) &= \arg\max_{s_i\in\mathcal{S}}\left\{2\,\log p_S(s_i) - (\mathbf{x}-\boldsymbol{\mu}_i)^T\,\mathbf{C}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_i)\right\}\\ &= \arg\min_{s_i\in\mathcal{S}}\left\{-2\,\log p_S(s_i) + \|\mathbf{x}-\boldsymbol{\mu}_i\|^2_{\mathbf{C}}\right\} \qquad (1.37)\end{aligned}$$

where

$$\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\mathbf{C}} \equiv (\mathbf{x}-\boldsymbol{\mu})^T\,\mathbf{C}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})$$

is called the squared *Mahalanobis distance*. We can then drop constant terms and arrive at

$$\begin{aligned}\delta_{\mathrm{MAP}}(\mathbf{x}) &= \arg\max_{s_i\in\mathcal{S}}\{\mathbf{x}^T\underbrace{(\mathbf{C}^{-1}\boldsymbol{\mu}_i)}_{\boldsymbol{\alpha}_i} + \underbrace{\log p_S(s_i) - \|\boldsymbol{\mu}_i\|^2_{\mathbf{C}}/2}_{\beta_i}\},\\ \delta_{\mathrm{MAP}}(\mathbf{x}) &= \arg\max_{s_i\in\mathcal{S}}\{\mathbf{x}^T\boldsymbol{\alpha}_i + \beta_i\}, \qquad (1.38)\end{aligned}$$

which is still a linear classifier (compare it with Eq. (1.36)).

Dependent observations with covariances differing by a scale factor: This is a slight variation from the last case. Consider that the covariance matrices of the classes only differ from each other by some scale factor, that is, we can write $\mathbf{C}_i = \gamma_i \mathbf{C}$. In this case,

$$\log \det(\mathbf{C}_i) = \log \det(\gamma_i \mathbf{C}) = n \log \gamma_i + \log \det(\mathbf{C}).$$

Since $\log \det(\mathbf{C})$ is a constant, we still obtain a linear classifier of the form

$$\delta_{\text{MAP}}(\mathbf{x}) = \arg \max_{s_i \in \mathcal{S}} \{\mathbf{x}^T \boldsymbol{\alpha}_i + \beta_i\}, \tag{1.39}$$

with the $\boldsymbol{\alpha}_i$ and β_i now given by

$$\begin{aligned} \boldsymbol{\alpha}_i &= \mathbf{C}^{-1} \boldsymbol{\mu}_i \\ \beta_i &= \gamma_i \log p_S(s_i) - \frac{1}{2} \|\boldsymbol{\mu}_i\|^2_{\mathbf{C}} - \frac{1}{2} n \gamma_i \log \gamma_i. \end{aligned} \tag{1.40}$$

Example 1.4.2

Let us recall Example 1.4.1, but now consider that each binary digit is represented by a previously specified sequence of n values (i.e., a discrete signal) rather than one single value. Let the sequence representing bit "0" be $\{s_0(1), ..., s_0(n)\}$ and that associated with bit "1" be $\{s_1(1), s_1(2), ...s_1(n)\}$. These values can be collected and stacked into two vectors, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. As in example 1.4.1, transmission takes place through a noisy channel that adds (to all values sent through it) independent samples of a Gaussian random variable of zero mean and variance σ^2 (white noise). Denoting by $\mathbf{x}$ the sequence of values observed by the receiver, the class-conditional probability density functions are $f_{\mathbf{X}}(\mathbf{x}|s_0) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_0, \sigma^2\mathbf{I})$ and $f_{\mathbf{X}}(\mathbf{x}|s_1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \sigma^2\mathbf{I})$. The notation $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$ now stands for a multivariate Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. This is clearly a (binary) classification problem with Gaussian likelihood falling into the first special category "independent observations with common variance" and so the optimal Bayesian detector is linear and given by Eq. (1.36).

If the noise is not white, i.e. if the n consecutive noise samples are not independent (the covariance matrix $\mathbf{C}$ is not a diagonal matrix), the signal detection problem falls into our second special category "dependent observations with common covariance". The optimal Bayes' detector is then the one in Eq. (1.38).

End of Example 1.4.2

1.4.4 General Costs

It is obviously possible to consider a more general cost structure than the simple "0/1" cost function. For any $M - class$ problem, where $\mathcal{A} = \mathcal{S} =$

$\{s_1, s_2, ...s_M\}$, an $M \times M$ *cost matrix* $\mathbf{L}$ can be defined where element $L_{i,j} = L(s_i, s_j)$ specifies the cost associated with choosing s_j when the true class was s_i. Although this does not lead to as simple a criterion as the MAP, many situations require this additional possibility. A classical example of this this type of situation is one of medical decision making: the unknown state of nature corresponds to the presence or absence of some serious disease (let us write $\mathcal{S} = \{\text{“disease”}, \text{“no disease”}\}$), which, if surgery is not performed becomes terminal; the decision here is whether or not to do the surgery (so $\mathcal{A} = \{\text{“surgery”}, \text{“no surgery”}\}$). It is natural in such a situation to take $L(\text{“disease”}, \text{“surgery”}) = 0$, $L(\text{“no disease”}, \text{“no surgery”}) = 0$, and $L(\text{“disease”}, \text{“no surgery”}) \gg L(\text{“no disease”}, \text{“surgery”})$ expressing the fact that the consequences of not performing surgery on a patient that does have the disease are more serious than those of performing it on a patient that turns out not to have it. In the signal detection literature, these situations are sometimes called “miss” and “false alarm”.

In the binary case, it is simple to show that the threshold in Eq. (1.29) can be modified as

$$t = \frac{p_S(s_2)}{p_S(s_1)} \cdot \frac{L(s_2, s_1) - L(s_2, s_2)}{L(s_1, s_2) - L(s_1, s_1)}. \tag{1.41}$$

This expression shows how the *a priori* probabilities and the cost structure (the consequences) are combined to yield a decision threshold. It is easy to find this type of reasoning in everyday life, i.e., weighting probabilities against consequences. The chances of wining a lottery are extremely low; however, the cost of not wining is also very low, while the consequences of wining are extremely good; this is why people do buy lottery tickets. As another example of the weighting of consequences versus probabilties, who wouldn't be willing to play some game where the chances of winning are 5/6? So, why is it so difficult to find people willing to play Russian roulette?

In general M-ary situations, if no particular structure is assumed for the cost matrix, it is not possible to further simplify Eq. (1.23).

Example 1.4.3

Let us now consider a less technically oriented example. Consider that someone is betting on the outcomes of coin tosses, owned and tossed by a friend. The first few outcomes are all heads, and the player starts suspecting that the coin is double-headed; however, he had some *a priori* confidence on the honesty of his friend. Moreover, he must take into account the cost of a wrong decision; if he says “stop, this is a double-headed coin!”, and the coin turns out to be a fair one, that will cost him the other person's friendship; whereas, if the coin is in fact double-headed and he says nothing, that will only cost him a small amount of money. The question that has to be answered is how many consecutive heads should be allowed before the coin is declared double-headed. Let us see how this can be approached with

Bayesian decision tools. Firstly, the two possible states of nature are simply $\mathcal{S} = \{s_1, s_2\}$, where s_1 = "double-headed coin" and s_2 = "fair coin". The *a priori* knowledge about the coin owner's (dis)honesty is expressed by $p_S(s_1) = p_1$ (with $p_2 = p_S(s_2) = 1 - p_S(s_1)$). A reasonable cost structure, in view of what was said above, is $L(s_1, s_1) = L(s_2, s_2) = 0$ and $L(s_2, s_1) \gg L(s_1, s_2)$. Now let $\mathbf{x} = (x_1, x_2, ..., x_n)$, with each $x_i \in \{$"heads", "tails"$\}$, be the observed sequence of outcomes. The two class-conditional probability functions are

$$f_{\mathbf{X}}(\mathbf{x}|s_1) = \begin{cases} 1 & \Leftarrow \quad \mathbf{x} \text{ is a sequence of } n \text{ heads} \\ 0 & \Leftarrow \quad \mathbf{x} \text{ has at least one tail outcome,} \end{cases} \tag{1.42}$$

and $f_{\mathbf{X}}(\mathbf{x}|s_2) = (1/2)^n$, for any sequence $\mathbf{x}$. The resulting *a posteriori* probability function is

$$p_S(s|\mathbf{x}) \propto \begin{cases} \left\{ \begin{array}{lll} p_1 & \Leftarrow & \mathbf{x} \text{ is a sequence of } n \text{ heads} \\ 0 & \Leftarrow & \mathbf{x} \text{ has at least one tail outcome,} \end{array} \right\} & \Leftarrow \quad s = s_1 \\ (1 - p_1)\left(\frac{1}{2}\right)^n & \Leftarrow \quad s = s_2, \end{cases}$$

where we have used "$\propto$" because the expression is not normalized. What MAP decisions can be obtained from this? Clearly, if the observed sequence is not all heads, then s_1 gets *a posteriori* zero probability, as would be expected since a double-headed coin can not produce tail outcomes, and s_2 is the only possible decision. But, returning to the above scenario, what happens when a sequence of heads is observed? How should this data be balanced against our prior belief in the fairness of the coin, and the given cost structure? Denoting as $\mathbf{x}_h(n)$ a sequence of n heads, we obtain the following *a posteriori expected losses*

$$\rho(p_S(s), s_1|\mathbf{x}_h(n)) \quad = \quad L(s_2, s_1)(1 - p_1)\left(\frac{1}{2}\right)^n \tag{1.43}$$

$$\rho(p_S(s), s_2|\mathbf{x}_h(n)) \quad = \quad L(s_1, s_2)p_1; \tag{1.44}$$

and s_1 is chosen if $\rho(p_S(s), s_1|\mathbf{x}_h(n))/\rho(p_S(s), s_2|\mathbf{x}_h(n)) < 1$. Simple manipulation yields the following rule: decide for s_1 if

$$n > \log_2 \frac{p_2}{1 - p_2} + \log_2 \frac{L(s_2, s_1)}{L(s_1, s_2)}. \tag{1.45}$$

This rule quantifies how much "benefit of the doubt" should be granted, as a function of the amount of *a priori* belief p_2 in the fairness of the coin, and on how the player weights the money he is loosing against the friendship which he may loose. Quotients of the type $p/(1 - p)$ are usually known as *odds ratios*. Finally, just to get a concrete number, let us assume that the player was 80% sure about the honesty of his friend, thus $p_2 = 0.8$. Moreover, he weights his friendship to be worth 8 times more than the money he may loose, then $L(s_2, s_1)/L(s_1, s_2) = 8$. With this number, the player should wait until more than five consecutive heads appear before declaring that the coin is double-headed.

End of Example 1.4.3

1.4.5 Discriminant Functions and Decision Regions

In classification problems it is quite common to express the decision rule with the help of the so-called *discriminant functions*. Let us consider an M-class classification problem, with $\mathcal{S} = \mathcal{A} = \{s_1, \ldots, s_M\}$, and a decision rule $\delta(\mathbf{x})$ obtained from some (not necessarily Bayesian) criterion. Consider now a set of M real functions $\{g_i(\mathbf{x}) : \mathcal{X} \to I\!R, \;\; i = 1, 2, .., M\}$, which verify the following relation with $\delta(\mathbf{x})$:

$$\delta(\mathbf{x}) = s_i \Leftrightarrow g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad \forall_{j \neq i}, \tag{1.46}$$

that is, given $\mathbf{x}$, class s_i is chosen when the corresponding discriminant function at that point is greater than all the other ones.

For a classification problem under a "0/1" loss function, the discriminant functions can obviously be set to $g_i(\mathbf{x}) = p_S(s_i/\mathbf{x})$, or to $g_i(\mathbf{x}) = \log\left(p_S(s_i/\mathbf{x})\right)$, or, in fact, to any monotonic increasing function of $p_S(s_i/\mathbf{x})$; any such function will preserve the relative magnitudes of the several $p(s_i/\mathbf{x})$, for $i = 1, 2, ..M$.

The decision rule (or, equivalently, the discriminant functions) partitions the observation space $\mathcal{X}$ into a set of *decision regions* $\{R_i, \; i = 1, \ldots, M\}$, according to

$$\begin{aligned} R_i &= \{\mathbf{x} \in \mathcal{X} : \;\; \delta(\mathbf{x}) = s_i\} & (1.47) \\ &= \{\mathbf{x} \in \mathcal{X} : \;\; g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall_{j \neq i}\}\,. & (1.48) \end{aligned}$$

These regions are mutually disjoint, i.e.

$$R_i \bigcap R_j = \phi, \forall_{i \neq j}, \tag{1.49}$$

with ϕ denoting the empty set, since $\delta(\mathbf{x})$ is a (well defined) function. They are also *exhaustive*, i.e.,

$$\bigcup_{i=1}^{M} R_i = \mathcal{X}, \tag{1.50}$$

because the decision rule is defined for all points of the observation space $\mathcal{X}$.

Example 1.4.4

The decision rule for the binary detection problem of Example 1.4.1 can easily be rewritten in such a way that the decision regions appear explicitly.

After some simple manipulations (assuming $s_0 > s_1$) we can rewrite Eq. (1.29) as

$$\delta_{\text{MAP}}(x) = \begin{cases} s_0 & \Leftarrow \quad x \geq t \\ s_1 & \Leftarrow \quad x < t \end{cases} \tag{1.51}$$

with

$$t = \frac{s_0 + s_1}{2} - \frac{\sigma^2 \log\left(p_1/p_0\right)}{s_1 - s_0}. \tag{1.52}$$

Eqs. (1.51) and (1.52) define a splitting of the real line into two regions $R_0 = \{x : \;\; x \geq t\}$ and $R_1 = \{x : \;\; x < t\}$ separated by a threshold value (see Figure 1.7).

End of Example 1.4.4

Example 1.4.5

Let us now consider a binary decision problem ($\mathcal{S} = \{s_1, s_2\}$), where the two class-conditional likelihoods are bivariate (i.e., $\mathcal{X} = \mathbb{R}^2$) Gaussian with different covariances, i.e., $f_{\mathbf{X}}(\mathbf{x}|s_1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \mathbf{C}_1)$ and $f_{\mathbf{X}}(\mathbf{x}|s_2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \mathbf{C}_2)$ (see Eq. (1.33)). Let $\boldsymbol{\mu}_1 = [3 \;\; 3]^T$, $\boldsymbol{\mu}_2 = [6 \;\; 6]^T$, and

$$\mathbf{C}_1 = \begin{bmatrix} 1.2 & -0.4 \\ -0.4 & 1.2 \end{bmatrix} \qquad \mathbf{C}_1 = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 1.2 \end{bmatrix}. \tag{1.53}$$

The two corresponding class-conditional densities are plotted in Figure 1.2 (of course these functions are defined for $\mathbf{x} \in \mathbb{R}^2$, not just on the square region shown in the plots).

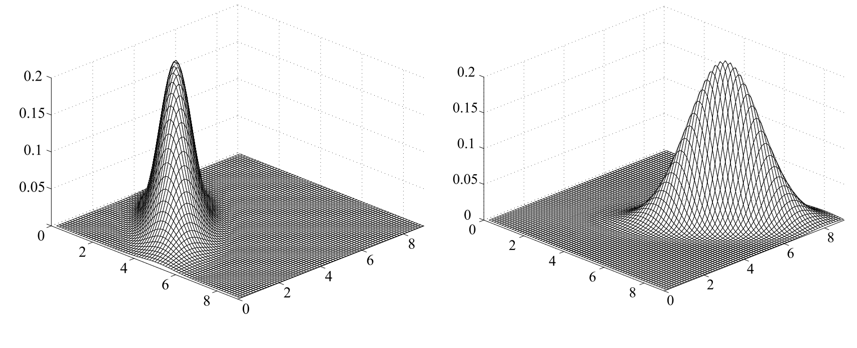

FIGURE 1.2. The two class-conditional Gaussian densities considered in Example 1.4.5 (see text).

To gain some insight into the aspect of the decision regions and into how they depend on the *a priori* probabilities $p_S(s_1) \equiv p_1$ (notice that $p_S(s_2) = 1 - p_S(s_1)$), Figure 1.3 plots the boundary between R_1 and R_2 for several values of p_1. The decision region R_1 includes all the points below the boundary while R_2 is its complement. Notice how increasing p_1 pushes the boundary away from $\boldsymbol{\mu}_1$; observe also that these boundaries are not straight lines, which is due to the fact that the two covariance matrices are different.

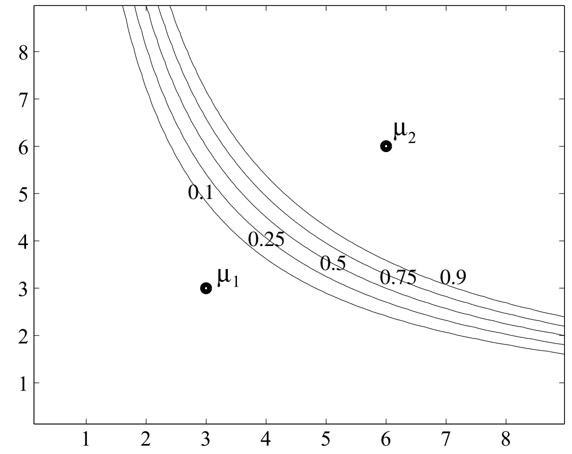

FIGURE 1.3. Boundary between the two decision regions in Example 1.4.5, for several values of the *a priori* probability $p_S(s_1)$.

End of Example 1.4.5

In the case of the linear decision rules for M-ary classification problems with Gaussian class-conditional densities of common covariance, studied in Section 1.4.3, the decision regions can be obtained explicitly. Notice that in these cases, the discriminant functions can be directly identified in Eqs. (1.36) and (1.38) and have the general form

$$g_i(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\alpha}_i + \beta_i. \tag{1.54}$$

To see how the decision regions look, we start by making the observation that any M-ary decision problem can be solved by solving all the partial 2-class problems it contains; in other words, if it is always possible to choose a winner from any pair of hypotheses, then it is possible to find a final winner (for example, in a 4-class problem, if a is better than b, and c is better than d, and c is better than a, then c is better than a, b, and d). So let us first understand which kind of decision regions are associated with any 2-class problem of the type: choose s_i or s_j? Clearly, s_i is chosen over s_j when $g_i(\mathbf{x}) > g_j(\mathbf{x})$, i.e., when

$$\mathbf{x}^T(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j) + (\beta_i - \beta_j) > 0. \tag{1.55}$$

This condition can be rewritten as (see [38])

$$(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j)^T(\mathbf{x} - \mathbf{x}_0) > 0, \tag{1.56}$$

where

$$\mathbf{x}_0 = \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} - \frac{\sigma^2(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j)}{\|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j\|^2} \log \frac{p_S(s_i)}{p_S(s_j)}. \tag{1.57}$$

The inequality in Eq. (1.56) splits the observation space $\mathcal{X} = I\!\!R^n$ into two semi-spaces separated by a hyperplane; this hyperplane is perpendicular to the straight line from $\boldsymbol{\alpha}_i$ to $\boldsymbol{\alpha}_j$ and contains point $\mathbf{x}_0$ which is somewhere on that line. Notice that if $p_S(s_i) = p_S(s_j)$, $\mathbf{x}_0$ is located halfway between $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$; increasing $p_S(s_i)$ pulls the hyperplane away from $\boldsymbol{\alpha}_i$ making the choice for s_i more likely. Notice that Eq. (1.52) was simply a scalar version of Eq. (1.57). Finally, putting all these 2-class criteria together to yield the complete M-ary decision rule, shows that the decision regions are intersections of semi-spaces separated by hyperplanes, resulting in piecewise hyper-planar boundaries. A deeper look at these decision regions and separating hyperplanes can be found in several pattern recognition references; see, e.g., [38] for several very interesting illustrations and examples. For illustration purposes only, we include some simple examples.

Example 1.4.6

Let us now return to the binary decision problem of Example 1.4.5, but now letting the class-conditional covariance matrices be equal

$$\mathbf{C}_1 = \mathbf{C}_2 = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 1.0 \end{bmatrix} \tag{1.58}$$

The corresponding class-conditional densities are now plotted in Figure 1.4.

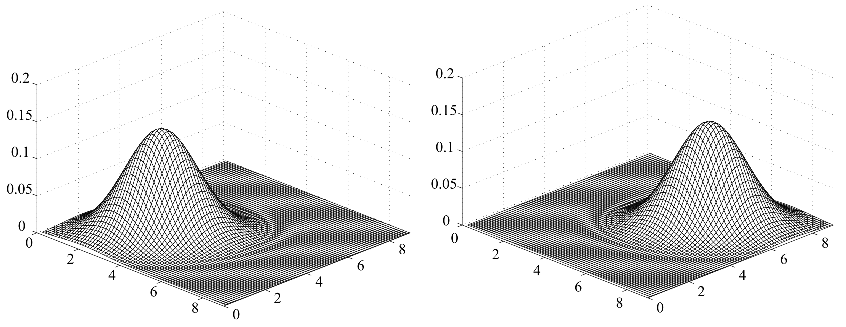

FIGURE 1.4. The two class-conditional Gaussian densities considered in Example 1.4.6 (see text).

The boundary between the decision regions (for several values of $p_S(s_1)$) is plotted in Figure 1.5. As in Example 1.4.5, R_1 includes all the points below the boundary while R_2 is its complement. Notice that these boundaries are now straight lines, as a consequence of the fact that the two covariance matrices are equal. In higher dimensions, the boundaries are hyper-planes (straight-lines, in 2 dimensions).

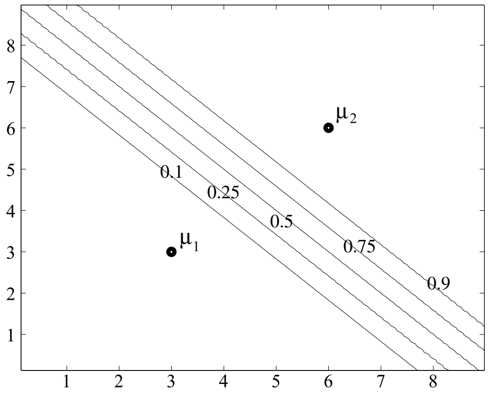

FIGURE 1.5. Boundary between the two decision regions in Example 1.4.6, for several values of the *a priori* probability $p_S(s_1)$.

End of Example 1.4.6

Example 1.4.7

In M-ary classification problems where all the observations are independent and have common variance (that is, all class-conditional covariance matrices are equal to $\sigma^2\mathbf{I}$), we have seen in Eq. (1.35) that the classifier finds a compromise between selecting the class whose mean is closer to the observed data vector and whose *a priori* probability is higher. In the case where all the classes are *a priori* equiprobable, this reduces to a *nearest class*, or *minimum distance*, classification rule; the corresponding decision regions are in this case defined as

$$R_i = \{\mathbf{x} : \| \mathbf{x} - \boldsymbol{\mu}_i \| < \| \mathbf{x} - \boldsymbol{\mu}_j \|, \; j \neq i\}.$$

As we have just seen, since the covariance is the same for all classes, the boundaries between these regions are hyper-planar and define what is known as a Voronoi partition; the corresponding regions are called Voronoi regions. In Figure 1.6, an example of such a partition is shown; it is a 10-class problem, and the class-conditional means are $\boldsymbol{\mu}_1 = [1,0]^T$, $\boldsymbol{\mu}_2 = [0,1]^T$, $\boldsymbol{\mu}_3 = [-1,0]^T$, $\boldsymbol{\mu}_4 = [0,-1]^T$, $\boldsymbol{\mu}_5 = [2,2]^T$, $\boldsymbol{\mu}_6 = [-2,-2]^T$, $\boldsymbol{\mu}_7 = [-2,2]^T$, $\boldsymbol{\mu}_8 = [2,-2]^T$, $\boldsymbol{\mu}_9 = [4,4]^T$, and $\boldsymbol{\mu}_{10} = [-4,-4]^T$.

End of Example 1.4.7

The Gaussian observation model is not the only one leading to linear discriminants and linear classifiers; in the next two examples we consider two other such cases.

Example 1.4.8

Consider the problem of deciding which one among a given set of radioactive substances $\mathcal{S} = \{s_1, s_2, ..., s_M\}$ is contained in a given sample. To this

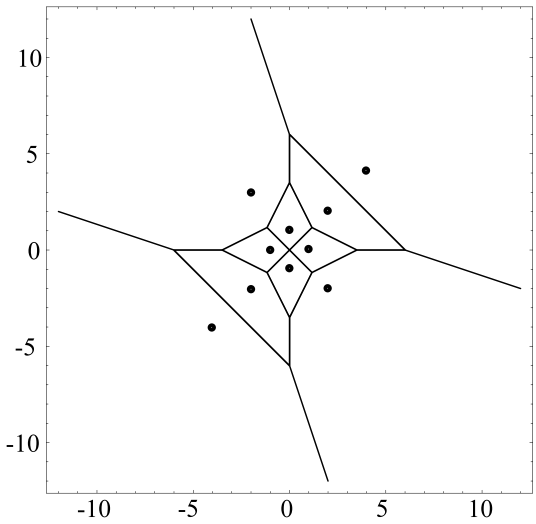

FIGURE 1.6. Voronoi regions corresponding to the points listed in Example 1.4.7; the black dots are located at the class-conditional means.

end, a set of n measurements is performed. In each of these, the number of emissions x_j is counted during a period T_j (seconds), for $j = 1, 2, ..., n$, with the intervals T_j being possibly different from each other. Each substance is characterized by its emission rate λ_i (emissions/second), and it is known that the emission process obeys a Poisson distribution (see Appendix A). Letting $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ denote the vector of counts (the observed data), assumed mutually independent, the class-conditional probability functions are

$$f_{\mathbf{X}}(\mathbf{x}|s_i) = \prod_{j=1}^{n} \frac{e^{-\lambda_i T_j} (\lambda_i T_j)^{x_j}}{x_j!}. \tag{1.59}$$

Inserting this into the (natural) logarithmic version of the MAP classifier (Eq. (1.27)) yields (after dropping constants)

$$\delta_{\text{MAP}}(\mathbf{x}) = \arg\max_{s_i \in \mathcal{S}} \left\{ \log(\lambda_i) \sum_{j=1}^{n} x_j - \lambda_i T + \log p_S(s_i) \right\}, \tag{1.60}$$

where $T = T_1 + T_2 + ... + T_n$ is the total observation time; these are clearly linear discriminant functions.

End of Example 1.4.8

Example 1.4.9

A final example of a decision problem involving linear discriminants considers a set of M (not necessarily fair) coins $\mathcal{S} = \{s_1, s_2, ..., s_M\}$; each of these coins is characterized by its probability of heads outcomes θ_i (assumed known). The objective is to decide which of the coins was used in an observed sequence of n tosses, whose outcomes are denoted as $\mathbf{x} = (x_1, x_2, ..., x_n)$; here $x_i = 1$ and $x_i = 0$ represent heads and tails outcomes, respectively. The n tosses being independent, the observation model is

$$f_{\mathbf{X}}(\mathbf{x}|s_i) = \prod_{j=1}^{n} \theta_i^{x_j} (1-\theta_i)^{1-x_j} = \theta_i^h (1-\theta_i)^{n-h} \tag{1.61}$$

which is a Bernoulli distribution, where $h = x_1 + x_2 + ... + x_n$ is the total number of heads outcomes. Again using logarithms, we can write the MAP classifier as

$$\delta_{\text{MAP}}(\mathbf{x}) = \arg\max_{s_i \in \mathcal{S}} \left\{ h \log \frac{\theta_i}{1-\theta_i} + n \log(1-\theta_i) + \log p_S(s_i) \right\}, \tag{1.62}$$

which is linear in h (the observed heads count).

End of Example 1.4.9

1.4.6 Error Probabilities

Let us consider the Bayes' risk (or integrated risk), as defined in Eq. (1.15), for a binary classification problem (i.e., a problem where $\mathcal{A} = \mathcal{S} = \{s_1, s_2\}$ and $p_S(s_1) = 1 - p_S(s_2)$) for which a "0/1" loss function is adopted. This risk can be rewritten as

$$\begin{aligned} r(p(s), \delta(\mathbf{x})) &= \int_{\mathcal{X}} [L(s_1, \delta(\mathbf{x}))\, p(s_1|\mathbf{x})\, p(\mathbf{x}) \\ &\qquad + L(s_2, \delta(\mathbf{x}))\, p(s_2|\mathbf{x})\, p(\mathbf{x})]\; d\mathbf{x} \\ &= \int_{R_2} p(s_1|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} + \int_{R_1} p(s_2|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} \\ &= p(s_1) \int_{R_2} p(\mathbf{x}|s_1)\, d\mathbf{x} + p(s_2) \int_{R_1} p(\mathbf{x}|s_2)\, d\mathbf{x}, \end{aligned} \tag{1.63}$$

where R_1 and R_2 are the decision regions associated with s_1 and s_2, respectively. The two integrals in Eq. (1.63) have the following clear meaning: they are the probabilities of making incorrect decisions (errors) conditioned on each of the two possible states of nature. In fact, the first integral in Eq. (1.63) simply expresses the probability that an observation produced under s_1 will fall inside R_2 which is the decision region associated with s_2; the second integral quantifies the probability of the second type of error. Accordingly, Eq. (1.63) can be rewritten as

$$\begin{aligned} r(p(s), \delta(\mathbf{x})) &= P(\text{“error”}|s_1)\, p(s_1) + P(\text{“error”}|s_2)\, p(s_2) \\ &= P(\text{“error”}). \end{aligned} \tag{1.64}$$

This interpretation can be extended to M-ary classification problems, under the "0/1" loss function, with Eq. (1.63) generalizing to

$$\begin{aligned} r(p(s,\delta(\mathbf{x}))) &= \sum_{i=1}^{M} p(s_i) \int_{R_i^C} p(\mathbf{x}|s_i)\, d\mathbf{x} \\ &= \sum_{i=1}^{M} P(\text{“error”}|s_i)\, p(s_i) \\ &= P(\text{“error”}), \end{aligned} \tag{1.65}$$

where R_i^C denotes the complement of the decision region R_i, i.e.

$$R_i^C = \bigcup_{j \neq i} R_j. \tag{1.66}$$

It can then be stated that the Bayes' risk associated with the "0/1" loss function equals the probability of error; as an immediate corollary, a decision rule minimizing the Bayes' risk (and consequently the *a posteriori* expected loss) under a "0/1" loss function is also minimizing the probability of error. This is, of course, not at all surprising: the "0/1" loss function can be seen as the *indicator function* of the "error" event[2]

$$\text{“error”} = \{(s,\mathbf{x}) \in \mathcal{S} \times \mathcal{X} : \delta(\mathbf{x}) \neq s\} \subseteq \mathcal{S} \times \mathcal{X}, \tag{1.67}$$

because

$$L(s,\delta(\mathbf{x})) = \begin{cases} 1 & \Leftarrow \quad (s,\mathbf{x}) \in \text{ “error”} \\ 0 & \Leftarrow \quad (s,\mathbf{x}) \notin \text{ “error”}, \end{cases} \tag{1.68}$$

and the probability of an event is equal to the expected value of its indicator function (see Appendix A).

For general (not necessarily binary) problems ($M \geq 2$), it is sometimes simpler to obtain the probability of error via $P(\text{“error”}) = 1 - P(\text{“correct decision”})$, with

$$\begin{aligned} P(\text{“correct decision”}) &= \sum_{i=1}^{M} P(\text{“correct decision”}|s_i)\, p(s_i) \\ &= \sum_{i=1}^{M} p(s_i) \int_{R_i} p(\mathbf{x}|s_i)\, d\mathbf{x} \\ &= \sum_{i=1}^{M} \int_{R_i} p(s_i|\mathbf{x})\, p(\mathbf{x})\; d\mathbf{x} \\ &= \int_{\mathcal{X}} \left(\max_{i=1,\dots,M} p(s_i|\mathbf{x}) \right) p(\mathbf{x})\, d\mathbf{x}, \end{aligned} \tag{1.69}$$

[2] Recall that an event is a subset of the sample space, and that an indicator function for a given set is one that equals one for points in that set, and zero for points outside the set.

where the fact that $R_i = \{\mathbf{x} : \; p(s_i|\mathbf{x}) > p(s_j|\mathbf{x}), \; j \neq i\}$ was used.

The probability of error is an important and useful characterization of the performance of a decision rule for a given classification problem. However, the probability of error is very difficult to compute, and exact closed-form expressions are only attainable in some very simple situations. Closed-form expressions for P(“correct decision”) can only be obtained (even in the binary case) in very simple and particular scenarios, with the nonlinear term ($\max p(s_i|\mathbf{x})$) being mainly responsible for this difficulty. This has stimulated a considerable amount of research in the derivation of bounds and approximations for this term (see standard texts on statistical pattern recognition, such as [75] or [45], and [4] for more recent results and further references).

Example 1.4.10

For a scalar Gaussian observation, these probabilities of error can be easily visualized. Let us return to the the binary detection/classification problem in Example 1.4.1, with (*a priori*) probabilities p_0 and $p_1 = 1 - p_0$. Recall that the likelihood functions are $p(x|s_0) = \mathcal{N}(x|s_0, \sigma^2)$ and $p(x|s_1) = \mathcal{N}(x|s_1, \sigma^2)$; the decision regions in this case are given by (see Example 1.4.4) $R_0 = \{x : \;\; x \geq t\}$ and $R_1 = \{x : \;\; x < t\}$, with the threshold t being given by Eq. (1.52). It is now easy to write, from Eq. (1.63),

$$\begin{aligned} P(\text{“error”}) &= p_1\, P(\text{“error”}|s_1) + p_0\, P(\text{“error”}|s_0) \\ &= p_1 \int_t^{+\infty} p(x|s_1) + p_0 \int_{-\infty}^{t} p(x|s_0). \end{aligned} \qquad (1.70)$$

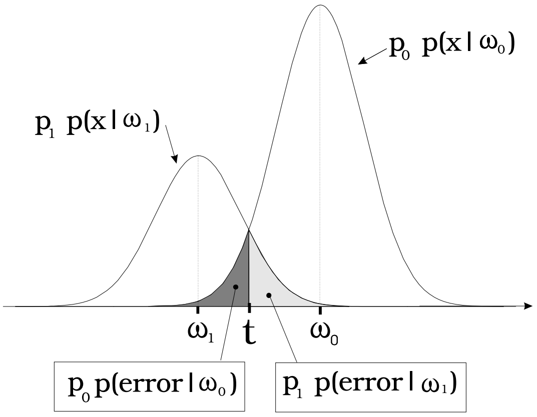

FIGURE 1.7. Illustration of the probabilities of error for a simple binary classification problem with Gaussian likelihood functions.

This probability of error is given by the sum of the two shaded areas depicted in Figure 1.7; the different heights of the two Gaussians means that

$p_0 > p_1$. Notice how any other value of t would lead to a higher probability of error: for example, if we move t to the left, $p_0 P(\text{“error”}|s_0)$ decreases by some amount but $p_1 P(\text{“error”}|s_1)$ increases by a larger amount. It is simple to show that the optimal threshold given by Eq. (1.52) is the abscissa where the two Gaussian functions intersect each other, as suggested by the previous argument.

Notice that even in this very simple case, it is not possible to obtain an elementary analytical expression for the probability of error; all that can be done is rewrite Eq. (1.70) as

$$P(\text{“error”}) = p_1 \operatorname{erfc}\left(\frac{t - s_1}{\sigma}\right) + p_0 \operatorname{erfc}\left(\frac{s_0 - t}{\sigma}\right), \qquad (1.71)$$

where

$$\operatorname{erfc}(x) \equiv \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\frac{u^2}{2}}\, du \qquad (1.72)$$

is called the *complementary error function.* This function, although only computable through numerical integration, is widely available in tables and in most mathematical software libraries and packages. One important particular instance of Eq. (1.71) arises when $p_0 = p_1$; in that case (see Eq. (1.52)) the threshold becomes $t = (s_0 + s_1)/2$ which implies that $(t - s_1) = (s_0 - t) = (s_0 - s_1)/2$ and Eq. (1.71) simplifies to $P(\text{“error”}) = \operatorname{erfc}((s_0 - s_1)/2\sigma)$ which is plotted in Figure 1.8. Notice how it rapidly plunges to very small values as soon as its arguments becomes larger than one (i.e., when the difference between s_0 and s_1 is larger than twice the noise standard deviation). Observe also that as the quotient between $(s_0 - s_1)$ and σ approaches zero, the probability of error approaches 1/2 (not 1). In the communications literature, the quotient $(s_0 - s_1)/\sigma$ is usually called the *signal to noise ratio* (SNR) and measured in dB according to $\text{SNR}_{\text{dB}} = 20 \log_{10}(s_0 - s_1)/\sigma$.

End of Example 1.4.10

Example 1.4.11

We now look once more at the situation described in Example 1.4.2. Let us consider that bit “0” is represented by a null (length n) sequence, $\boldsymbol{\mu}_0 = [0, 0, ...0]^T$, while bit “1” corresponds to a simple constant sequence, say, $\boldsymbol{\mu}_1 = [A, A, ..., A]^T$. As before, the channel adds (to all transmitted values) independent Gaussian noise samples of zero mean and variance σ^2, and the sequence of received values is denoted as $\mathbf{x} = [x_1, x_2, ..., x_n]^T$. The optimal Bayesian detector is linear and is given by Eq. (1.36); some simple manipulations allows representing this classification rule in a simpler form

$$\delta_{\text{MAP}}(\mathbf{x}) = \begin{cases} \text{bit “0”} & \Leftarrow \quad l(\mathbf{x}) \le t \\ \text{bit “1”} & \Leftarrow \quad l(\mathbf{x}) > t \end{cases} \qquad (1.73)$$

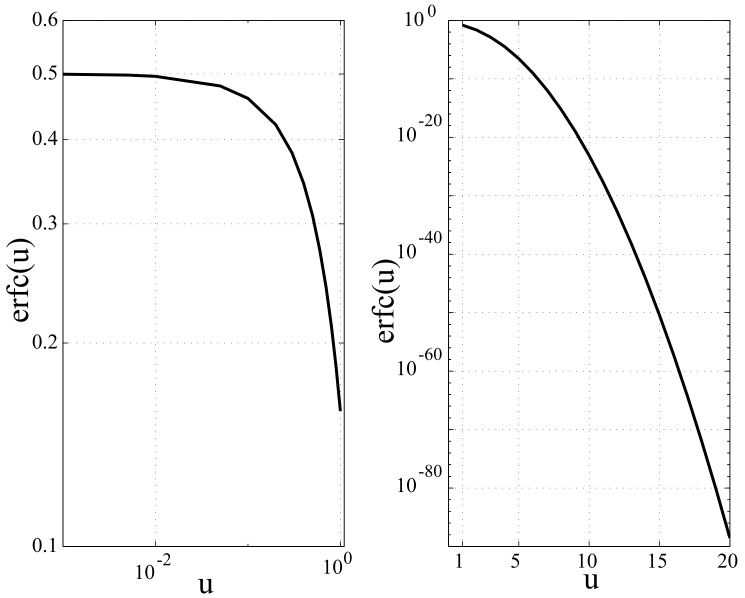

FIGURE 1.8. The erfc(u) function. Left plot: from $u = 10^{-3}$ to $u = 1$; on the right: for u from 1 to 20.

where

$$l(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1.74}$$

and the threshold t is

$$t = \frac{A}{2} + \frac{\sigma^2}{nA}\log\frac{p_0}{p_1}. \tag{1.75}$$

The meaning of this criterion is clear: if $p_0 = p_1$, the sample average of the observations is being compared with $A/2$; increasing p_0 will increase this threshold, pushing it away from bit "0", or vice-versa. To compute the probability of error, let us assume that $p_0 = p_1$; then, the threshold is simply $A/2$, and $P(\text{“error”}) = P(\text{“error”}|\text{bit “0”}) = P(\text{“error”}|\text{bit “1”})$, where

$$P(\text{“error”}|\text{bit “0”}) = P\left[l(\mathbf{x}) > t|\text{bit “0”}\right] = \int_{\frac{A}{2}}^{\infty} f_T(t|\text{bit “0”})dt. \tag{1.76}$$

Since t is the sample mean of the observations given that bit "0" was sent, it is a sample of a Gaussian random variable of zero mean and variance σ^2/n, and Eq. (1.76) can be rewritten as an erfc($\cdot$) function

$$P(\text{“error”}) = \text{erfc}\left(\frac{A\sqrt{n}}{2\sigma}\right) \tag{1.77}$$

(see Figure 1.8), whose meaning is clear: if the length n of the "signal" increases, the probability of error decreases. This type of problem has been exhaustively studied in the communications literature, where being able to compute error probabilities is of extreme importance.

End of Example 1.4.11

Recall that when the class-conditional densities are Gaussian with a common covariance matrix $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$, the resulting classifier can be expressed via a scalar linear function of the observations (Eq. (1.36)). This fact allows obtaining a simple expression for the probability of error in the two-class (binary) case, exploiting the fact that the linear discriminant function is, for each class, still a Gaussian random variable. Omitting the details (see, for example, [45] or [38]), it is easy to show that the probability of error is given by:

$$P(\text{``}error\text{''}) = \operatorname{erfc}\left(\frac{d}{2}\right) \tag{1.78}$$

where d is the Mahalanobis distance between the two classes:

$$d = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\mathbf{C}} = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}.$$

Notice that Eq. (1.77) is nothing but a particular case of Eq. (1.78) because with $\mathbf{C} = \sigma^2\mathbf{I}$, $\boldsymbol{\mu}_1 = \mathbf{0}$, and $\boldsymbol{\mu}_2 = [A, A, ..., A]^T$ (n-dimensional), then

$$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\mathbf{C}} = \frac{A\sqrt{n}}{\sigma}.$$

1.5 Bayesian Estimation

1.5.1 Introduction

In estimation problems, nature takes values in a continuous set; for now, we will only consider scalar states of nature, $\mathcal{S} \subseteq I\!R$. The goal is, of course, to decide which is the true state of nature (i.e., $\mathcal{A} = \mathcal{S}$) from the observed data $\mathbf{x}$. The prior $p_S(s)$ here is a probability density function because $\mathcal{S}$ is continuous, and the denominator in Bayes' theorem appears in its continuous version (see Eq. (1.9)); again, this is true regardless of $f_{\mathbf{X}}(\mathbf{x})$ and $f_X(\mathbf{x}|s)$ being probability density (continuous $\mathcal{X}$) or mass (discrete $\mathcal{X}$) functions. In this section we review the three most widely used (partly for mathematical tractability reasons) and well studied loss functions for estimation problems: the "0/1" (also called *uniform*), the *"quadratic error"*, and the *"absolute error"* loss functions. These three loss functions are depicted in Figure 1.9.

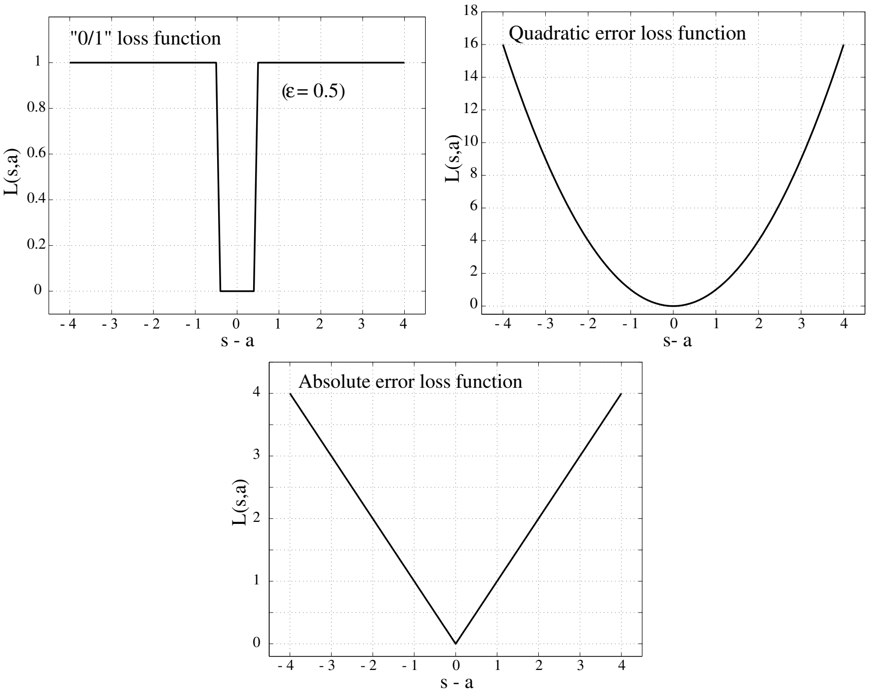

FIGURE 1.9. The three classical loss functions.

1.5.2 The "0/1" loss function

The "0/1" loss function for estimation problems is defined with the help of a parameter ε

$$L_\varepsilon(s,a) = \left\{ \begin{array}{ccc} 1 & \Leftarrow & |s-a| \geq \varepsilon \\ 0 & \Leftarrow & |s-a| < \varepsilon, \end{array} \right. ; \tag{1.79}$$

see Figure 1.9. The resulting Bayes decision rule (see Eqs. (1.8) and (1.11)) under this loss function becomes

$$\begin{aligned} \delta_\varepsilon(\mathbf{x}) &= \arg\min_{d\in\mathcal{S}} \int_{s\in\mathcal{S}} L_\varepsilon(s,d) p_S(s|\mathbf{x})\, ds \\ &= \arg\min_{d\in\mathcal{S}} \left(1 - \int_{s:|s-d|<\varepsilon} p_S(s|\mathbf{x})\, ds\right) \\ &= \arg\max_{d\in\mathcal{S}} \int_{d-\varepsilon}^{d+\varepsilon} p_S(s|\mathbf{x})\, ds. \end{aligned} \tag{1.80}$$

The meaning of the integral in this expression is depicted in Figure 1.10 (a); for some finite ε, this estimation criterion returns a value such that the probability of finding S inside an interval of width 2ε around it is maximal.

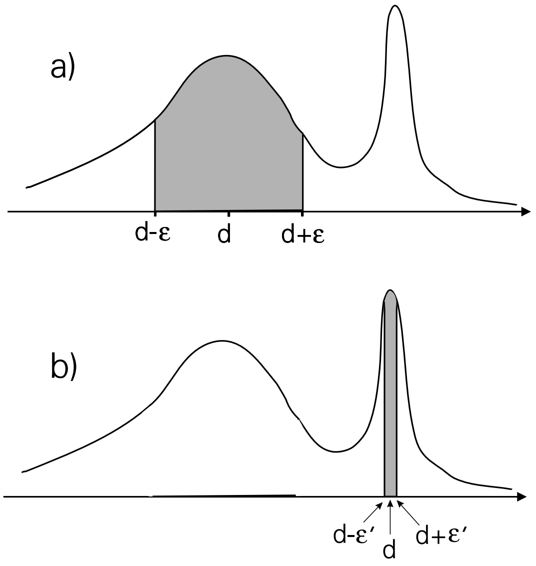

FIGURE 1.10. Illustration of how an estimate under the L_ε loss function converges to the higher mode of the *a posteriori* probability density function, as the width of the considered interval vanishes ($\varepsilon' \ll \varepsilon$).

Of special interest is the case of infinitesimal ε,

$$\lim_{\varepsilon \to 0} \arg\max_{d \in \mathcal{S}} \int_{d-\varepsilon}^{d+\varepsilon} p_S(s|\mathbf{x})\, ds = \arg\max_{d \in \mathcal{S}} p_S(d|\mathbf{x}) \equiv \delta_{\mathrm{MAP}}(\mathbf{x}) \tag{1.81}$$

which is called the *maximum a posteriori* (MAP) estimator. Figure 1.10 illustrates what happens as constant ε decreases; from (a) to (b), as ε decreases, higher modes of the *a posteriori* density are chosen, and their widths loose importance. In the limit, as ε goes to zero, the highest mode (or one of them if there are several with the same height) is chosen, regardless of its width which may be arbitrarily small.

Since for a given observation $\mathbf{x}$, the marginal $f_{\mathbf{X}}(\mathbf{x})$ is a constant (just as in classification problems), the MAP estimator also often appears under the following alternative forms

$$\begin{aligned} \delta_{\mathrm{MAP}}(\mathbf{x}) &= \arg\max_{s \in \mathcal{S}} \left\{ f_{\mathbf{X}}(\mathbf{x}|s) p_S(s) \right\} && (1.82)\\ &= \arg\max_{s \in \mathcal{S}} \left\{ \log f_{\mathbf{X}}(\mathbf{x}|s) + \log p_S(s) \right\} && (1.83)\\ &= \arg\max_{s \in \mathcal{S}} \left\{ f_{\mathbf{X},S}(\mathbf{x}, s) \right\}. && (1.84) \end{aligned}$$

all obviously equivalent.

Example 1.5.1

A classical example that will give some further insight into MAP estimation is the one where both the prior and the likelihood are Gaussian. More specifically, let us consider the goal of estimating an unknown real quantity s from a single measurement (an observation) x which is related to s via a Gaussian model of known variance σ^2, $f_X(x|s) = \mathcal{N}(x|s, \sigma^2)$ (this is a very common model for measurement errors). Moreover, knowledge about s is modeled by the prior $p_S(s) = \mathcal{N}(s|s_0, \phi^2)$. Simple manipulation allows showing that the posterior is still a Gaussian with mode and mean at

$$\delta_{\mathrm{MAP}}(\mathbf{x}) = \frac{\frac{s_0}{\phi^2} + \frac{x}{\sigma^2}}{\dfrac{1}{\phi^2} + \dfrac{1}{\sigma^2}} = \frac{s_0\sigma^2 + x\phi^2}{\sigma^2 + \phi^2}; \tag{1.85}$$

the interpretation of Eq. (1.85) is clear: the estimate is a compromise, in the form of a weighted mean, between the observed value x and the *expected* prior mean s_0; the weights are functions of the degree of confidence placed on each of these two elements, as measured by the inverse of the respective variances. In order to take one step further this view of the variance as a measure of how much to trust some belief, let us look at the variance of the *a posteriori* p.d.f. which is given by

$$E_S\left[(s - \delta_{\mathrm{MAP}}(\mathbf{x}))^2|x\right] = \frac{\sigma^2\phi^2}{\sigma^2 + \phi^2} = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\phi^2}}; \tag{1.86}$$

The readers familiar with basic electric circuit theory, will recognize that this expression is similar to the one giving the resistance of two resistors connected in parallel. One important property of this expression is that

$$\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\phi^2}} < \min\{\sigma^2, \phi^2\} \tag{1.87}$$

which can be interpreted as follows: the MAP estimate is more trustful then either the observations or the prior alone.

End of Example 1.5.1

Example 1.5.2

Now consider that, rather than just one, there are n independent and identically distributed (i.i.d.) observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$; the likelihood can now be written as

$$f_{\mathbf{X}}(\mathbf{x}|s) \propto \prod_{i=1}^{n} \exp\left\{-\frac{(x_i - s)^2}{2\sigma^2}\right\} = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - s)^2\right\} \tag{1.88}$$

and the corresponding MAP estimate becomes

$$\delta_{\mathrm{MAP}}(\mathbf{x}) = \frac{\dfrac{s_0}{\phi^2} + \dfrac{\bar{x}n}{\sigma^2}}{\dfrac{1}{\phi^2} + \dfrac{n}{\sigma^2}} = \frac{s_0\frac{\sigma^2}{n} + \bar{x}\phi^2}{\frac{\sigma^2}{n} + \phi^2}, \tag{1.89}$$

where $\bar{x} \equiv (x_1 + x_2 + ... + x_n)/n$ is the sample mean of the observations. Clearly, as the number of observations increases, the role of the prior decreases; it is clear from Eq. (1.89) that $\lim_{n\to\infty} \delta_{\text{MAP}}(\mathbf{x}) = \bar{x}$. This fact is illustrated in Figure 1.11; there, $s = s_0 = 0$, i.e., the prior mean coincides with the true parameter value. Two different values for ϕ^2 were considered: 0.1 and 0.01, expressing two different degrees of confidence on the belief that s should be close to zero. Notice how the MAP estimates are compromises between the observed data, which determines $\bar{x}$, and the prior knowledge that s should be near zero. Also, observe how the MAP estimator with $\phi = 0.01$ is more strongly tied to the prior mean.

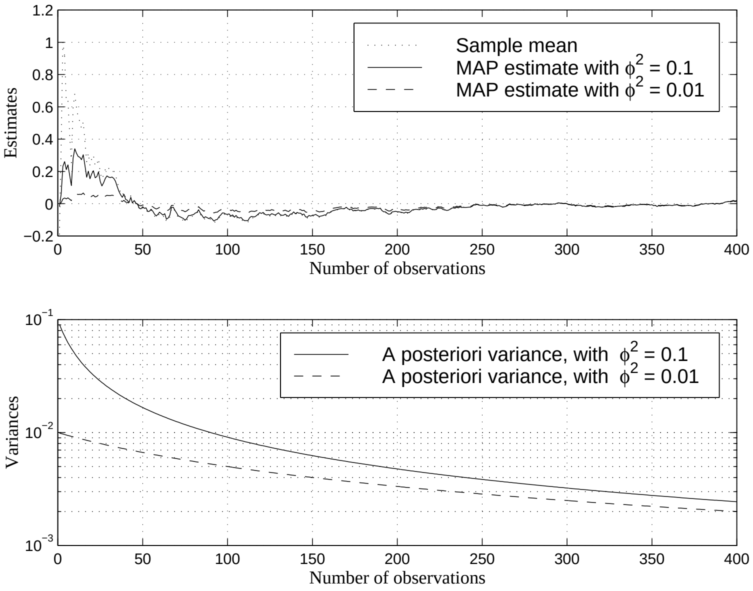

FIGURE 1.11. Upper plot shows the evolution of $\bar{x}$ (dotted line) and $\delta_{\text{MAP}}(\mathbf{x})$ for two values of the prior variance ($\phi^2 = 0.1$, solid line, and $\phi^2 = 0.01$, dashed line) versus the total number of observations n. Lower graph shows the evolution of the *a posteriori* variances ($\phi^2 = 0.1$, solid line, and $\phi^2 = 0.01$, dashed line)

Here again we can look at the *a posteriori* variance, which has a purely deterministic evolution with n,

$$E_S\left[(s - \delta_{\text{MAP}}(\mathbf{x}))^2|\mathbf{x}\right] = \frac{\sigma^2\phi^2}{\sigma^2 + n\phi^2} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\phi^2}} \tag{1.90}$$

(compare with Eq. (1.86)), as shown in the lower plot in Figure 1.11. The term n/σ^2, which can be interpreted as the degree of confidence of the observations, grows linearly with n, this being a reasonable property; as a consequence, as $n \to \infty$, the *a posteriori* variance goes to zero, which means that when we have a very large amount of data we may absolutely trust the estimate, and the influence of the prior vanishes.

End of Example 1.5.2

Example 1.5.3

Let us now consider a set of n, not necessarily independent, observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, whose possible dependence is expressed by a known covariance matrix $\mathbf{C}$, with common unknown (to be estimated) mean s:

$$f_{\mathbf{X}}(\mathbf{x}|s) = ((2\pi)^n \det(\mathbf{C}))^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - s\mathbf{u})^T \mathbf{C}^{-1}(\mathbf{x} - s\mathbf{u})\right\}, \quad (1.91)$$

where $\mathbf{u} = [1, 1, ..., 1]^T$. The adopted prior, as above, is $p_S(s) = \mathcal{N}(s|s_0, \phi^2)$. Some computations lead to the following MAP estimate:

$$\delta_{\text{MAP}}(\mathbf{x}) = \left(\mathbf{u}\mathbf{C}^{-1}\mathbf{u} + \frac{1}{\phi^2}\right)^{-1} \left(\mathbf{x}^T\mathbf{C}^{-1}\mathbf{u} + \frac{s_0}{\phi^2}\right). \quad (1.92)$$

This is still a weighted average of the observations and the prior expected value s_0; to make this fact more obvious, let us denote the inverse of the covariance matrix as $\mathbf{D} = \mathbf{C}^{-1}$, and define a vector $\mathbf{w} = \mathbf{D}\mathbf{u}$; notice that the $i-$th element of $\mathbf{w}$ equals the sum of all the elements in row i of matrix $\mathbf{D}$. With this notation, Eq. (1.92) can be rewritten to clearly show that it expresses a weighted average:

$$\delta_{\text{MAP}}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{w} + \frac{s_0}{\phi^2}}{\mathbf{u}^T\mathbf{w} + \frac{1}{\phi^2}} = \frac{\sum_{i=1}^{n} x_i w_i + \frac{s_0}{\phi^2}}{\sum_{i=1}^{n} w_i + \frac{1}{\phi^2}}. \quad (1.93)$$

Three special cases help understanding the meaning of these weights:

- Firstly, if $\mathbf{C} = \sigma^2\mathbf{I}$ then we have the situation studied in the previous example, since $\mathbf{D} = \mathbf{I}/\sigma^2$ and so $w_i = 1/\sigma^2$.

- If $\mathbf{C}$ is diagonal but with different elements, $\mathbf{C} = \text{diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_n^2\}$, this means that the observations are independent but each has its own variance; in this case, $\mathbf{D} = \text{diag}\{(\sigma_1^2)^{-1}, (\sigma_2^2)^{-1}, ..., (\sigma_n^2)^{-1}\}$ and so $w_i = 1/\sigma_i^2$, i.e., each weight is inversely proportional to the variance of the corresponding observation, which is intuitively reasonable.

- Finally, to study the general case, let us focus on $n = 2$ for which simple results can be obtained: the inverse of the covariance matrix of a bivariate Gaussian can be written as (see Appendix A)

$$\mathbf{D} = \frac{1}{1-\rho^2}\left[\begin{array}{cc} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}} \\ -\frac{\rho}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}} & \frac{1}{\sigma_2^2} \end{array}\right] \quad (1.94)$$

where $\rho \in [-1, +1]$ is the *correlation coefficient* which measures the degree of dependence between the two components of the random vector. So,

$$w_i = \frac{1}{1-\rho^2}\left(\frac{1}{\sigma_i^2} - \frac{\rho}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}}\right), \quad i = 1, 2. \tag{1.95}$$

If the variables are uncorrelated, $\rho = 0$, the weights coincide with those of the independent observations case ($w_i = 1/\sigma_i^2$). To study the influence of non-zero correlation on these weights, let us further assume that both observations have the same variance, say $\sigma_1^2 = \sigma_2^2 = \sigma^2$; this assumption leads to

$$\delta_{\text{MAP}}(\mathbf{x}) = \frac{\dfrac{x_1 + x_2}{\sigma^2(1+\rho)} + \dfrac{s_0}{\phi^2}}{\dfrac{2}{\sigma^2(1+\rho)} + \dfrac{1}{\phi^2}}. \tag{1.96}$$

Now, if $\rho \to 1$, this means that X_2 tends to follow X_1 (or vice versa) deterministically and in the same direction (i.e., when one is larger than the mean, so is the other); consequently, both observations carry almost the same information as just one of them, and therefore (in the limit) each gets one half of the weight it would get in the independent case. On the other hand, if $\rho \to -1$, this means that X_2 still tends to accompany X_1 deterministically but now on opposite directions with respect to their common mean; i.e., when $x_1 = s + d$, there is high probability of $x_2 \simeq s - d$ and so we know that $s \simeq (x_1 + x_2)/2$ (with exact equality, in the limit) and the prior can be (completely, in the limit) disregarded. More formally,

$$\lim_{\rho \to -1} \delta_{\text{MAP}}(\mathbf{x}) = \frac{x_1 + x_2}{2}.$$

End of Example 1.5.3

1.5.3 *The "quadratic error" loss function*

This loss function, suited to (scalar) estimation problems ($\mathcal{A} = \mathcal{S} = I\!R$), is defined as $L(s, d) = k\,(s - a)^2$, where k is an arbitrary constant (see Figure 1.9). Inserting this loss function into the Bayes' decision rule (Eq. (1.11)) leads to

$$\begin{aligned} \delta_{\text{PM}}(\mathbf{x}) &= \arg\min_{a \in \mathcal{A}} E_S\left[(s-a)^2 | x\right] \\ &= \arg\min_{a \in \mathcal{A}} \left(E_S[s^2|\mathbf{x}] + a^2 - 2aE_S\left[s|\mathbf{x}\right]\right) \qquad (1.97) \\ &= E_S\left[s|\mathbf{x}\right], \qquad (1.98) \end{aligned}$$

which is the *posterior mean* (PM) or *posterior expected value*. To go from Eq. (1.97) to Eq. (1.98), notice that given some $\mathbf{x}$, $E_S\left[s^2|\mathbf{x}\right]$ is a constant; we can then solve for a by taking the derivative and setting it to zero. If the *a posteriori* probability density function is Gaussian, the PM and MAP estimates coincide because the mean and the mode of a normal density are at the same location.

This result has a very interesting generalization which makes the PM estimate particularly notable (see [104], for a proof). If the cost function is symmetric in the sense that

$$L(s,-d) = L(s,d)$$

and strictly convex (see Appendix A), i.e.,

$$L(s, \lambda d_1 + (1-\lambda)d_2) < \lambda L(s,d_1) + (1-\lambda)L(s,d_2)$$

and $p_S(s|\mathbf{x})$ is symmetric around its mean $\delta_{\mathrm{PM}}(\mathbf{x})$, then the optimal Bayes estimate is still $\delta_{\mathrm{PM}}(\mathbf{x})$. This shows that for this class of symmetric posteriors, not only the quadratic error loss function, but any symmetric strictly convex loss function (e.g., $(s-a)^n$, with n even) leads to $\delta_{\mathrm{PM}}(\mathbf{x})$ as the optimal Bayes' estimator.

Situations where $\mathcal{S} = \mathcal{A} \subset I\!\!R$ is a discrete numeric set, rigorously, should be addressed as classification problems. However, when the cardinality of set $\mathcal{S}$ is *large* (and so its discrete nature is not very relevant) and s posseses a clear quantitative meaning, this may still be called an estimation problem and the quadratic error loss function may be adopted. For examples, think of $\mathcal{S}$ as resulting from the discretization of some intrinsically continuous quantity, such as gray levels in a digital image which are usually quantized into, say, $\mathcal{S} = \mathcal{A} = \{0, 1, ..., 255\}$. The optimal decision rule (restarting from Eq. (1.97) because here it is not possible to obtain a derivative given that $\mathcal{A}$ is a discrete set) is, in this case,

$$\begin{aligned}\delta_{\mathrm{TPM}}(\mathbf{x}) &= \arg\min_{a\in\mathcal{A}} \left(a^2 - 2aE\left[s|\mathbf{x}\right]\right) &(1.99)\\ &= \arg\min_{a\in\mathcal{A}} \left\{\left(a - \delta_{\mathrm{PM}}(\mathbf{x})\right)^2\right\}, &(1.100)\end{aligned}$$

which is called *thresholded posterior mean* (TPM) [74]; notice that a discrete $\mathcal{A}$ is not guaranteed to contain $\delta_{\mathrm{PM}}(\mathbf{x}) = E\left[s|\mathbf{x}\right]$ and so $\delta_{\mathrm{TPM}}(\mathbf{x})$ returns the closest value from $\mathcal{A}$.

Finally, it is worth mentioning the following fact: consider a *weighted* quadratic loss function $L(s,a) = W(s)\ (s-a)^2$; the optimal Bayes estimator, called the *weighted posterior mean* (WPM), is

$$\begin{aligned}\delta_{\mathrm{WPM}}(\mathbf{x}) &= \arg\min_{a\in\mathcal{A}} \left(E\left[W(s)\,|\mathbf{x}\right]a^2 - 2aE\left[W(s)\ s|\mathbf{x}\right]\right) &(1.101)\\ &= \frac{E\left[W(s)\,s|\mathbf{x}\right]}{E\left[W(s)\,|\mathbf{x}\right]}. &(1.102)\end{aligned}$$

This fact expresses a *duality*-type property that the quadratic loss function enjoys with respect to the prior; in fact, notice that Eq. (1.102) coincides with the posterior mean that would be obtained with a modified prior $p'_S(s) \propto W(s)\, p_S(s)$. This type of duality between prior and loss function is typical of Bayesian inference, and a previous instance of it was very clear in Example 1.4.3 (see Eq. (1.45)).

Note that, since the *a posteriori* probability density functions in Examples 1.5.1, 1.5.2 and 1.5.3 are all Gaussian, the PM and MAP estimates coincide; thus, the same examples also illustrate PM estimation under normal likelihoods and priors.

Example 1.5.4

This example shows a situation where the MAP and PM produce different estimates. Let us consider again, an unknown real quantity s whose prior information is expressed by $p_S(s) = \mathcal{N}(s|s_0, \phi^2)$. Let us consider that this unknown is observed through the following mechanism: with probability β, a constant ϵ is added to it, while with probability $1-\beta$, it is left unchanged. In both cases, the resulting value is additively contaminated by a Gaussian perturbation of zero mean and variance σ^2; this mechanism can be represented by a so-called *mixture* (of Gaussians) *model*, in this case with two components,

$$f_X(x|s) = \frac{\beta}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-s-\epsilon)^2}{2\sigma^2}\right\} + \frac{1-\beta}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-s)^2}{2\sigma^2}\right\},$$

with parameters ϵ, β, s_0, and σ^2 considered known. It is a simple task to verify that the *a posteriori density* is

$$\begin{aligned} p_S(s|x) \quad \propto \quad & \beta \exp\left\{-\frac{(x-s-\epsilon)^2}{2\sigma^2} - \frac{(s-s_0)^2}{2\phi^2}\right\} \\ & + (1-\beta)\exp\left\{-\frac{(x-s)^2}{2\sigma^2} - \frac{(s-s_0)^2}{2\phi^2}\right\}. \quad (1.103) \end{aligned}$$

To see how such a function looks like, let us consider some specific values: $s_0 = 0$, $\beta = 0.6$, $\phi^2 = 4$, $\sigma^2 = 1/2$, and an observation $x = 0.5$. With these values, $p_S(s|x = 0.5)$ is as shown in Figure 1.12.

The mean is $\delta_{\text{PM}}(0.5) = 1.2$, while the mode is at $\delta_{\text{MAP}}(\mathbf{x}) \simeq 0.4$ (see the arrows in Figure 1.12). This is an instance of the typical behavior of the MAP and PM criteria in presence of multi-modal *a posteriori* probability densities: in general, the PM criterion behaves conservatively and outputs a compromise solution between the (two, in this case) modes; in contrast, the MAP criterion chooses the largest mode and outputs the location of its peak, ignoring all the other modes.

End of Example 1.5.4

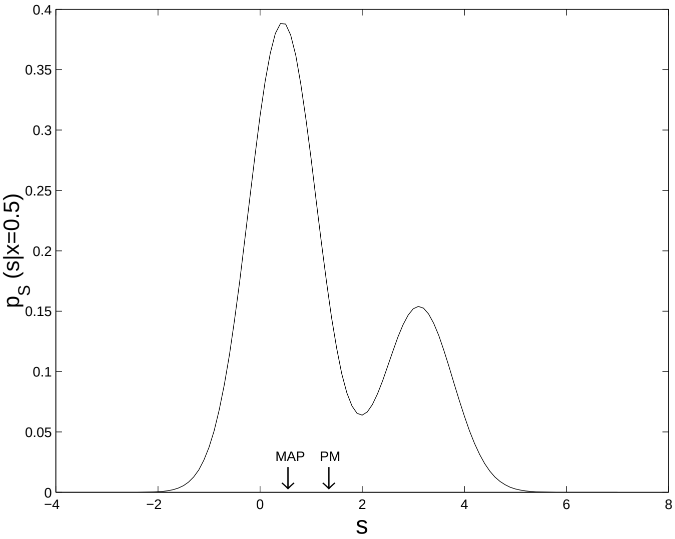

FIGURE 1.12. MAP and PM estimates for Example 1.5.4.

1.5.4 The "absolute error" loss function

The "absolute error" loss function (also depicted in Figure 1.9), defined as $L(s,a) \propto |s-a|$, is also suited to scalar estimation problems, i.e., when $\mathcal{A} = \mathcal{S} = I\!R$. Again, from the definition of the Bayes' decision rule (Eq. (1.11)),

$$\begin{aligned}\delta(\mathbf{x}) &= \arg\min_{a\in\mathcal{A}} E_s\left[(|s-a|)|\mathbf{x}\right] \\ &= \arg\min_{a\in\mathcal{A}} \left(\int_{-\infty}^{a} (a-s)p(s|\mathbf{x})\,ds + \int_{a}^{+\infty} (s-a)p(s|\mathbf{x})\,ds \right).\end{aligned}$$

By invoking Leibniz's rule[3] to compute the derivative with respect to a, one is lead to

$$\delta_{\text{MPD}}(\mathbf{x}) = \underset{d\in\mathcal{A}}{\text{solution}} \left\{ \int_{-\infty}^{d} p(s|\mathbf{x})\,ds = \int_{d}^{+\infty} p(s|\mathbf{x})\,ds \right\}, \qquad (1.104)$$

which is, by definition, the *median of the posterior density* (MPD) $p(s|\mathbf{x})$, i.e., $\delta_{\text{MPD}}(\mathbf{x}) = \text{median}\left[p(s|\mathbf{x})\right]$. As is clear from Eq. (1.104), the median is the point that splits the total probability in two equal halves.

[3] Leibniz's rule states that:

$$\frac{d}{dx}\left(\int_{\alpha(x)}^{\beta(x)} f(x,t)dt\right) = \int_{\alpha(x)}^{\beta(x)} \frac{df(x,t)}{dx}dt + \frac{d\beta(x)}{dx}f(x,\beta(x)) - \frac{d\alpha(x)}{dx}f(x,\alpha(x)).$$

Finally, notice that for Gaussian *a posteriori* densities, the MPD estimate coincides with the PM and MAP estimates, because the mean, the mode and the median are the same. So, once more, Examples 1.5.1, 1.5.2 and 1.5.3 can also be thought of as utilizing the MPD criterion.

Example 1.5.5

We will compare the behavior of the MAP, PM, and MPD criteria, using the simplest instance of what are known as change-point location problems. Consider a sequence of independent observations $\mathbf{x} = (x_1, ..., x_s, x_{s+1}, ...x_n)$, where $x_1, ..., x_s$ are samples of Gaussian density with mean μ_A and variance σ_A^2, while $x_{s+1}, ..., x_n$ have mean μ_B and variance σ_B^2. Assume that μ_A, σ_A^2, μ_B, and σ_B^2 are known, and the goal is to estimate the "change-point" s; clearly, $\mathcal{S} = \{1, 2, ..., n-1\}$. From the independence assumption, the likelihood function is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|s) &= (2\pi)^{-\frac{n}{2}}(\sigma_A^2)^{-\frac{s}{2}}(\sigma_B^2)^{-\frac{n-s}{2}} \\ & \quad \exp\left\{-\sum_{i=1}^{s}\frac{(x_i-\mu_A)^2}{2\sigma_A^2} - \sum_{i=s+1}^{n}\frac{(x_i-\mu_B)^2}{2\sigma_B^2}\right\} \end{aligned} \quad (1.105)$$

Assuming a uniform prior $p(s) = 1/n$, expressing no preference for any location, the MAP estimate becomes simply

$$\widehat{s}_{\text{MAP}} = \arg\max_s p_S(s|\mathbf{x}) = \arg\max_s f_{\mathbf{X}}(\mathbf{x}|s).$$

The PM estimate (actually, it is a TPM estimate), is given by

$$\widehat{s}_{\text{TPM}} = \text{round}\left[\left(\sum_{s=1}^{n} s\, f_{\mathbf{X}}(\mathbf{x}|s)\right)\left(\sum_{s=1}^{n} f_{\mathbf{X}}(\mathbf{x}|s)\right)^{-1}\right],$$

where the function round[$\cdot$] returns the closest integer to its real argument. Finally, the MPD estimate is

$$\begin{aligned} \widehat{s}_{\text{MPD}} &= \text{median}\,[p_S(s|\mathbf{x})] \\ &= \min\left\{s : \sum_{i=1}^{s} p_S(i|\mathbf{x}) > \frac{1}{2}\right\}. \end{aligned} \quad (1.106)$$

Notice that the median of a probability mass function is not well defined; in general, there is no point in S spliting the total probablity mass into two exact halves. As with the TPM, a thresholding scheme is required; Eq. (1.106) is one such scheme.

To study the behavior of these three criteria, consider a situation where $\mu_A = 0$, $\mu_B = 2$, $\sigma_A^2 = \sigma_B^2 = 1$, $n = 80$, and $s = 40$; an example of a sequence of observations thus generated is shown in Figure 1.13. In Figure 1.14, histograms of the results produced by the three estimation criteria

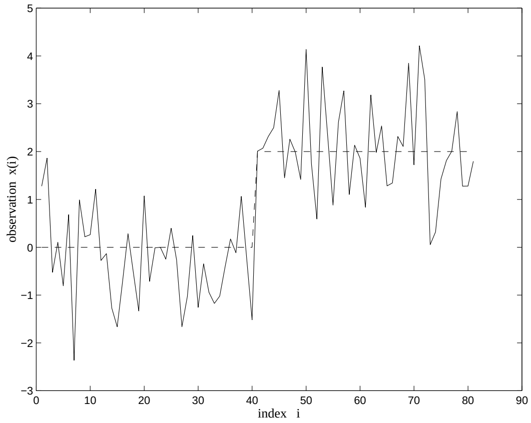

FIGURE 1.13. Examples of a sequences of independent Gaussian observations all with unit variance; the mean (shown by the dashed line) up to $i = 40$ equals zero, and from $i = 41$ to $i = n = 80$ equals 2.

are shown, for a set of 5000 trials. As can be seen from Figure 1.13, it is fairly obvious where the change occurs, and the three criteria perform comparably well; the only noticeable difference is that the TPM criterion produces fewer correct estimates.

Different behaviors start showing when we choose a more difficult setting with $\mu_B = 0.5$; the sequence of observations in Figure 1.15 show that now it is not very clear where the change is located. This added difficulty is reflected in the histograms presented in Figure 1.16. Now it is possible to perceive qualitative differences between the criteria: the MAP produces many more correct estimates than the other two; this may be traced back to the "0/1" loss function for which the "price paid" for (any) wrong estimate is high when compared to the zero cost of a correct decision.

The TPM (or PM) criterion is based on a (quadratic) loss function where the "cost" does depend on how far the estimate is from the true value. This makes it a more "conservative" criterion: though it produces fewer correct decisions than the MAP, the wrong estimates are clustered around the true value. Finally, the MPD criterion can be placed somewhere between these two behaviors; its loss function does not penalize wrong estimates as strongly as the quadratic loss function, but it is still sensitive to the estimation error.

End of Example 1.5.5

1.6 Summary

Statistical decision theory consists of a set of formal tools to deal with decision making problems under uncertainty: the *state of nature s* which is

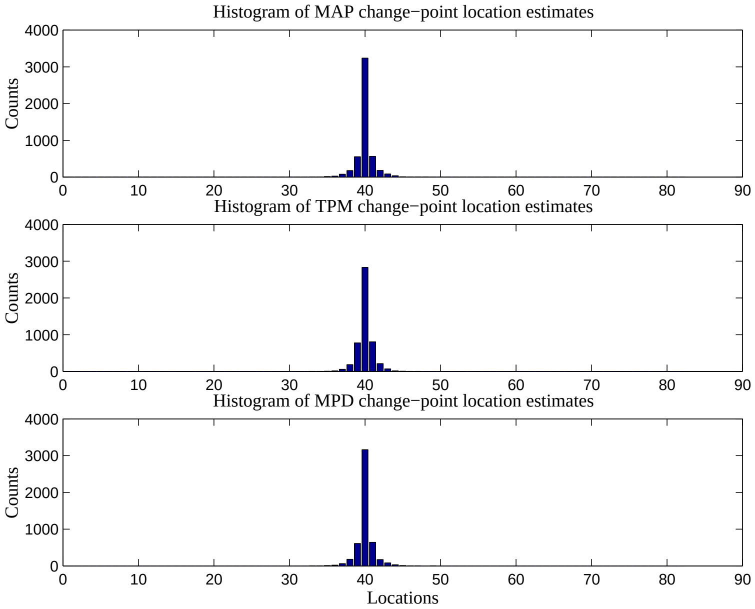

FIGURE 1.14. Histograms of the MAP, TPM, and MPD estimates of the change point location; see text for details.

an unknown element of a configuration space $\mathcal{S}$; a probabilistic model, in the form of a *conditional probability function* $f_{\mathbf{X}}(\mathbf{x}|s)$ (the *likelihood function*), of the mechanism by which observations $\mathbf{x} \in \mathcal{X}$ are obtained given the (unknown) state of nature s; a set $\mathcal{A}$ of possible *actions* or *decisions*; a quantification of the consequences of choosing action $a \in \mathcal{A}$ when the true state of nature is $s \in \mathcal{S}$, by means of a *loss function* $L(s, a)$. The goal of decision theory is to propose and evaluate decision rules $\delta(\mathbf{x})$, which are functions that map the observation space into the action set $\mathcal{A}$.

In the *frequentist* perspective, decision rules are evaluated by measuring how well (in terms of average loss) they perform when repeatedly used under the same state of nature for all possible observations. We saw that this approach does not allow deriving closed form expressions for optimal decision rules.

The Bayesian approach brings a new element into play: *a priori* knowledge under the form of a probability function $p_S(s)$ defined on $\mathcal{S}$. This *a priori* probability, together with the likelihood function, are used by Bayes' law to yield the *a posteriori* probability function $p_S(s|\mathbf{x})$. This course of action adheres to one of the fundamental principles of the Bayesian philosophy, *conditionality*, which advocates that any decision should be based (conditioned) on what has actually been observed. Candidate decision rules, according to this principle, must be evaluated by how well they perform on average, according to $p_S(s|\mathbf{x})$, for the given observed data $\mathbf{x}$. In conclusion,

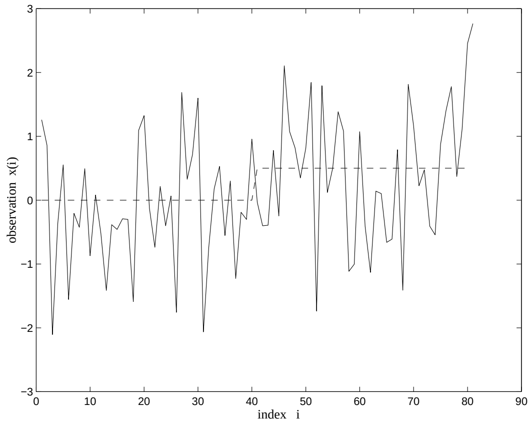

FIGURE 1.15. Examples of a sequences of independent Gaussian observations all with unit variance; the mean (shown by the dashed line) up to $i = 40$ equals zero, and from $i = 41$ to $i = n = 80$ equals 0.5.

the main points to be learned from the material presented in this chapter can be summarized as follows:

(a) model the available knowledge and uncertainty with probabilities;

(b) use the basic laws of probability theory (namely Bayes' law);

(c) condition on what is known;

(c) average over what is unknown.

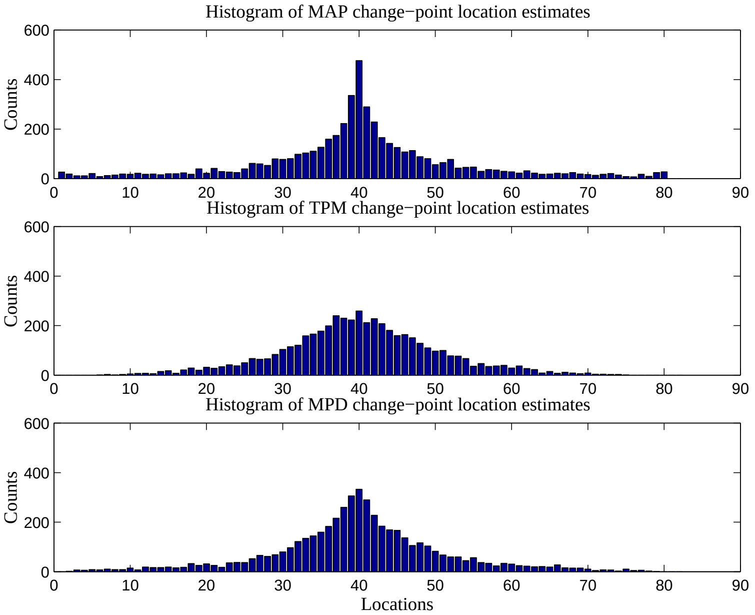

FIGURE 1.16. Histograms of the MAP, TPM, and MPD estimates of the change point location; see text for details.

2
Topics in Bayesian Inference

In the previous Chapter we introduced the basic concepts of Bayesian classification and estimation theory. To keep the presentation as clear and uncluttered as possible, we focused on essential aspects, postponing other (arguably not so essential) topics to the current Chapter. How to specify priors is the subject that we will address first; then we will look at sufficient statistics and exponential families of probability distributions, two highly interrelated concepts which are equally important for non-Bayesian approaches. The Chapter will end with a section on foundational/philosophical aspects of Bayesian inference.

2.1 Specifying Priors

As we have mentioned in Section 1.3, the Bayesian paradigm generally implies a *personal* (or *subjective*) view of probabilities; from this standpoint, the priors formally (quantitatively) express degrees of personal belief. One of the main criticisms of Bayesian inference is aimed precisely at the arbitrariness which resides in the personal (or subjective) choice of priors. As an alternative to this subjective *prior elicitation*, formal (objective) rules to build priors have been put forward by several authors. The research devoted to formal procedures for specifying priors may be divided into two (not independent) areas:

- In the first area, efforts are made for finding priors which are said to be *non-informative*, reflecting a desire to remove subjectiveness from Bayesian procedures; a *non-informative* prior may be seen as a *reference* against which any subjective prior can be compared [14]. Non-informative priors obtained from invariance arguments (including the so-called *Jeffreys' priors*) are arguably the most simple (but profound) results in this area.

- In the second area, the goal is to obtain (informative) priors which are compatible with partial *a priori* knowledge. Here, we find *conjugate priors*, which are obtained by invoking computational/analytical tractability arguments, and *maximum entropy priors*. The latter result from applying information theoretical arguments in building priors that, while compatible with the partial prior information available, are as non-informative as possible. Another information theoretical approach which, when seen under a Bayesian light, can be interpreted as yielding objective priors is Rissanen's *minimum description length* (MDL) principle which we will also consider in this chapter.

We must point out that the issue of prior selection (including the subjective/objective discussion) is one of the most controversial aspects of Bayesian theory. An interesting and readable overview of the conceptual and philosophical aspects underlying these issues can be found in [54]. A recent survey on methods for formal selection of priors, including a very comprehensive list of references, is [61].

2.2 Improper Priors and Maximum Likelihood Estimation

For Bayesian inference, the prior knowledge about the quantity of interest is formalized by considering it a random variable S characterized by its *a priori* probability (density or mass) function $p_S(s)$. However, it is not vital to believe in a strict probabilistic interpretation of the prior. In many circumstances, the prior is just a (technical) way of expressing available information leading to an inference procedure; in particular, the Bayesian paradigm (formally) allows considering priors which are not normalized (nor normalizable), which are then called *improper priors*. These priors are characterized by

$$\int_{\mathcal{S}} p_S(s)\, ds = +\infty, \tag{2.1}$$

or, in discrete contexts where $\mathcal{S}$ is infinite, by

$$\sum_{s_i \in \mathcal{S}} p_S(s_i) = +\infty, \tag{2.2}$$

thus failing to obey one of the basic laws of probability. By inserting an improper prior into Bayes' theorem, as given by Eq. (1.9), a conventional (proper) posterior may still be obtained as long as the marginal in the denominator is well defined. Although there is some controversy surrounding improper priors (see [54]), mainly because they can not be interpreted as conventional probability functions, their acceptance has important desirable consequences [8], [14], [54], [93]. In practice, most improper priors of interest can be interpreted as limits of proper ones (e.g., Gaussians of infinite variance, uniform densities on infinite intervals) thus bringing a "closure" property to Bayesian inference.

Example 2.2.1

Consider an estimation problem ($\mathcal{S} = \mathcal{A} = I\!\!R$), with the likelihood function $f_X(x|s) = \mathcal{N}(x|s,1)$ and the prior $p_S(s) = c \neq 0$, which is clearly improper; nevertheless, since

$$\begin{aligned} p_S(s|x) &= \frac{f_X(x|s)\; p_S(s)}{\displaystyle\int_{I\!\!R} f_X(x|s)\; p_S(s) ds} \\ &= \frac{\frac{c}{\sqrt{2\pi}} \exp\left\{-\frac{(x-s)^2}{2}\right\}}{\displaystyle\int_{I\!\!R} \frac{c}{\sqrt{2\pi}} \exp\left\{-\frac{(x-s)^2}{2}\right\} ds} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-s)^2}{2}\right\} \;\equiv\; f_X(x|s), \end{aligned} \tag{2.3}$$

the posterior exists and coincides formally with the likelihood function. Now assume that $p_S(s) = \mathcal{N}(s|s_0, \phi^2)$ as in Example 1.5.1. From Eq. (1.85), it is clear that $\delta(\mathbf{x}) \overset{\phi\rightarrow\infty}{\longrightarrow} x$, i.e. the improper prior leads to an estimator coinciding with a limit situation of a family of proper priors. This is how improper priors can be seen as providing "closure" to the Bayesian inference setting.

End of Example 2.2.1

Let us generalize the previous example to any estimation problem ($\mathcal{S} = \mathcal{A}$); if $f_{\mathbf{X}}(\mathbf{x}|s)$ denotes the likelihood function and the prior is uniform (maybe improper) $p_S(s) = c \neq 0$, then the posterior $p_S(s|\mathbf{x})$ is proportional to the likelihood, as long as the marginal $f_{\mathbf{X}}(\mathbf{x})$ is well defined (finite); in fact,

$$\begin{aligned} p_S(s|\mathbf{x}) &= \frac{f_X(\mathbf{x}|s)\; p_S(s)}{\displaystyle\int_{I\!\!R} f_X(\mathbf{x}|s)\; p_S(s) ds} \\ &= \frac{c\; f_{\mathbf{X}}(\mathbf{x}|s)}{c\displaystyle\int_{I\!\!R} f_{\mathbf{X}}(\mathbf{x}|s)\; ds} \;\propto\; f_{\mathbf{X}}(\mathbf{x}|s), \end{aligned} \tag{2.4}$$

as long as the denominator is finite. In particular, with such a uniform prior, the MAP criterion yields

$$\begin{aligned}\delta_{\text{MAP}}(\mathbf{x}) &= \arg\max_{s\in\mathcal{S}} p_S(s|\mathbf{x}) \\ &= \arg\max_{s\in\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}|s) \equiv \delta_{\text{ML}}(\mathbf{x}), \qquad (2.5)\end{aligned}$$

which is called the *maximum likelihood* (ML) estimate. Its designation stems from the decision criterion it stipulates: choose s so that it maximizes the probability (or probability density) of the actually observed data, i.e., the density that is most *likely* to have generated the observed data. The use of improper priors allows recovering the ML criterion (which is a widely used criterion independent of any Bayesian considerations) as a limit situation (i.e., on the "boundary") of Bayesian inference.

In classification problems with a finite number M of hypotheses, the uniform prior is not improper, it is simply $p(s) = 1/M$, for $s \in \mathcal{S} = \{s_1 \dots, s_M\}$. Inserting it into Eq. (1.25) reduces the MAP criterion to

$$\delta_{\text{MAP}}(\mathbf{x}) = \arg\max_{s\in\mathcal{S}} p(\mathbf{x}|s) \equiv \delta_{\text{ML}}(\mathbf{x}) \qquad (2.6)$$

which is (naturally) known as the *maximum likelihood* (ML) classifier.

Example 2.2.2

Considering Example 1.4.1 again, if the two binary digits are *a priori* equiprobable, $p_0 = p_1 = 1/2$, the threshold becomes 0 (see Eq. (1.32)), and the decision rule simply checks which of the two values (s_0 or s_1) is closer to the observation x. In the communications literature, this is known as the *ML detector*.

End of Example 2.2.2

With the *a priori* equiprobability assumption, the optimal Bayes' classifiers for the Gaussian observation models in Eqs. (1.34), (1.35), and (1.37), become simpler since the term containing $p_S(s_i)$ can be dropped. In this case they become ML classifiers. Particularly interesting are Eqs. (1.35) and (1.37) which now simply return the class whose mean vector $\boldsymbol{\mu}_i$ is closest (in Euclidean or Mahalanobis distance, respectively) to the observed vector $\mathbf{x}$. Two other equations which become particularly meaningful under the equiprobability assumption are Eqs. (1.52) and (1.57); they represent simply the mid-point between the two competing hypotheses.

Example 2.2.3

Going back to Examples 1.5.1, 1.5.2, and 1.5.3, if we let the prior variance go to infinity, $\phi^2 \to \infty$, we recover the corresponding maximum likelihood estimators.

End of Example 2.2.3

2.3 Conjugate priors

There are many situations in which the prior knowledge about the state of nature is not concrete enough to specify an *a priori* probability function; however, to follow a Bayesian approach, one is still needed. In this cases there is some freedom which can be exploited to allow mathematical tractability considerations to come into play: given the likelihood function, one can look for a prior which, on the one hand, is compatible with the available knowledge and, on the other hand, leads to an *a posteriori* probability function satisfying certain (computational convenience) conditions. Notice that when combining priors and likelihoods via Bayes' rule, one may often arrive at intractable *a posteriori* probability functions for which closed form expressions may not even exist. These concerns have motivated the study of the so-called *conjugate families*, or *conjugate priors*. The formalization of this concept is as follows:

Let $\mathcal{F} = \{f_{\mathbf{X}}(\mathbf{x}|s),\ s \in \mathcal{S}\}$ *be a class of likelihood functions; let* $\mathcal{P}$ *be a class (set) of probability (density or mass) functions; if, for any* $\mathbf{x}$*, any* $p_S(s) \in \mathcal{P}$*, and any* $f_{\mathbf{X}}(\mathbf{x}|s) \in \mathcal{F}$*, the resulting a posteriori probability function* $p_S(s|\mathbf{x}) \propto f_{\mathbf{X}}(\mathbf{x}|s)\, p_S(s)$ *is still in* $\mathcal{P}$*, then* $\mathcal{P}$ *is called a conjugate family, or a family of conjugate priors, for* $\mathcal{F}$*.*

Of course, a trivial conjugate family for any class of likelihoods is the set of all probability functions on $\mathcal{S}$. Interesting conjugate families should be as small as possible and, more importantly, parameterized; when this is the case, computing the posterior from the prior is simply a matter of updating the associated parameters. Some examples will help to elucidate these ideas.

Example 2.3.1

When estimating s from a set of n i.i.d. normal observations of mean s and known variance σ^2, the likelihood function is

$$f_{\mathbf{X}}(\mathbf{x}|s) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - s)^2}{2\sigma^2}\right\}; \tag{2.7}$$

if the prior is normal, $p(s) = \mathcal{N}(s|s_0, \phi^2)$, then the posterior is also normal; its mean is given by Eq. (1.89) and its variance is $\psi^2 = \sigma^2\phi^2/(n\phi^2 + \sigma^2)$. That is, for Gaussian likelihoods (with respect to the mean) the family of normal priors is a conjugate one since the posterior is also normal. This fact explains the adoption of Gaussian priors in Examples 1.5.1, 1.5.2, and 1.5.3. As a final note, observe that the *a posteriori* mean (and consequently the MAP, PM, and MPD estimates) converges, as $n \to \infty$, to $\bar{x} = (x_1 + x_2 + ... + x_n)/n$ (the sample mean) which is the ML estimate.

End of Example 2.3.1

With the conjugacy concept in hand, we can venture into considering other estimation examples, beyond the simple Gaussian prior and Gaussian likelihood cases studied above.

Example 2.3.2

Let θ and $1-\theta$ denote the (unknown) probabilities of heads and tails, respectively, of a given coin under study. The outcomes of an observed sequence of n tosses is denoted by $\mathbf{x} = (x_1, \ldots, x_n)$, with $x_i = 1$ standing for a head, and $x_i = 0$ for a tail. The likelihood function is then a Bernoulli distribution; i.e., letting $n_h(\mathbf{x}) = x_1 + x_2 + ... + x_n$ denote the number of heads outcomes, it can be written as

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})}. \tag{2.8}$$

Ignorance about θ can be modeled by a uniform prior $p_\Theta(\theta) = 1$, for $\theta \in [0,1]$ (notice this is not improper, due to the boundedness of the parameter space); with this flat prior, the *a posteriori* density is

$$p_\Theta(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)}{\displaystyle\int_0^1 f(\mathbf{x}|\theta)\,d\theta} = \frac{\Gamma(2+n)\,\theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})}}{\Gamma(1+n_h(\mathbf{x}))\,\Gamma(1+n-n_h(\mathbf{x}))} \tag{2.9}$$

which, apart from the normalizing constant (where Γ is the Euler gamma function[1]), has the same form as the likelihood. The MAP and ML estimates are both simply $\delta_{\mathrm{ML}}(\mathbf{x}) = \delta_{\mathrm{MAP}}(\mathbf{x}) = n_h(\mathbf{x})/n$, while $\delta_{\mathrm{PM}}(\mathbf{x}) = (n_h(\mathbf{x})+1)/(n+2)$. Now consider that in a particular experiment, in a total of, say, 4 tosses, all outcomes are heads, we obtain $\delta_{\mathrm{ML}}(\mathbf{x}) = \delta_{\mathrm{MAP}}(\mathbf{x}) = 1$, and $\delta_{\mathrm{PM}}(\mathbf{x}) = 5/6$. These values are incompatible with the common belief about heads and tails probabilities, *a priori* expected to be around 1/2. To formalize this belief, a prior with the maximum at 1/2, symmetric around 1/2, and going to zero as s goes to 0 or 1, has to be used. Of course, many functions satisfy these conditions, but most of them lead to intractable posteriors; a convenient conjugate prior turns out to be the Beta density

$$\begin{aligned} p_\Theta(\theta|\alpha,\beta) &= \mathrm{Be}(\theta|\alpha,\beta) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}, \end{aligned} \tag{2.10}$$

[1] The Euler gamma function is defined as

$$\Gamma(z) = \int_0^\infty t^{z-1}\,e^{-t}\,dt,$$

valid for any complex number z. For positive integer arguments, $\Gamma(n) = (n-1)!$, and it may thus be seen as generalizing the factorial function to non-integer arguments.

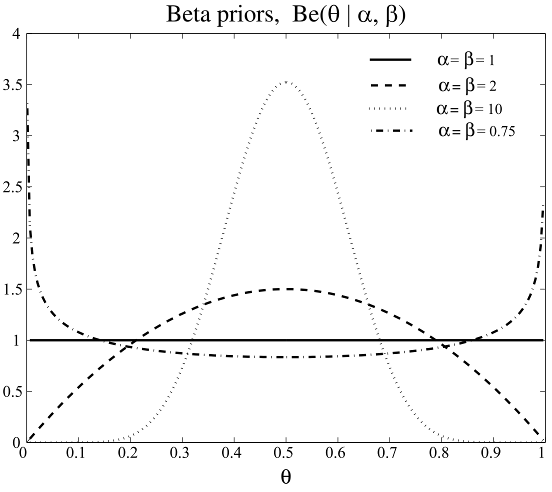

FIGURE 2.1. Beta densities for different parameter choices: $\alpha = \beta = 0.75$, 1, 2, and 10. Notice how for $\alpha = \beta \leq 1$, the behavior is qualitatively different, with the mode at 1/2 disappearing.

defined for $\theta \in [0, 1]$ and $\alpha, \beta > 0$; the main features of this density are

$$\begin{aligned} E[\theta|\alpha,\beta] &= \frac{\alpha}{\alpha+\beta} \quad \text{(mean)} && (2.11)\\ E\left[\left(\theta - \frac{\alpha}{\alpha+\beta}\right)^2 \middle| \alpha,\beta\right] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad \text{(variance)} \\ \arg\max_\theta p_\Theta(\theta|\alpha,\beta) &= \frac{\alpha-1}{\alpha+\beta-2} \quad \text{(mode, if } \alpha > 1). && (2.12) \end{aligned}$$

Accordingly, our desire of "pulling" the estimate towards 1/2 can be expressed by choosing $\alpha = \beta$, while their common magnitude allows controlling how "strongly we pull". Several Beta densities are depicted in Figure 2.1. Notice the uniform density is recovered for $\alpha = \beta = 1$, showing that it is a special case of Beta.

The corresponding *a posteriori* is still a Beta density, easily identifiable after multiplying the prior in Eq. (2.10) by the likelihood in Eq. (2.8)

$$p_\Theta(\theta|\mathbf{x},\alpha,\beta) = \mathrm{Be}\left(\theta|\alpha + n_h(\mathbf{x}), \beta + n - n_h(\mathbf{x})\right) \quad (2.13)$$

which leads to the following Bayesian estimates

$$\begin{aligned} \widehat{\theta}_{\mathrm{PM}} &= \delta_{\mathrm{PM}}(\mathbf{x}) = \frac{\alpha + n_h(\mathbf{x})}{\alpha+\beta+n} && (2.14)\\ \widehat{\theta}_{\mathrm{MAP}} &= \delta_{\mathrm{MAP}}(\mathbf{x}) = \frac{\alpha + n_h(\mathbf{x}) - 1}{\alpha+\beta+n-2}. && (2.15) \end{aligned}$$

For example, with $\alpha = \beta = 2$, and $n = n_h(\mathbf{x}) = 4$ (as above), we have $\delta_{\text{PM}}(\mathbf{x}) = 0.75$ and $\delta_{\text{MAP}}(\mathbf{x}) =\simeq 0.833$. which are more moderate estimates. With $\alpha = \beta = 10$, a stronger prior, we would obtain $\delta_{\text{PM}}(\mathbf{x}) = 0.58$ and $\delta_{\text{MAP}}(\mathbf{x}) \simeq 0.59$. In Figure 2.2, we plot a typical evolution of the Beta *a posteriori* density for two priors: uniform and $\text{Be}(\theta|5, 5)$. The observed data was generated using $\theta = 0.7$. Notice that as the amount of data increases, the influence of the prior decreases and both densities approach each other. In the limit of an infinitely long sequence of tosses, both estimates converge to $\lim_{n\to\infty} n_h(\mathbf{x})/n$; this quantity converges itself to the true θ, according to the weak law of large numbers. Parameter α (with $\alpha = \beta$) may be seen as a measure of how large $n_h(\mathbf{x})$ and n have to be before the observed data dominates the Bayesian estimates.

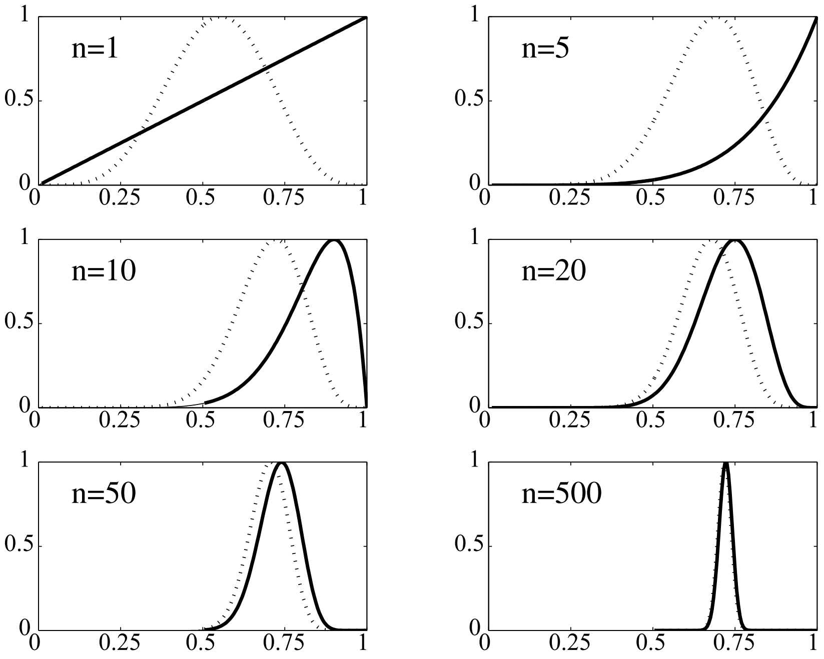

FIGURE 2.2. Typical evolution of the *a posteriori* Beta densities (non-normalized, merely scaled to 1) as the number n of Bernoulli trials increases; the solid line corresponds to a uniform prior (its maximum being thus the ML estimate), while the dotted line corresponds to a $\text{Be}(\theta|5, 5)$ prior. Notice how both densities increasingly concentrate their mass around the true value $\theta = 0.7$. The horizontal axes of these plots denote θ values.

End of Example 2.3.2

Example 2.3.3

Consider now, as another example, n i.i.d. zero-mean Gaussian observations of unknown variance σ^2; it happens that it is more convenient to reparameterize the problem into $\theta = \frac{1}{\sigma^2}$ (we will see another argument supporting this choice in Section 2.12). Accordingly, the following likelihood function is obtained

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \left(\frac{\theta}{2\pi}\right)^{\frac{n}{2}} exp\left\{-\frac{\theta}{2}\sum_{i=1}^{n} x_i^2\right\}. \tag{2.16}$$

It is easy to verify that a Gaussian prior on θ (the same would happen for σ^2, or even σ) does not lead to a Gaussian *a posteriori* density. For this particular parameterization, the Gamma density is a conjugate prior

$$\begin{aligned} p_\Theta(\theta|\alpha,\beta) &= \text{Ga}(\theta|\alpha,\beta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\,\theta^{\alpha-1}\exp\{-\beta\theta\} \end{aligned} \tag{2.17}$$

defined for $\theta \in [0,\infty)$ (as required by the meaning of $\theta = 1/\sigma^2$) and for $\alpha, \beta > 0$. Had we adopted the original σ^2 parameterization, and the conjugate prior would be the *inverse*-Gamma distribution, given by

$$\text{Inv-Ga}(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,(\sigma^2)^{-(\alpha+1)}\exp\left\{-\frac{\beta}{\theta}\right\}. \tag{2.18}$$

The main features of the Gamma distribution are

$$E[\theta|\alpha,\beta] = \frac{\alpha}{\beta} \quad \text{(mean)} \tag{2.19}$$

$$E\left[\left(\theta - \frac{\alpha}{\beta}\right)^2\middle|\,\alpha,\beta\right] = \frac{\alpha}{\beta^2} \quad \text{(variance)} \tag{2.20}$$

$$\arg\max_\theta p_\Theta(\theta|\alpha,\beta) = \frac{\alpha-1}{\beta} \quad \text{(mode)}, \tag{2.21}$$

where the mode expression is only valid if $\alpha \geq 1$, otherwise it has no mode. Figure 2.3 shows some plots of Gamma priors.

By multiplying together Eqs. (2.16) and (2.17) and identifying the result with a Gamma density, it becomes clear that

$$p_\Theta(\theta|x_1, x_2, ..., x_n) = \text{Ga}\left(\theta|\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n} x_i^2\right). \tag{2.22}$$

The above mentioned mean and mode of a Gamma density imply that

$$\widehat{\theta}_{\text{PM}} = \left(\frac{2\alpha}{n} + 1\right)\left(\frac{2\beta}{n} + \frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{-1} \tag{2.23}$$

$$\widehat{\theta}_{\text{MAP}} = \left(\frac{2\alpha}{n} + 1 - \frac{2}{n}\right)\left(\frac{2\beta}{n} + \frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{-1}. \tag{2.24}$$

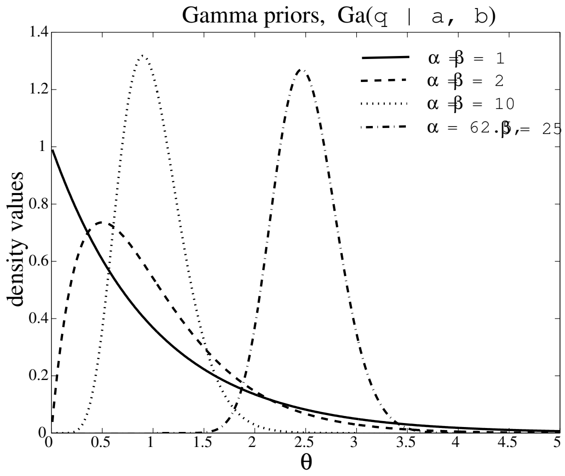

FIGURE 2.3. Gamma densities for different parameter choices. For $\alpha = \beta = 1$, 2, and 10, the mean equals 1, and the variances are 1, 0.5 , and 0.1, respectively. For $\alpha = 62.5$ and $\beta = 25$, the variance is 0.1, but the mean is 2.5.

Notice that, as expected, as n becomes larger both the PM and the MAP estimates approach the maximum likelihood one which is given by

$$\widehat{\theta}_{\text{ML}} = \frac{n}{\displaystyle\sum_{i=1}^{n} x_i^2};$$

i.e., the data term becomes dominant over the prior. This fact is illustrated in Figure 2.4 which is based on a sequence of 50 zero-mean unit-variance (i.e., $\theta = 1$) observations; two Gamma density priors were considered: $\alpha = \beta = 1$ and $\alpha = \beta = 10$. Both have mean equal to 1 (coinciding with the true θ) but the second one has a variance 10 times smaller than the first. Notice in the figure how the Bayesian estimates are more stable for small sample sizes (due to the presence of the prior) and how all three estimates approach each other and the true parameter value as the number of observations becomes large.

End of Example 2.3.3

Example 2.3.4

Let us now consider a Poisson observation model, with θ denoting the unknown rate (in counts/second, for example) of some phenomenon that follows Poisson statistics (e.g., radioactive emissions, as considered in Example 1.4.8). The observed data consist of a sequence of independent counts $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ registered during a set of observation inter-

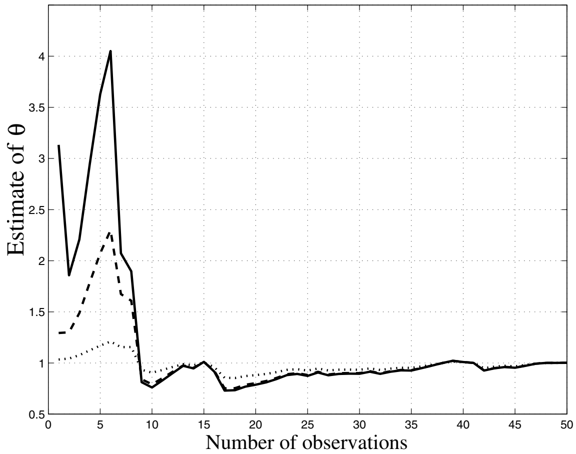

FIGURE 2.4. Evolution of the ML (solid line) and two MAP (for $\alpha = \beta = 1$: dashed line; and $\alpha = \beta = 10$: dotted line) estimates of θ versus the number of observed values.

vals $\{T_1, T_2, ..., T_n\}$ (measured in seconds). The likelihood function is then

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{j=1}^{n} \frac{e^{-\theta T_j}(\theta T_j)^{x_j}}{x_j!} \propto \theta^{t(\mathbf{x})} e^{-T\theta}, \tag{2.25}$$

where all factors independent of θ were dropped from the last expression, $t(\mathbf{x}) = x_1 + x_2 + ... + x_n$, and $T = T_1 + T_2 + ... + T_n$. To proceed with a Bayesian analysis of this problem, we should find a family of conjugate priors that allows formalizing any available prior knowledge. As for the likelihood of the previous example, Gamma densities are conjugate priors for Poisson likelihoods (see Eq. (2.17)). The posterior is then still a Gamma density, easily identifiable after multiplying Eq. (2.25) by Eq. (2.17),

$$p_\Theta(\theta|\mathbf{x}) = \text{Ga}(\theta|\alpha + t(\mathbf{x}), \beta + T) \tag{2.26}$$

and the corresponding Bayesian estimates are then

$$\widehat{\theta}_{\text{PM}} = \frac{\alpha + t(\mathbf{x})}{\beta + T} = \frac{\alpha + \sum_{j=1}^{n} x_j}{\beta + T} \tag{2.27}$$

$$\widehat{\theta}_{\text{MAP}} = \frac{\alpha + t(\mathbf{x}) - 1}{\beta + T} = \frac{\alpha - 1 + \sum_{j=1}^{n} x_j}{\beta + T}. \tag{2.28}$$

Again, these Bayesian criteria converge to the maximum likelihood; notice that since the intervals T_i have non-zero length, if $n \to \infty$, then $T \to \infty$ and

$t(\mathbf{x}) \to \infty$ (although some more care is required to establish the validity of the last limit, because $t(\mathbf{x})$ is a random quantity); as a consequence, both the MAP and the PM estimates approach the one provided by the ML criterion,

$$\widehat{\theta}_{\text{ML}} = \frac{\sum_{i=j}^{n} x_j}{T} \tag{2.29}$$

which is simply the sample average rate.

End of Example 2.3.4

Example 2.3.5

Consider a set of (real non-negative) observations $\mathbf{x} = (x_1, x_2, ..., x_n)$ which are assumed independent and identically distributed according to the uniform density on the interval $[0, \theta]$, with θ the unknown to be estimated. Formally,

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \begin{cases} \theta^{-n}, & \max\{x_1, ...x_n\} \leq \theta \\ 0, & \text{otherwise.} \end{cases} . \tag{2.30}$$

Notice that the condition $\max\{x_1, ...x_n\} \leq \theta$ is equivalent to the conjunction of $\{x_1 \leq \theta, x_2 \leq \theta, ..., x_n \leq \theta\}$.

Let us take as prior the so-called Pareto distribution whose density is given by

$$p_\Theta(\theta) = \text{Pa}(\theta|\alpha, \beta) = \begin{cases} \alpha\beta^\alpha\theta^{-(\alpha+1)}, & \theta \geq \beta \\ 0, & \theta < \beta. \end{cases} \tag{2.31}$$

Several examples of this density, for different values of the parameters, are shown in Figure 2.5. Its mode is obviously located at $\theta = \beta$, while its mean is given by

$$E[\theta|\alpha, \beta] = \frac{\alpha\beta}{\alpha - 1},$$

if $\alpha > 1$, otherwise the mean does not exist.

Multiplication of the likelihood in Eq. (2.30) by the prior in Eq. (2.31) leads to the following *a posteriori* probability density function:

$$p_\Theta(\theta|\mathbf{x}) = \text{Pa}\left(\theta|n + \alpha, \max\{\beta, \max\{x_1, ..., x_n\}\}\right). \tag{2.32}$$

showing that the Pareto density is conjugate with respect to the uniform observation model. Finally, the MAP and PM estimates are given by

$$\widehat{\theta}_{\text{MAP}} = \max\{\beta, \max\{x_1, ..., x_n\}\} \tag{2.33}$$

$$\widehat{\theta}_{\text{PM}} = \frac{(\alpha + n)\max\{\beta, \max\{x_1, ..., x_n\}\}}{\alpha + n - 1} \tag{2.34}$$

Notice that, unlike in previous examples, these Bayesian rules do not converge to the maximum likelihood estimate as n approaches infinity.

End of Example 2.3.5

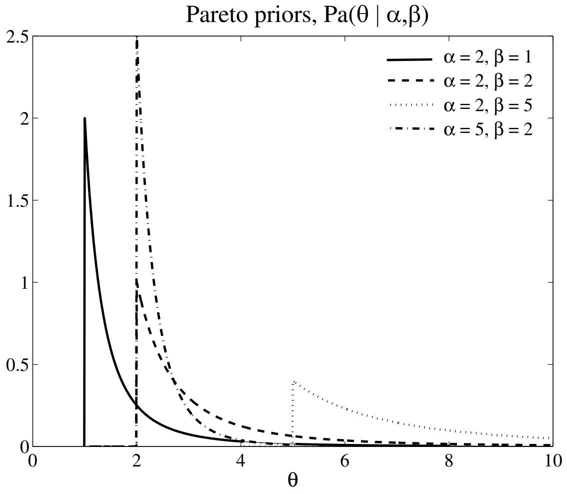

FIGURE 2.5. Pareto densities for different parameter choices. Notice how the density is zero for $\theta < \beta$, and the fact that a larger value of α leads to a narrower prior.

2.4 Mixtures of Conjugate Priors

It may be the case that the standard conjugate family for the likelihood function at hand is not expressive enough to model a certain kind of *a priori* belief. It is then possible to extend this family through the use of *mixtures* of conjugate priors [14], [26], [35], [93]. As we shall see, these mixtures are still conjugate with respect to the given likelihood function, and can be used to achieve more freedom and flexibility in formalizing prior knowledge.

Let $\mathcal{F}$ be some class of likelihood functions, and $\mathcal{P}$ a family of conjugate priors for $\mathcal{F}$. Consider a family $\mathcal{Q}^{(m)}$ of m-dimensional finite mixture models supported on $\mathcal{P}$, i.e.

$$\mathcal{Q}^{(m)} = \left\{ q_S(s) = \sum_{i=1}^{m} \lambda_i \, p_S^{(i)}(s) : \quad \sum_{i=1}^{m} \lambda_i = 1; \quad p_S^{(i)}(s) \in \mathcal{P} \right\}.$$

Notice that the *a posteriori* probability function $q_S(s|\mathbf{x})$ resulting from a prior $q_S(s)$ verifies

$$\begin{aligned} q_S(s|\mathbf{x}) \propto q_S(s) f_{\mathbf{X}}(\mathbf{x}|s) &= f_{\mathbf{X}}(\mathbf{x}|s) \sum_{i=1}^{m} \lambda_i \, p_S^{(i)}(s) \\ &= \sum_{i=1}^{m} \lambda_i \, p_S^{(i)}(s) f_{\mathbf{X}}(\mathbf{x}|s); \end{aligned}$$

moreover, each $p_S^{(i)}(s) f_{\mathbf{X}}(\mathbf{x}|s)$, adequately normalized, does belong to $\mathcal{P}$. So, $q_S(s|\mathbf{x})$ is a mixture of densities from $\mathcal{P}$, thus still belonging to $\mathcal{Q}^{(m)}$.

In conclusion, if $\mathcal{P}$ is a conjugate family for the class of densities $\mathcal{F}$, so is any class of m-dimensional mixture priors $\mathcal{Q}^{(m)}$ built from elements of $\mathcal{P}$.

An important feature of the families of finite mixtures of conjugate priors is their universal approximation property. As shown in [26], [35], any *a priori* probability density function (verifying some weak constraints) can be approximated arbitrarily well by a mixture of conjugate priors (of course, if the function has a complex behavior, we may need a large m to obtain a good approximation). Unfortunately, the proofs of this fact in [26] and [35] are not constructive and so we are left without a formal procedure to construct such mixtures.

If the likelihood is a mixture of elements from $\mathcal{F}$, the *a posteriori* density which results from combining this mixture likelihood with a prior from $\mathcal{P}$ does not belong to $\mathcal{P}$. This fact was evident in Example 1.5.4, where a mixture likelihood and a Gaussian prior led to a mixture *a posteriori* density.

However, it is still possible to consider mixture likelihoods and find an interesting conjugate family for them: let $\mathcal{G}$ be the family of all finite mixture likelihoods whose components are members of $\mathcal{F}$, i.e.

$$\mathcal{G} = \left\{ g_{\mathbf{X}}(\mathbf{x}|s) = \sum_{i=1}^{m} \lambda_i \, f_{\mathbf{X}}^{(i)}(\mathbf{x}|s) : \;\; m < \infty; \;\; \sum_{i=1}^{m} \lambda_i = 1; \;\; f_{\mathbf{X}}^{(i)}(\mathbf{x}|s) \in \mathcal{F} \right\}.$$

Now consider the family $\mathcal{M}$ of all finite mixture priors whose components are members of $\mathcal{P}$,

$$\mathcal{M} = \left\{ q_S(s) = \sum_{j=1}^{n} \alpha_j \, p_S^{(j)}(s) : \;\; n < \infty; \;\; \sum_{j=1}^{n} \alpha_j = 1; \;\; p_S^{(j)}(s) \in \mathcal{P} \right\}.$$

An *a posteriori* density resulting from multiplying a prior in $\mathcal{M}$ by a likelihood in $\mathcal{G}$ has the form

$$q_S(s|\mathbf{x}) \propto \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_j \, \lambda_i \, f_{\mathbf{X}}^{(i)}(\mathbf{x}|s) p_S^{(j)}(s),$$

which is, of course, still a finite mixture, specifically of dimension mn. Each of its components clearly belongs to $\mathcal{P}$ and so this *a posteriori* mixture does belong to $\mathcal{M}$, showing that this is in fact a conjugate family for $\mathcal{G}$.

Example 2.4.1

This example illustrates how a mixture of conjugate priors can be used to model a kind of prior knowledge out of the reach of a simple conjugate prior. Let s be an unknown quantity which is observed n times contaminated by independent additive Gaussian perturbations (noise) of zero mean and known variance σ^2; the corresponding likelihood function is the one in

Eq. (1.88) (Example 1.5.2). Now let us suppose that we want to express the prior knowledge that s should be close to zero with high probability, but, with lower probability, it can be very far from zero. This type of behavior calls for what is known as a *heavy-tailed* density; i.e., one that is concentrated around its mean, but does not fall to zero as fast as a Gaussian does. One way to achieve such a behavior is through the adoption of a mixture prior involving a low variance and a high variance Gaussian densities, as has been recently proposed for signal processing applications in [25]; specifically,

$$p_S(s) = w_0 \mathcal{N}(s|0, \phi_0^2) + (1 - w_0)\mathcal{N}(s|0, \phi_1^2) \tag{2.35}$$

An example of such a mixture density (with $w_0 = 0.2$, $\phi_0 = 1$, and $\phi_1 = 20$)is shown in Figure 2.6. The *a posteriori* density turns out to be

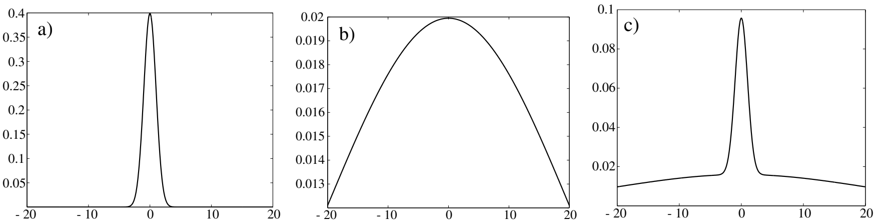

FIGURE 2.6. Mixture of conjugate priors: (a,b) the two components of the mixture; (c) the mixture. See text for parameters.

$$p_S(s|\mathbf{x}) = \frac{\left[w_0 \mathcal{N}(s|0, \phi_0^2) + (1 - w_0)\mathcal{N}(s|0, \phi_1^2)\right] \prod_{i=1}^n \mathcal{N}(x_i|s, \sigma^2)}{w_0 \prod_{i=1}^n \mathcal{N}(x_i|0, \phi_0^2 + \sigma^2) + (1 - w_0) \prod_{i=1}^n \mathcal{N}(x_i|0, \phi_1^2 + \sigma^2)}$$

which, after some straightforward computations, can be given a mixture form

$$\begin{aligned} p_S(s|\mathbf{x}) \quad = \quad & w_0'(\mathbf{x})\mathcal{N}\left(s|\frac{\bar{x}\phi_0^2}{\phi_0^2 + \frac{\sigma^2}{n}}, \frac{\sigma^2\phi_0^2}{n\phi_0^2 + \sigma^2}\right) + \\ & (1 - w_0'(\mathbf{x}))\mathcal{N}\left(s|\frac{\bar{x}\phi_1^2}{\phi_1^2 + \frac{\sigma^2}{n}}, \frac{\sigma^2\phi_1^2}{n\phi_1^2 + \sigma^2}\right) \end{aligned} \tag{2.36}$$

where $\bar{x} = (x_1 + x_2 + ... + x_n)/n$ is the observed sample mean. Notice that each of the Gaussian components of this *a posteriori* density is similar to the one that would be obtained with a single Gaussian prior with the corresponding parameters (see Example 1.5.2). The (*a posteriori*) weight $w_0'(\mathbf{x})$, which is a function of the observations, is given by

$$w_0'(\mathbf{x}) = \frac{w_0 \prod_{i=1}^n \mathcal{N}(x_i|0, \phi_0^2 + \sigma^2)}{w_0 \prod_{i=1}^n \mathcal{N}(x_i|0, \phi_0^2 + \sigma^2) + (1 - w_0) \prod_{i=1}^n \mathcal{N}(x_i|0, \phi_1^2 + \sigma^2)}.$$

From Eq. (2.36) it is now possible to obtain the PM and MAP estimates. The PM estimate is (since the mean of a mixture is simply the weighted average of the means of its components) simply

$$\widehat{s}_{\text{PM}} = w_0'(\mathbf{x}) \frac{\bar{x}\phi_0^2}{\phi_0^2 + \frac{\sigma^2}{n}} + (1 - w_0'(\mathbf{x})) \frac{\bar{x}\phi_1^2}{\phi_1^2 + \frac{\sigma^2}{n}}.$$

The MAP estimate is not so simple to obtain due to the complex modal behavior of mixtures of Gaussians. Specifically, the probability density of the mixture of two Gaussians can be either unimodal or bimodal, depending on the particular arrangement of the involved parameters (for details, see [103]). In any case, it can be obtained numerically by a simple line search algorithm.

Unlike what was obtained with a single Gaussian prior, both the PM and MAP estimates are now non-linear functions of the observations (due to the dependence of $w_0'(\mathbf{x})$ on $\mathbf{x}$. Further insight into these estimators is obtained from plots of $\widehat{s}_{\text{PM}}$ and $\widehat{s}_{\text{MAP}}$ versus x (for simplicity, we take $n = 1$, thus $\bar{x} = x$), shown in Figure 2.7. In both cases, $\phi_0 = 1$, $\phi_1 = 10$, $w_0 = 0.5$, and $\sigma = 4$. Looking first at the MAP criterion, its switching nature is very evident; it behaves as if the data chose one of the two modes of the mixture as prior. For small values of x, the low variance component dominates, and the estimate is a strongly shrunk version of the observation (the slope of the curve near the origin is small); when the observation is large enough, the situation is reversed and the resulting estimate is a slightly shrunk version of the observation (far from the origin, the slope of the curve is close to one).

The response of the PM criterion can be seen as a "smoothed" version of the MAP. Rather than a hard choice of one of the modes of the posterior (as in the MAP), it simply recomputes the weights as functions of the observation. This is another manifestation of the more "conservative" nature of the PM criterion.

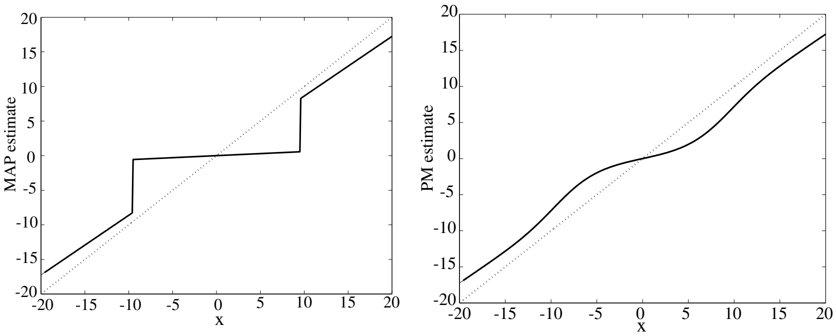

FIGURE 2.7. MAP and PM estimates versus observed value, for the mixture prior considered in Example 2.4.1.

End of Example 2.4.1

2.5 Asymptotic Behavior Of Bayes' Rules

An important conclusion that seems to result from most of the examples presented in the last section is that Bayesian estimates converge to maximum likelihood ones as the amount of observed data increases. In fact, as we shall see below, this is a reasonably general property for the MAP estimate, and also (although not so general) for the PM estimate (as, in fact, Example 2.5.1 seems to suggest).

The asymptotic behavior, when the amount of observed data goes to infinity, of Bayes' estimation rules was first studied by Laplace and more recently by von Mises [105] and Bernstein (see also [70]). The main results concern *consistency* and *asymptotical efficiency*, concepts that we will now briefly review. Let θ_0 be the true (but unknown) parameter, and $\mathbf{x}_{(n)}$ an observed data vector, containing n observations, which is a sample of the observation model $f_{\mathbf{X}}(\mathbf{x}|\theta_0)$. An estimator $\widehat{\theta} = \delta(\mathbf{x}_{(n)})$ is said to be *consistent* if

$$\lim_{n\to\infty} \delta(\mathbf{x}_{(n)}) = \theta_0, \tag{2.37}$$

where the convergence we are referring to is *in probability*. Notice that $\delta(\mathbf{x}_{(n)})$ is a random quantity because $\mathbf{x}_{(n)}$ is itself a sample from a random variable whose probability function is $f_{\mathbf{X}_{(n)}}(\mathbf{x}_{(n)}|\theta_0)$. The concept of convergence *in probability*, used in defining *consistency*, is as follows: if Y_n is a sequence of random variables, we say that this sequence converges, in probability, to some y, if, for any (arbitrarily small) $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left\{|Y_n - y| \geq \varepsilon|\right\} = 0. \tag{2.38}$$

Without going into technical details (for a deeper look at this subject see, *e.g.*, [14], [70]), the key ideas can be captured as follows: As long as the prior is continuous and not zero at the location of the ML estimate, then, the MAP estimate converges to the ML estimate. Accordingly, to establish the consistency of the MAP estimate, it is necessary to guarantee consistency of the ML estimate. Conditions for this are more technical and the interested reader is again referred to [70]. To show the convergence of the PM estimate to the ML estimate, it is generally required that some other conditions hold. Namely, the observation models $f_{\mathbf{X}}(\mathbf{x}|\theta)$ must have common support, *i.e.*, $\{\mathbf{x} : \; f_{\mathbf{X}}(\mathbf{x}|\theta) > 0\}$ must not be a function of θ (it happens that this is also a condition necessary to prove consistency of the ML estimate). Next, we present a simple example where this result applies, but where a similar convergence property does not hold for the PM criterion.

Example 2.5.1

Let us get back to uniform observation model studied in Example 2.3.5. With an arbitrary continuous prior $p_\Theta(\theta)$ (not the Pareto, which is not continuous), the *a posteriori* probability density function becomes

$$p_\Theta(\theta|\mathbf{x}) \propto p_\Theta(\theta)\, f_{\mathbf{X}}(\mathbf{x}|\theta) = p_\Theta(\theta) \begin{cases} \theta^{-n}, & \theta \geq \max(\mathbf{x}) \\ 0, & \text{otherwise} \end{cases} \tag{2.39}$$

where $\max(\mathbf{x})$ stands for the maximum of all observations. The corresponding MAP and PM estimates are

$$\begin{aligned} \widehat{\theta}_{\text{MAP}} &= \arg \max_{\theta \geq \max(\mathbf{x})} \left\{ p_\Theta(\theta)\, \theta^{-n} \right\} & (2.40) \\ \widehat{\theta}_{\text{PM}} &= \frac{\displaystyle\int_{\max(\mathbf{x})}^{\infty} p_\Theta(\theta)\, \theta^{-n+1}\, d\theta}{\displaystyle\int_{\max(\mathbf{x})}^{\infty} p_\Theta(\theta)\, \theta^{-n}\, d\theta}. & (2.41) \end{aligned}$$

This MAP estimator converges to the ML criterion as n approaches infinity; in fact,

$$\begin{aligned} \lim_{n\to\infty} \widehat{\theta}_{\text{MAP}} &= \lim_{n\to\infty} \arg \max_{\theta \geq \max(\mathbf{x})} \{\log p_\Theta(\theta) - n \log \theta\} \\ &= \lim_{n\to\infty} \arg \min_{\theta \geq \max(\mathbf{x})} \{n \log \theta\} \; = \; \max(\mathbf{x}) \; = \; \widehat{\theta}_{\text{ML}}. \end{aligned}$$

As seen in Example 2.3.5, this is not true under the conjugate Pareto prior.

End of Example 2.5.1

Another very important characteristic that may be present is *asymptotic normality.* When the necessary conditions are met (see, *e.g.*, [14]), this means that the *a posteriori* probability density function will tend to a Gaussian, as the amount of data grows to infinity. Although this is a fairly technical subject, and the reader interested in further studying the subject is referred, for example, to [14] and the references therein, we will have something more to say about it in Section 2.7.2.

2.6 Non-informative Priors and Invariance

In Example 2.2.1, we have used a uniform (improper) prior to express ignorance about the (unknown) mean of a Gaussian observation; this prior led to the *maximum likelihood* criterion. This may seem an obvious choice but it in fact touches on one of the delicate points of Bayesian theory: non-informative priors [8], [14], [27], [93]. The goal of non-informative priors (pioneered by Jeffreys [59], [60] although used, in simpler forms, as early as by Bayes and Laplace) is to formalize the concept of "ignorance". The key idea is that, associated with each parametric observation model (a likelihood function), there is a certain prior expressing ignorance about the involved parameter(s), i.e., which is non-informative. Non-informative priors do not describe *a priori* beliefs; rather, they should be seen as a way of letting the data "speak for itself", while staying inside a Bayesian approach [54]. For this reason they are often termed *non-subjective* priors.

What may not be obvious at a first thought is that the non-informative nature of a prior does depend strongly on the meaning and role of the unknown parameter, i.e., it depends on how the likelihood function is parameterized. To illustrate this fact, let us refer again to Example 2.2.1: it seemed rather obvious that the flat prior $p_S(s) = c$ expresses "total ignorance" about s. Now, let the problem be reparameterized in terms of, say, $u = \exp\{s\}$ (a one-to-one transformation which should not influence the estimation results); since we are as ignorant about u as we were about s, we are tempted to also adopt a uniform (flat) prior for it. However, the prior probability density function on u, resulting from $p_S(s) = c$, and from the application of the transformation rule for densities of functions of random variables[2], is

$$f_U(u) = p_S(\exp^{-1}\{u\}) \frac{1}{|\exp'\{\exp^{-1}\{u\}\}|} = \frac{c}{u} \propto \frac{1}{u}. \tag{2.43}$$

because $p_S(\cdot) = c$, the inverse function of the exponential is the natural logarithm $\exp^{-1}\{\cdot\} = \log\{\cdot\}$, and its derivative is still an exponential $\exp^{'}\{x\} = \exp\{x\}$; thus $\exp'\{\exp^{-1}\{u\}\} = u$. In conclusion, the non-informative prior for u resulting from a uniform prior for s is not uniform, which means that ignorance is not necessarily expressed by a uniform prior. Moreover, it also became clear that the uniform prior is not invariant under an exponential transformation; but is it invariant under any class of parameter transformations? And is this class relevant? These questions are still surrounded by some controversy. In fact, for many problems, different considerations may lead to different non-informative priors, and it may not be clear which is the relevant invariance that should be imposed. However, there are three cases that are less controversial: discrete problems (e.g., classification), location parameters, and scale parameters. We next briefly address these three classes of problems to illustrate the type of reasoning involved in designing non-informative prior densities through invariance

[2] Recall that if X is a continuous r.v. with p.d.f. $f_X(x)$ and $g(\cdot)$ is a one-to-one continuous function, then $Y = g(X)$ has p.d.f. given by

$$f_Y(y) = f_X(g^{-1}(y)) \frac{1}{|g'(g^{-1}(y))|}, \tag{2.42}$$

where $g^{-1}(\cdot)$ denotes the inverse function of $g(\cdot)$ (which exists because $g(\cdot)$ is one-to-one) and $g^{'}(\cdot)$ is its derivative.

arguments. For a more complete treatment of this topic, see the excellent review in [27] or more advanced Bayesian theory textbooks such as [14] or [93].

2.6.1 Discrete Problems

For discrete finite (e.g., classification) problems, the (one-to-one) transformations under which it is natural to require invariance are index permutations. In simpler words, if one knows nothing *a priori*, permutations of the set $\mathcal{S}$ should add no knowledge or in any way influence the result of the inference procedure. The only probability mass function that is invariant under permutations is the uniform one $p(s) = 1/|\mathcal{S}|$ (where, recall, $|\mathcal{S}|$ is the number of elements in $\mathcal{S}$). When $\mathcal{S}$ is discrete but infinite, the argument is more elaborate, but the same conclusion is still valid [14].

There are, however, some difficulties with uniform priors on discrete sets when we consider non one-to-one transformations. Consider that the set of possible states of nature is initially believed to be $\mathcal{S} = \{s_1, s_2\}$; complete ignorance is modeled by equiprobability, $p_S(s_1) = p_S(s_2) = 1/2$. Suppose that it is then found that s_2 is in fact composed of two sub-states, say s_2^a and s_2^b, so that now $\mathcal{S} = \{s_1, s_2^a, s_2^b\}$; now ignorance about this new set of possible states of nature would be expressed by $p_S(s_1) = p_S(s_2^a) = p_S(s_2^b) = 1/3$ which is not compatible with $p_S(s_2) = p_S(s_2^a) + p_S(s_2^b)$.

2.6.2 Location Parameters

Consider an estimation problem ($\mathcal{S} = \mathcal{A} = I\!R$) where the likelihood function has the form $f_X(x|s) = \phi(x - s)$, for some function $\phi(\cdot)$; s then is called a *location parameter*. A simple example is $f_X(x|s) = \mathcal{N}(x|s, \sigma^2)$, with known σ^2. Now, suppose that rather than observing one outcome of X, say x, a shifted version $y = x + k$ (with k known) is observed, the new goal being to estimate $u = s + k$; clearly, the likelihood function for this new problem is still $f_Y(y|u) = \phi(y - u)$. Obviously, estimating u from y or s from x are equivalent problems and none of them should be privileged with respect to the other; accordingly, there is no reason why the non-informative prior density would not be the same for both of them. Denoting as $p_S(s)$ the non-informative prior density for the (x, s) problem, and as $p_U(u)$ the one for the (y, u) problem, we are then forced to require that, for any interval $[\alpha_1, \alpha_2]$,

$$\int_{\alpha_1}^{\alpha_2} p_S(s)\, ds = \int_{\alpha_1}^{\alpha_2} p_U(u)\, du. \tag{2.44}$$

But

$$\int_{\alpha_1}^{\alpha_2} p_U(u)\, du = \int_{\alpha_1}^{\alpha_2} p_S(u - k)\, du = \int_{\alpha_1 - k}^{\alpha_2 - k} p_S(s)\, ds. \tag{2.45}$$

Equality between the left hand side of Eq. (2.44) and the third integral in Eq. (2.45) is clearly a translation invariance condition, i.e. $p_S(s) = p_S(s - k)$, for any s and k. The only solution is $p_S(s) = a$, with arbitrary a (the same would, of course, be concluded about $p_U(u)$). In summary, for a location parameter, translation invariance arguments lead to a uniform (improper) non-informative prior density (see Figure 2.8).

2.6.3 Scale Parameters

Consider now that s is a *scale parameter*, i.e., the likelihood function has the form $f_X(x|s) = s^{-1}\psi(x/s)$, for some function $\psi(\cdot)$. For example, if $f_X(x|s) = \mathcal{N}(x|0, s^2)$, the standard deviation s is a *scale parameter*; of course now $\mathcal{S} = \mathcal{A} = (0, +\infty)$. In this case, it is natural to expect the non-informative prior density to express *"scale ignorance"*, and so the relevant reparameterization is $y = kx$ which corresponds to $u = ks$, i.e., a change of scale. Scale invariance means that the units in which some quantity are measured should not influence any conclusions drawn from it. Using the same line of thought as above, since the two problems are equivalent, the two non-informative prior densities have to coincide; so, for any (α_1, α_2)

$$\int_{\alpha_1}^{\alpha_2} p_S(s)\, ds = \int_{\alpha_1}^{\alpha_2} p_U(u)\, du. \tag{2.46}$$

But

$$\int_{\alpha_1}^{\alpha_2} p_U(u)\, du = \int_{\alpha_1}^{\alpha_2} p_S(u/k)\frac{1}{k}\, du = \int_{\alpha_1/k}^{\alpha_2/k} p_S(s)\, ds. \tag{2.47}$$

which obviously expresses scale invariance (as shown in Figure 2.8). Now, the equality between the left hand side of Eq. (2.46) and the second integral in Eq. (2.47) requires that $p_S(s) = p_S(s/k)(1/k)$ for any s and k; in particular, since it also has to be true for $s = k$, it requires that $p_S(s) = p_S(1)(1/k) \propto (1/k)$. In conclusion, a scale invariant non-informative prior density is not uniform, though it is still improper.

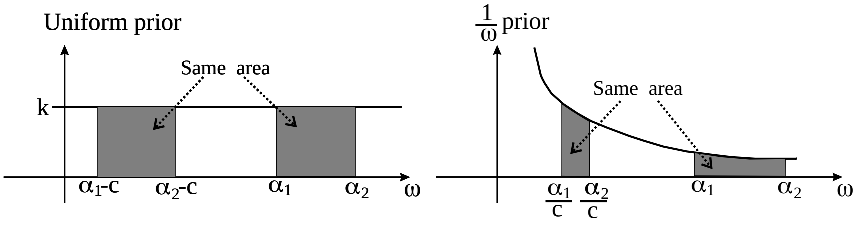

FIGURE 2.8. Non-informative priors for location (left) and scale (right) parameters.

At this point, the following question seems to be relevant: since the use of conjugate priors was advocated on the basis of analytical and computational feasibility, what about this non-informative scale-invariance prior we just studied? Does it exhibit any kind of conjugacy? We will now take a brief look at this issue by revisiting previous examples for which we considered conjugate priors. We shall see that in all three cases, the non-informative prior density belongs to the family of conjugate densities; possibly, lying at the "boundary" of the family as a limit case.

A similar fact was already verified for location parameters: simply note that the uniform prior may be seen as the limit of a family of Gaussian densities, as the variance goes to infinity.

Example 2.6.1

In Example 2.3.2, we took the uniform density on the interval $[0,1]$ as a non-informative prior for the Bernoulli parameter, which led to its ML estimate. Now we can notice that by inserting $\alpha = \beta = 1$ in the Beta prior we do obtain this uniform prior showing that it is, in fact, a particular instance of the conjugate family. In this case, because the parameter space is bounded (it is the interval $[0,1]$), this prior density is not improper.

Example 2.6.2

Let us now take another look at Example 2.3.3. How can we obtain a non-informative conjugate prior? In other words, is there a Gamma density (recall Eq. (2.17)) which expresses ignorance about the unknown parameter θ. Recall that the mean, the variance, and the mode of a Gamma density, $\text{Ga}(\theta|\alpha,\beta)$, are equal to α/β, α/β^2, and $(\alpha-1)/\beta$, respectively (as expressed in Eqs. (2.19), (2.20), and (2.21)). Moreover, the mode only exists if $\alpha \geq 1$. By letting $\alpha \to 0$ and $\beta \to 0$, the mean becomes indeterminate, the variance goes to $+\infty$ and the limit density has no mode, which seems to be a reasonably vague prior. This limit Gamma density is improper; in fact, the inverse of the normalizing constant in Eq. (2.17) verifies

$$\lim_{\alpha,\beta\to 0} \frac{\beta^\alpha}{\Gamma(\alpha)} = 0. \tag{2.48}$$

To obtain the "shape" of this improper prior, one simply ignores the normalizing constant,

$$\lim_{\alpha,\beta\to 0} \theta^{\alpha-1} e^{-\beta\theta} = \frac{1}{\theta} \tag{2.49}$$

which is our improper non-informative prior (notice in Figure 2.3, how for $\alpha = \beta = 1$ the Gamma density starts resembling the $1/\theta$ function). Recalling that $\theta = 1/\sigma^2$, i.e., $\sigma = \theta^{-1/2}$, we obtain the corresponding prior for σ simply by using the transformation rule (Eq. (2.42)); the result is a similar prior $p_\Sigma(\sigma) \propto 1/\sigma$, which interestingly agrees with the one obtained via scale invariance arguments.

Another way to look at this prior is by considering the transformation $\omega = \log(\sigma)$, i.e., parameter ω represents scale in logarithmic units. By a direct application of the density transformation rule, we obtain $p_\Omega(\omega) \propto 1$, a uniform prior over $I\!R$; notice that the $\omega = log(\sigma)$ transformation maps $\sigma \in (0, +\infty)$ onto $\omega \in (-\infty, +\infty)$. An intuitively appealing explanation for this fact is that it is more natural to think of scale parameters in logarithmic terms; with respect to that (most natural) parameterization, the non-informative prior becomes uniform.

Finally, to obtain the Bayesian rules corresponding to this non-informative prior, let $\alpha, \beta \to 0$ in Eqs. (2.23) and (2.24):

$$\lim_{\alpha,\beta\to 0} \widehat{\theta}_{\mathrm{PM}} = n\left(\sum_{i=1}^{n} x_i^2\right)^{-1} = \widehat{\theta}_{\mathrm{ML}} \tag{2.50}$$

$$\lim_{\alpha,\beta\to 0} \widehat{\theta}_{\mathrm{MAP}} = (n-2)\left(\sum_{i=1}^{n} x_i^2\right)^{-1} < \widehat{\theta}_{\mathrm{ML}}. \tag{2.51}$$

The PM and ML estimates become coincident, but the MAP criterion outputs a smaller value (in what may be seen as the effect of the $1/\theta$ prior "pulling-down" the estimate).

Example 2.6.3

Is it possible to follow the line of thought of the previous example, but now for the Poisson observation model of Example 2.3.4? We saw that the conjugate family for the Poisson likelihoods (parameterized as in Eq. (2.25)) is also the set of Gamma priors (see Eq. (2.17)). So, the natural thing to do in order to obtain a non-informative prior density is again to consider the limit $\alpha, \beta \to 0$. As in the previous example, the limit prior thus obtained is

$$\lim_{\alpha,\beta\to 0} \theta^{\alpha-1} e^{-\beta\theta} = \frac{1}{\theta} \tag{2.52}$$

which seems to suggest that the rate parameter of a Poisson distribution should also be treated as a scale parameter; i.e., a non-informative prior for θ should express scale invariance. But what kind of scale invariance is being invoked? The answer is clear if one notices that θ is a "rate" and so its units are "counts/time-interval" (where "time-interval" may be replaced by some "length", "area", or "volume" if we are talking about a spatial Poisson process, rather than a time one). Of course, the units used to measure time intervals should not influence the results of any inference procedure and so the relevant reparameterization is in fact a change of scale; the $1/\theta$ prior does express this scale invariance.

The Bayesian estimators of θ under this non-informative prior are obtained as the limits when $\alpha, \beta \rightarrow 0$ of those in Eqs. (2.27) and (2.28),

$$\lim_{\alpha,\beta\rightarrow 0} \widehat{\theta}_{\mathrm{PM}} = \frac{\sum_{j=1}^{n} x_j}{T} = \widehat{\theta}_{\mathrm{ML}} \tag{2.53}$$

$$\lim_{\alpha,\beta\rightarrow 0} \widehat{\theta}_{\mathrm{MAP}} = \frac{-1 + \sum_{j=1}^{n} x_j}{T} < \widehat{\theta}_{\mathrm{ML}}. \tag{2.54}$$

As in the previous example, the PM estimate converges to the ML one. The limit of the MAP estimate is slightly smaller than the ML estimate but this difference naturally vanishes as the amount of data (and T) increases.

2.7 Jeffreys' Priors and Fisher Information

A common criticism of the invariance approach to obtaining non-informative priors is that it relies on the choice of an invariance structure; although for discrete problems and for location and scale parameters there seems to be no controversy (probably because most approaches lead to the same results), this may not be the case for other types of parameters. An alternative technique which does not rely (explicitly) on the choice of an invariance structure (although it often leads to similar results) is the one proposed by Jeffreys [60].

2.7.1 Fisher Information

Before addressing the central topic of this section, the so-called *Jeffreys' non-informative priors*, the underlying concept of *Fisher information* has to be introduced. If an estimation problem is characterized by the likelihood function $f_{\mathbf{X}}(\mathbf{x}|\theta)$, the associated *Fisher information* is defined as

$$\mathcal{I}(\theta) = E_{\mathbf{X}}\left[\left(\frac{\partial \log f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial\theta}\right)^2\right]. \tag{2.55}$$

The Fisher information is a very important concept introduced by Fisher in 1922 [42]. Its best known use is in the Cramer-Rao (lower) bound (see, e.g., [63], [98] or [104]) which relates to basic frequentist concepts which we will now briefly review. Let θ be a real (deterministic) parameter to be estimated from observations $\mathbf{x}$ which are generated according to some likelihood function $f_{\mathbf{X}}(\mathbf{x}|\theta)$, and let $\widehat{\theta}(\mathbf{x})$ be some estimator of θ. If this estimator verifies

$$E_{\mathbf{X}}\left[\widehat{\theta}(\mathbf{x})\right] = \int_{-\infty}^{+\infty} \widehat{\theta}(\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} = \theta \qquad (2.56)$$

it is said to be *unbiased.* Notice the frequentist nature of this concept; it is an average with respect to all possible observations, with fixed θ. Also, compare this with Eq. (1.2). The most important (frequentist) characterization of a (unbiased) parameter estimator is its variance,

$$E_{\mathbf{X}}\left[\left(\widehat{\theta}(\mathbf{x}) - \theta\right)^2\right] = \int_{-\infty}^{+\infty} \left(\widehat{\theta}(\mathbf{x}) - \theta\right)^2 f_{\mathbf{X}}(\mathbf{x}|\theta)\ d\mathbf{x}, \qquad (2.57)$$

which has been very widely used to support optimality criteria in statistical signal processing [63], [98]. A particularly important relation involving the variance of an unbiased estimator and the Fisher information is the *Cramer-Rao bound*[3] which states that, if $\widehat{\theta}(\mathbf{x})$ is an unbiased estimator,

$$E_{\mathbf{X}}\left[\left(\widehat{\theta}(\mathbf{x}) - \theta\right)^2\right] \geq \frac{1}{\mathcal{I}(\theta)}. \qquad (2.58)$$

Estimators verifying Eq. (2.58) with strict equality are called *efficient.* In order to have an intuitive interpretation of the Cramer-Rao inequality, it is convenient to rewrite the Fisher information in its alternative form; in fact, it is not difficult to verify that the Fisher information can be rewritten as (see, *e.g.*, [104])

$$E_{\mathbf{X}}\left[\left(\frac{\partial \log f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial\theta}\right)^2\right] = -E_{\mathbf{X}}\left[\frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial\theta^2}\right]. \qquad (2.59)$$

Now this form has a clearer meaning: it can be roughly described as the "average concavity" (second derivative) of the logarithm of the likelihood function, i.e., it measures how prominent its maximum is; it is intuitively acceptable that a parameter can be more accurately estimated if the associated log-likelihood has a clear maximum.

At this point, let us go back and compute Cramer-Rao bounds for some of the previous examples, in order to further clarify this concept.

Example 2.7.1

Recalling Example 2.3.1, when θ is the common mean of a set of n i.i.d. normal observations of known variance σ^2, the log-likelihood function is

$$\log f_{\mathbf{X}}(\mathbf{x}|\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2. \qquad (2.60)$$

[3] See, e.g., [63], [98], or [104] for derivations; see also [23] for a different approach from an information theoretical viewpoint

Now consider the ML estimator of θ, i.e., $\widehat{\theta}_{\mathrm{ML}}(\mathbf{x})$ which is easily seen to be unbiased:

$$E_{\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E_{X_i}[x_i] = \theta.$$

The Fisher information can be found by applying Eq. (2.59),

$$\mathcal{I}(\theta) = \frac{n}{\sigma^2} \tag{2.61}$$

which can be read as: the amount of information about θ carried by the likelihood function is proportional to the size of the observation set and decreases as the observation variance increases; this is an intuitively pleasing result. The Cramer-Rao bound, in this case, will state that no unbiased estimator can have a variance less than σ^2/n; moreover, the fact that $\mathcal{I}(\theta)$ in Eq. (2.61) is not a function of θ says that the estimation precision is not expected to depend on the true θ. Some further simple manipulations allow us to show that the ML estimator is in fact *efficient*, i.e., its variance equals σ^2/n.

End of Example 2.7.1

Example 2.7.2

Let us revisit Example 2.3.2; there, θ and $1-\theta$ denoted the (unknown) probabilities of heads and tails, respectively, of a given coin under study. The outcomes of an observed sequence of n tosses are denoted by $\mathbf{x} = (x_1, \ldots, x_n)$, with $x_i = 1$ standing for a head, and $x_i = 0$ for a tail. The likelihood function is then a Bernoulli distribution; i.e., letting $n_h(\mathbf{x}) = x_1 + x_2 + ... + x_n$ denote the number of heads outcomes, it can be written as

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})}. \tag{2.62}$$

The maximum likelihood estimate of θ was seen to be $\widehat{\theta}(\mathbf{x}) = n_h(\mathbf{x})/n$; to check that this is an unbiased estimate, let us compute its frequentist expected value

$$E_{\mathbf{X}}[n_h(\mathbf{x})/n] = \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{X}} n_h(\mathbf{x})\,\theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})}$$

where $\mathcal{X}$ is the space of all possible sequences of heads and tails. This sum can be rearranged by counting how many sequences have each possible number of heads outcomes, i.e.

$$\begin{aligned}&\frac{1}{n}\sum_{\mathbf{x}\in\mathcal{X}} n_h(\mathbf{x})\theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})} = \\ &\qquad \frac{1}{n}\sum_{n_h(\mathbf{x})=0}^{n} n_h(\mathbf{x})\binom{n}{n_h(\mathbf{x})}\theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})} = \theta\end{aligned} \tag{2.63}$$

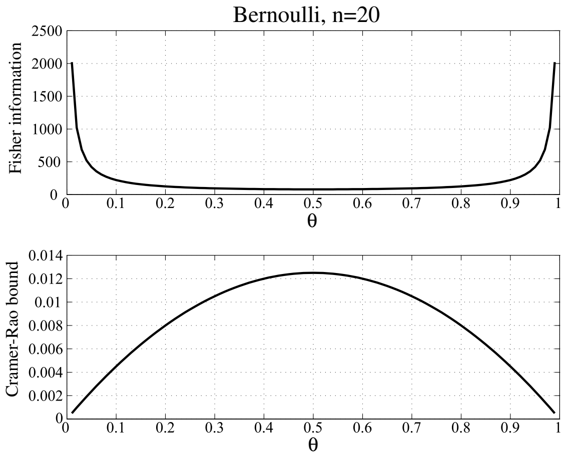

FIGURE 2.9. The Fisher information and the corresponding Cramer-Rao bound for the problem of estimating the parameter of a Bernoulli sequence of 20 trials.

because the summation on the right hand side can be easily identified as the mean of a Binomial random variable of "sample size" n and "probability of success" equal to θ (see Appendix A), which is $n\theta$.

Obtaining the Fisher information $\mathcal{I}(\theta)$ via Eq. (2.59) leads (after some computations) to

$$\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)} \tag{2.64}$$

which is plotted in Figure 2.9. Here we have a different situation: $\mathcal{I}(\theta)$ does depend on θ. This means that some values of the parameter θ will be more difficult to estimate than others, with $\theta = 0.5$ being the value for which estimators are expected to have higher variance.

End of Example 2.7.2

Example 2.7.3

In Example 2.3.3 we looked into the problem of estimating the inverse of the common variance of a set of zero mean i.i.d. Gaussian random variables. Let us now parameterize the problem directly on the variance σ^2. Here again, it is easy to check that the ML estimator is unbiased, and some further computations yield

$$\mathcal{I}(\sigma^2) = \frac{n}{2\sigma^4}. \tag{2.65}$$

Again, this is not a surprising result; it is natural that the variance of a variance estimate will depend on the fourth moment of the underlying variable.

End of Example 2.7.3

The expressions for the Fisher information found in the previous three examples (see Eqs. (2.61), (2.64), and (2.65)) seem to exhibit a common feature: they are proportional to n, the size of the observed data set. In fact, this is a general result for sets of i.i.d. observations; it is a simple consequence of the use of logarithm in the definition of Fisher information and of the important fact

$$E_{\mathbf{X}}\left[\frac{\partial \log f_{\mathbf{X}}(\mathbf{x})|\theta)}{\partial \theta}\right] = \int \frac{\partial \log f_{\mathbf{X}}(\mathbf{x})|\theta)}{\partial \theta}\, f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} = 0, \tag{2.66}$$

whose proof is simply (recall that for any function u, $d\log(u(x))/dx = (1/u(x))du(x)/dx$)

$$\begin{aligned}\int \frac{\partial \log f_{\mathbf{X}}(\mathbf{x})|\theta)}{\partial \theta}\, f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} &= \int \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta}\, \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{x}|\theta)}\, d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \underbrace{\int f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x}}_{1} = 0. \end{aligned} \tag{2.67}$$

Now, let the *elementary* (i.e., for one observation) likelihood be $f_X(x|\theta)$, with the corresponding Fisher information being denoted as $\mathcal{I}^{(1)}(\theta)$. Given n i.i.d. observations $\mathbf{x} = (x_1, \ldots, x_n)$, all obtained according to this same likelihood function, i.e., $f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$ ($f_{X_i}(x_i|\theta) = f_X(x_i|\theta)$, for any i), the associated Fisher information, denoted $\mathcal{I}^{(n)}(\theta)$, is

$$\begin{aligned}\mathcal{I}^{(n)}(\theta) &= E_{\mathbf{X}}\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f_X(x_i|\theta)\right)^2\right] \\ &= \sum_{i=1}^{n} \underbrace{E_{X_i}\left[\left(\frac{\partial \log f_{X_i}(x_i|\theta)}{\partial\theta}\right)^2\right]}_{\mathcal{I}^{(1)}(\theta),\ \text{independently of } i} \\ &\quad + \sum_{i\neq j} \underbrace{E_{X_i,X_j}\left[\frac{\partial \log f_{X_i}(x_i|\theta)}{\partial\theta}\,\frac{\partial \log f_{X_j}(x_j|\theta)}{\partial\theta}\right]}_{0} \qquad (2.68) \\ &= n\,\mathcal{I}^{(1)}(\theta), \qquad (2.69)\end{aligned}$$

where the second term in Eq. (2.68) is zero as a consequence of the independence of X_i and X_j and of Eq. (2.66).

Let us conclude these few paragraphs devoted to the Fisher information with a final example involving a Poisson likelihood function.

Example 2.7.4

For the Poisson observation model, according to Eq.(2.25), the log-likelihood function is (up to irrelevant additive constants)

$$\log f_{\mathbf{X}}(\mathbf{x}|\theta) = t(\mathbf{x})\log\theta - T\theta;$$

see Example 2.3.4 to recall the notation. The ML estimate of θ is given by Eq. (2.29). Checking that it is unbiased is simple,

$$E_{\mathbf{X}}\left[\frac{\sum_{i=1}^n x_i}{T}\right] = \frac{1}{T}\sum_{i=1}^n E_{X_i}[x_i] = \frac{1}{T}\sum_{i=1}^n \theta T_i = \theta,$$

since the expected number of counts of a Poisson process of rate θ during an observation interval T_i is simply θT_i (see Appendix A).

Now, some straightforward computations allow obtaining the Fisher information for this observation model, which turns out to be

$$\mathcal{I}(\theta) = \frac{T}{\theta}; \tag{2.70}$$

this expression is clearly similar to Eq. (2.61), if we recall that the variance (which coincides with the mean) of the number of Poisson counts during a unit interval is θ. Here, the total observation time T plays the role of "size of data set" n in the previous examples. In addition, this expression suggests that smaller values of θ can be estimated with less variance; this is a natural consequence of the fact that the mean and the variance of a Poisson distribution coincide.

End of Example 2.7.4

2.7.2 Fisher Information in Asymptotic Normality

Following [14], let us briefly see what can be stated about the asymptotic properties of the *a posteriori* probability density function, $p_S(\theta|\mathbf{x})$ (with θ a real parameter), obtained from a likelihood function $f_{\mathbf{X}}(\mathbf{x}|\theta)$ together with a prior $p_\Theta(\theta)$. Let us assume that the n observations are independent and identically distributed, *i.e.*,

$$\log f_{\mathbf{X}}(\mathbf{x}|\theta) = \sum_{i=1}^n \log f_X(x_i|\theta)$$

and write the *a posteriori density* as

$$p_\Theta(\theta|\mathbf{x}) \propto \exp\left\{\log p_\Theta(\theta) + \log f_{\mathbf{X}}(\mathbf{x}|\theta)\right\}. \tag{2.71}$$

Now, let θ_0 be the maximum of $p_\Theta(\theta)$ (assumed to exist and be unique) and $\widehat{\theta}_{(n)}^{\text{ML}}$ be the maximum likelihood estimate obtained from the n observations, that is, the maximum of $\log f_{\mathbf{X}}(\mathbf{x}|\theta)$ with respect to θ. Performing Taylor expansions of the logarithmic terms in Eq. (2.71), we can write

$$\begin{aligned}
\log p_\Theta(\theta) &= \log p_\Theta(\theta_0) - \frac{1}{2}(\theta-\theta_0)^2 \left.\frac{\partial^2 \log p_\Theta(\theta)}{\partial\theta^2}\right|_{\theta=\theta_0} + R_0 \\
\log f_{\mathbf{X}}(\mathbf{x}|\theta) &= \log f_{\mathbf{X}}(\mathbf{x}|\widehat{\theta}_{(n)}^{\text{ML}}) - \frac{1}{2}(\theta-\widehat{\theta}_{(n)}^{\text{ML}})^2 \left.\frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial\theta^2}\right|_{\theta=\widehat{\theta}_{(n)}^{\text{ML}}} + R_{(n)}
\end{aligned}$$

where R_0 and $R_{(n)}$ are remainder terms. Now, these expansions, if they are accurate, allow us to look at the prior and the likelihood as Gaussian; this is so if we can treat R_0 and $R_{(n)}$ as constants, which of course they are not. Observing that (see Eq. (2.69))

$$\left.\frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}^{\mathrm{ML}}_{(n)}} = \mathcal{I}^{(n)}\left(\widehat{\theta}^{\mathrm{ML}}_{(n)}\right) = n\,\mathcal{I}^{(1)}\left(\widehat{\theta}^{\mathrm{ML}}_{(n)}\right)$$

we obtain

$$p_\Theta(\theta|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}(\theta-\widehat{\theta}_n)^2\left(n\,\mathcal{I}^{(1)}\left(\widehat{\theta}^{\mathrm{ML}}_{(n)}\right) + \left.\frac{\partial^2 \log p_\Theta(\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\right)\right\}$$

where (compare with Eq. (1.85))

$$\widehat{\theta}_n = \frac{n\,\mathcal{I}^{(1)}\left(\widehat{\theta}^{\mathrm{ML}}_{(n)}\right)\widehat{\theta}^{\mathrm{ML}}_{(n)} + \left.\frac{\partial^2 \log p_\Theta(\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\theta_0}{n\,\mathcal{I}^{(1)}\left(\widehat{\theta}^{\mathrm{ML}}_{(n)}\right) + \left.\frac{\partial^2 \log p_\Theta(\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}}.$$

Now, as the amount of data increases, $n \to \infty$, the prior term looses importance, *i.e.*, the term $n\,\mathcal{I}^{(1)}\left(\widehat{\theta}^{\mathrm{ML}}_{(n)}\right)$ dominates over the second derivative of the log-prior; consequently,

$$\lim_{n\to\infty} p_\Theta(\theta|\mathbf{x}) = \lim_{n\to\infty} \mathcal{N}\left(\theta|\widehat{\theta}^{\mathrm{ML}}_{(n)}, \left(n\,\mathcal{I}^{(1)}\left(\widehat{\theta}^{\mathrm{ML}}_{(n)}\right)\right)^{-1}\right). \tag{2.72}$$

Of course, we have omitted all the detailed technical conditions required, and what we have presented is not a formal proof but only an heuristic view. It serves mainly the purpose of briefly showing the kind of results that are obtained in asymptotic analysis. There is a large body of work on this issue, and a comprehensive list of references can be found, for example, in [14].

2.7.3 Jeffreys' Priors

After this brief view of the Fisher information and Cramer-Rao bound, which was followed by a few examples, let us return to the Jeffreys' prior. We begin by considering an estimation problem supported on the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$. A reparameterization of the problem into $u = g(s)$, where $g(\cdot)$ is any one-to-one continuous function, corresponds to a new likelihood function[4], say $f'_{\mathbf{X}}(\mathbf{x}|u)$. Because $g(\cdot)$ is a one to one mapping, these two

[4] It is important to use a notation $f'_{\mathbf{X}}(\mathbf{x}|u)$ that distinguishes this new likelihood because its functional dependence on u is not the same as the functional dependence of $f_{\mathbf{X}}(\mathbf{x}|s)$ on s.

likelihoods are of course related through

$$f'_{\mathbf{X}}(\mathbf{x}|u) = f_{\mathbf{X}}(\mathbf{x}|g^{-1}(u)) \tag{2.73}$$

where $g^{-1}(\cdot)$ is the inverse function of $g(\cdot)$. Then, by the chain rule of the derivative

$$\begin{aligned} \frac{\partial \log f'_{\mathbf{X}}(\mathbf{x}|u)}{\partial u} &= \left.\frac{\partial \log f_{\mathbf{X}}(\mathbf{x}|s)}{\partial s}\right|_{s=g^{-1}(u)} \cdot \frac{dg^{-1}(u)}{du} \\ &= \left.\frac{\partial \log f_{\mathbf{X}}(\mathbf{x}|s)}{\partial s}\right|_{s=g^{-1}(u)} \cdot \frac{1}{g'(g^{-1}(u))}. \end{aligned} \tag{2.74}$$

Now, squaring both sides, taking expected values with respect to X (notice that $1/g'(g^{-1}(u))$ is not a function of X), and then extracting square roots, leads to

$$\sqrt{\mathcal{I}(u)} = \sqrt{\mathcal{I}(g^{-1}(u))}\frac{1}{|g'(g^{-1}(u))|}. \tag{2.75}$$

The reader should notice how this equation is similar to the random variable transformation rule in Eq. (2.42)). The implication of this similarity is clear: if s has a prior $p(s) \propto \sqrt{\mathcal{I}(s)}$, which is called the Jeffreys' prior, then any (one-to-one and continuous) reparameterization $u = g(s)$ will automatically lead to the prior $p(u) \propto \sqrt{\mathcal{I}(u)}$, still the Jeffreys' prior. This shows how the Jeffreys' prior generalizes the location and scale invariance-based non-informative priors to arbitrary (continuous and one-to-one) reparameterizations.

Finally, notice that, in passing, we also derived the transformation rule for the Fisher information; removing the square roots from Eq. (2.75),

$$\mathcal{I}(u) = \mathcal{I}(g^{-1}(u))\left(\frac{1}{|g'(g^{-1}(u))|}\right)^2. \tag{2.76}$$

Example 2.7.5

In Example 2.7.1, where θ is the common mean of a set of n i.i.d. normal observations of known variance σ^2, we saw that the Fisher information is expressed in Eq. (2.61); accordingly, Jeffreys' prior will be

$$p_{\Theta}(\theta) \propto k \tag{2.77}$$

where k is an arbitrary constant (since this prior is improper, it is only defined up to an arbitrary multiplicative factor). Notice that this prior coincides with the one obtained in Section 2.6.2 through the use of location invariance considerations.

End of Example 2.7.5

Example 2.7.6

For a set of n Bernoulli trials studied in Example 2.7.2, the Fisher information for the parameter θ was found to be as given in Eq. (2.64). Consequently, the non-informative prior according to Jeffreys' principle is

$$p_\Theta(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}. \tag{2.78}$$

The first interesting thing to notice about this prior, is that it is still a conjugate prior; notice (see Eq. (2.10)) that this is a Beta density, more precisely, $p_\Theta(\theta) = \text{Be}(\theta|1/2, 1/2)$. The next thing to notice is that $\text{Be}(\theta|1/2, 1/2)$ is not uniform on $[0, 1]$, as could be intuitively expected from a non-informative prior for this problem (recall Example 2.7.2); see [56] for an interesting discussion of this Jeffreys' prior. However, a look at the Bayesian estimation rules in Eqs. (2.14) and (2.15) reveals that, even for moderately large samples, the impact of this difference is small.

End of Example 2.7.6

Example 2.7.7

For the problem of estimating the common variance σ^2 of a set of zero mean i.i.d. Gaussian random variables, the Fisher information was found (see Example 2.7.3) to be $\mathcal{I}(\sigma^2) = n/\sigma^4$. The resulting Jeffreys' prior is then

$$p_{\Sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}; \tag{2.79}$$

at first look, this may seem different from the scale-invariant non-informative prior for σ, obtained in Example 2.6.2, which was $p_\Sigma(\sigma) \propto 1/\sigma$. To shorten the notation, let us write $s = \sigma^2$; direct application of the transformation rule in Eq. (2.42), with $s = g(\sigma) = \sigma^2$, leads to $p_S(s) = 1/s$, in accordance with Eq. (2.79). Once more, the priors based on invariance and obtained by Jeffreys' principle coincide. The resulting *a posteriori* density is

$$p(\sigma^2|\mathbf{x}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right\} \tag{2.80}$$

which is always normalizable but only has mean if $n > 2$ (because otherwise the integral of $\sigma^2\, p(\sigma^2|\mathbf{x})$ from 0 to ∞ does not converge); the PM and MAP estimates are

$$\widehat{\sigma^2}_{\text{MAP}} = \frac{1}{n+2}\sum_{i=1}^{n} x_i^2 \tag{2.81}$$

$$\widehat{\sigma^2}_{\text{PM}} = \frac{1}{n-2}\sum_{i=1}^{n} x_i^2, \quad \text{for } n > 2. \tag{2.82}$$

The MAP estimate obtained under the Jeffreys' prior for this problem has an interesting property, from a frequentist point of view. Consider all (linear) estimators of the form $\widehat{\sigma_k^2} = k\sum_{i=1}^n x_i^2$, and let us take the quadratic loss function $L(\widehat{\sigma_k^2}, \sigma^2) = (\widehat{\sigma_k^2} - \sigma^2)^2$. The frequentist risk (see Section 1.2.2) is found to be, after some computations (involving the fact that the fourth moment of a zero mean Gaussian of variance σ^2 equals $3(\sigma^2)^2$)

$$E_{\mathbf{X}}\left[\left(\sigma^2 - k\sum_{i=1}^n x_i^2\right)^2 \middle| \sigma^2\right] = (\sigma^2)^2\,[k^2(n^2+2n) - 2kn + 1].$$

Computing the derivative with respect to k, and equating to zero, leads to $k = 1/(n+2)$. In conclusion, the MAP criterion under the Jeffrey's prior coincides with the minimum (frequentist) risk estimator. Notice that in this case there is no minimax estimator (see Section 1.2.2), because $[k^2(n^2+2n) - 2kn + 1] > 0$, for any k, and so the supremum of the risk with respect to σ^2 is always equal to infinity.

End of Example 2.7.7

Example 2.7.8

Finally, the Poisson observation model was seen in Example 2.7.4 to have Fisher information $\mathcal{I}(\theta) = T/\theta$. The corresponding Jeffreys' prior is then

$$p_\Theta(\theta) \propto \frac{1}{\sqrt{\theta}} \qquad (2.83)$$

which is different from the one obtained in Example 2.6.3, and which had a scale-invariance interpretation. Notice that θ is not a pure scale parameter because both the mean and the variance of a Poisson random variable are proportional to θ. Nevertheless, this is still a conjugate prior in the Gamma family; notice that, ignoring the normalization factors as in Eq. (2.52),

$$\lim_{\beta\to 0} \mathrm{Ga}(\theta|1/2, \beta) \propto \lim_{\beta\to 0} \theta^{-1/2} e^{-\beta\theta} = \frac{1}{\sqrt{\theta}}. \qquad (2.84)$$

End of Example 2.7.8

The Jeffreys' prior approach is by no means universally accepted in all situations; many times it yields non-informative priors with which many people do not agree and/or that lead to poor parameter estimates; this is specially true when rather than a single scalar parameter, a set of parameters is being estimated [14], [93] (see Section 3.5.3). Other approaches have been proposed and the interested reader is referred to [61] for pointers to the relevant literature. A very comprehensive catalog of non-informative priors is avalaible in [110].

Finally, we would like to stress again that the Fisher information (and consequently the Jeffreys' prior, from a Bayesian perspective) is a fundamental concept. It plays the central role in the development of modern differential geometric theories of statistical inference [1], [62], [79], [94]; it is also a fundamental link in the interplay between statistics and information theory [23], [67], [91].

2.8 Maximum Entropy Priors

In many situations, *a priori* knowledge of a quantitative nature is available which can guide the building of the prior; nevertheless, this knowledge may be insufficient to uniquely specify a prior. A defensible and intuitively appealing criterion to be used in such situations is that the prior should be as non-informative as possible, while keeping compatibility with the *a priori* available information. This idea can be formalized via the *maximum entropy* (ME) criterion proposed by Jaynes (see [56], [57], [58]) which we will briefly review in this section. It is important to mention that, in addition to its intuitive appeal, entropy maximization can be formally shown to be the unique criterion satisfying a set of consistency axioms [100].

Entropy maximization has also been widely used as an estimation criterion *per se* [57], namely for image restoration/reconstruction [20], [44], [50], [112]; that approach (which is not the same as using the ME criterion to obtain priors for Bayesian inference) will not be covered in this section.

2.8.1 Maximum Entropy Priors for Discrete Problems

As its name implies, the *maximum entropy* criterion is supported on the concept of *entropy*; this is arguably the basic building block of *information theory*, the formal body of knowledge devoted to the mathematical study of information; comprehensive references on this subject are [17] and [23].

In a classification problem, the prior is a probability mass function $p_S(s)$, for $s \in \mathcal{S} = \{s_1, s_2, ..., s_M\}$. In this case, the *entropy* of this probability mass function is a quantity defined as

$$\begin{aligned} H\left(S\right) &= \sum_{s_i \in \mathcal{S}} p_S(s_i) \log \left(\frac{1}{p(s_i)} \right) \\ &= E_S\left[\log \left(\frac{1}{p_S(s)} \right) \right] \;=\; E_S\left[-\log\left(p_S(s)\right) \right]. \end{aligned} \tag{2.85}$$

Notice that this definition does not depend on the particular elements of $\mathcal{S}$, but only on their probabilities; the entropy is not a feature of a random variable, but of its probability distribution. This fact would justify the notation $H\left(p_S(s)\right)$, but we settle for the conventional $H(S)$. The entropy is the expected value of the function[5] $-\log(p_S(s))$, which can be interpreted as the information content of the elementary event s, in the following sense: a highly probable event carries little information while a rare event carries a lot of information (as seems to be well known by journalists and other news people). In passing, notice that since $p_S(s) \leq 1$, then $-\log p_S(s) \geq 0$ and so $H(S) \geq 0$ (the expected value of a non-negative function is necessarily non-negative) is true for any probability mass function. This average information interpretation is in accordance with, and is made clearer by, the following facts:

- For some (finite) configuration set $\mathcal{S}$, the probability mass function that maximizes the entropy is uniform $p(s) = 1/|\mathcal{S}|$, where $|\mathcal{S}|$ denotes the number of elements of $\mathcal{S}$. In fact, it is intuitively acceptable that the state of highest ignorance about a random variable occurs when that variable is uniformly distributed. The corresponding maximal entropy is
$$\max_{p_S(s)} \{H\left(S\right)\} = \log|\mathcal{S}|, \tag{2.86}$$
where the maximization is, of course, under the normalization constraint $\sum_i p_S(s_i) = 1$. This result can be obtained in several different ways; the most direct, probably, is simply to use Lagrange multipliers (see Appendix A) to combine the entropy being maximized with the normalization constraint; this leaves us with the minimization of the following function
$$-\sum_{s_i \in \mathcal{S}} p_S(s_i) \log p_S(s_i) + \lambda \sum_{s_i \in \mathcal{S}} p_S(s_i), \tag{2.87}$$
where λ is the Lagrange multiplier. Computing the derivative with respect to $p_S(s_i)$ and equating to zero yields $p_S(s_i) = \exp\{\lambda - 1\}$ (the same for any $s_i \in \mathcal{S}$); finally, invoking the normalization constraint to solve for λ results in $p_S(s_i) = 1/|\mathcal{S}|$, as expected.

- Consider now a probability mass function $p_S(s)$ such that $p_S(s_i) = 1 - \varepsilon$ with $\varepsilon = \sum_{k \neq i} p(s_k)$, where ε is a small ($\ll 1$) positive real number; i.e., we are considering that s_i is almost certain to occur (its probability is almost 1). If ε goes to zero, so must all the $p(s_k)$, for $k \neq i$, since they are non-negative quantities; then, we have that
$$\begin{aligned}\lim_{\varepsilon \to 0} H\left(S\right) &= \lim_{\varepsilon \to 0} \{(\varepsilon - 1) \log\left(1 - \varepsilon\right)\} \\ &\qquad - \sum_{k \neq i} \lim_{p(s_k) \to 0} \{p(s_k) \log\left(p(s_k)\right)\} = 0\end{aligned}$$

[5] The logarithm base only affects the units in which the entropy is measured. The classical choice is base 2, yielding entropies measured in *bits*; alternatively, base e leads to a unit called *nat* [23].

since $x \log x \to 0$, as x tends to zero[6]. This means that as one of the possible events becomes increasingly probable (certain), the entropy (uncertainty) approaches zero.

With the entropy concept in hand, we can now formally state what a *maximum entropy* (discrete) prior is. It is a probability mass function which maximizes the entropy (uncertainty) among all those satisfying the *a priori* available information. Of course, for this criterion to be applicable, the available information has to be formally expressed; the most common and often studied form is a set of $m+1$ equalities

$$\sum_{s_i \in \mathcal{S}} p_S(s_i) g_k(s_i) = \mu_k, \quad \text{for } k = 0, 1, ..., m, \tag{2.88}$$

where $g_0(s) = 1$ and $\mu_0 = 1$ (the zero-th constraint is always present and simply imposes normalization, $\sum p(s_i) = 1$). Under this type of constraint, the maximum entropy (ME) probability mass function has the form

$$p_S^{\text{ME}}(s) = \exp\{\lambda_0 + \sum_{k=1}^{m} \lambda_k g_k(s)\} \quad \text{for } s \in \mathcal{S} \tag{2.89}$$

where the parameters λ_k are obtained so that $p_S^{\text{ME}}(s)$ satisfies the constraints in Eq. (2.88); this result can be obtained by the technique of Lagrange multipliers (see, e.g., [23]).

Example 2.8.1

Let $\mathcal{S} = \{0, 1, 2, 3, ...\}$, the set of non-negative integer numbers, and consider a single ($m = 1$) restriction (apart from the zero-th one), on the expected value, i.e., $g_1(s) = s$, which is supposed to be μ_1. Then, according to Eq. (2.89),

$$p_S^{\text{ME}}(s) = \exp\{\lambda_0 + \lambda_1 s\} = \exp\{\lambda_0\} \left(\exp\{\lambda_1\}\right)^s. \tag{2.90}$$

Invoking the normalization condition, we easily obtain that $\exp\{\lambda_0\} = 1 - \exp\{\lambda_1\}$. Then the maximum entropy probability mass function is a so-called *geometric distribution*,

$$p_S^{\text{ME}}(s) = (1-\theta)\,\theta^s, \tag{2.91}$$

with $\theta = \exp\{\lambda_1\}$, whose expected value is $\theta/(1-\theta)$. The additional restriction (referred above) on the expected value finally leads to $\exp\{\lambda_1\} = \theta = \mu_1/(1+\mu_1)$.

End of Example 2.8.1

[6] This is an indeterminate limit, since $\lim_{x\to 0} x \log x = 0(-\infty)$, which can be easily solved by L'Hôpital's rule leading to $\lim_{x\to 0} x \log x = 0$.

2.8.2 The Kullback-Leibler Divergence

Discrete Variables

When maximizing the discrete entropy, although this was not explicit in Eq. (2.85), one was in fact looking for the prior $p_S(s)$ which is *"most similar"* to a non-informative one $q_S(s)$. The *"dissimilarity"* measure being used is the (discrete version of the) *Kullback-Leibler divergence* (also called *relative entropy*), denoted as $D[p_S(s)\|q_S(s)]$, and given by

$$D[p_S(s)\|q_S(s)] = \sum_{s\in\mathcal{S}} p_S(s) \log\left(\frac{p_S(s)}{q_S(s)}\right). \tag{2.92}$$

Since in discrete problems the non-informative prior is $q_S(s) = 1/|\mathcal{S}|$, the Kullback-Leibler divergence between any candidate prior $p_S(s)$ and this non-informative one is

$$\begin{aligned} D\left[p_S(s)\left\|\frac{1}{|\mathcal{S}|}\right.\right] &= \sum_{s\in\mathcal{S}} p_S(s)\left(\log p_S(s) + \log|\mathcal{S}|\right) \\ &= -H\left(S\right) + \log|\mathcal{S}|; \end{aligned} \tag{2.93}$$

thus, minimizing the Kullback-Leibler divergence with respect to $p_S(s)$ is in fact equivalent to maximizing the entropy $H(S)$

The Kullback-Leibler divergence (notice that $D[p_S(s)\|q_S(s)]$ may be different from $D[q_S(s)\|p_S(s)]$, precluding it from being called a *distance*) is a very important information theoretical concept. It verifies an inequality (known as the *information inequality* or the *Gibbs inequality*, depending on the context in which it is being used) which justifies its adoption as a dissimilarity measure and which has many important consequences and implications; specifically,

$$D[p_S(s)\|q_S(s)] \geq 0 \tag{2.94}$$

$$D[p_S(s)\|q_S(s)] = 0 \Leftrightarrow p_S(s) = q_S(s), \forall_{s\in\mathcal{S}}. \tag{2.95}$$

Notice that it is being assumed that[7] $0\log(0/q) = 0$, while $p\log(p/0) = \infty$; i.e., an event s that has zero probability under $p_S(s)$ has a zero contribution to the total divergence, while any event s with non-zero probability under $p_S(s)$, but zero under $q_S(s)$, makes the divergence go to infinity. The proof of this fundamental inequality is sufficiently simple to be shown here. First, start by noticing that

$$D[p_S(s)\|q_S(s)] = \sum_{s\in\mathcal{S}} p_S(s)\log\left(\frac{p_S(s)}{q_S(s)}\right) = \sum_{s\in\mathcal{S}:p_S(s)\neq 0} p_S(s)\log\left(\frac{p_S(s)}{q_S(s)}\right)$$

[7] This agrees with $\lim_{x\to 0} x\log x = 0$.

because $0 \log(0/q) = 0$. Then, Eq. (2.94) results directly from

$$\begin{aligned} -D[p_S(s) \| q_S(s)] &= \sum_{s \in \mathcal{S}: p_S(s) \neq 0} p_S(s) \log \left(\frac{q_S(s)}{p_S(s)} \right) \\ &\leq \sum_{s \in \mathcal{S}: p_S(s) \neq 0} p_S(s) \left(\frac{q_S(s)}{p_S(s)} - 1 \right) \qquad (2.96) \\ &= \sum_{s \in \mathcal{S}: p_S(s) \neq 0} q_S(s) - \sum_{s \in \mathcal{S}: p_S(s) \neq 0} p_S(s) \\ &\leq 0. \qquad (2.97) \end{aligned}$$

Eq. (2.96) is based on the fact that $\log x \leq (x - 1)$, while the inequality in Eq. (2.97) results from

$$\sum_{s \in \mathcal{S}: p_S(s) \neq 0} q_S(s) \leq 1 \quad \text{and} \quad \sum_{s \in \mathcal{S}: p_S(s) \neq 0} p_S(s) = 1. \qquad (2.98)$$

Strict equality in Eq. (2.96) is only valid if $q_S(s) = p_S(s)$, whenever $p_S(s) \neq 0$, because $\log x = (x - 1)$ if and only if $x = 1$. Strict equality in Eq. (2.97) is true if $q_S(s) = 0$, whenever $p_S(s) = 0$. Consequently, both are true if and only if $p_S(s) = q_S(s)$, for any s.

Example 2.8.2

As an illustration of the power of the information inequality, consider the inequality $H(S) \leq \log |\mathcal{S}|$; in the previous section we showed it by means of Lagrange multipliers. Now, let $q_S(s) = 1/|\mathcal{S}|$ be the uniform distribution over $\mathcal{S}$. The Kullback-Leibler divergence between $p_S(s)$ and this $q_S(s)$ is (see Eq. (2.93))

$$D[p_S(s) \| q_S(s)] = \log |\mathcal{S}| - H(S);$$

then, $\log |\mathcal{S}| \geq H(S)$ is a direct consequence of the *information inequality.* Also the ME distribution in Eq. (2.89) can be obtained by Lagrange multipliers; a much simpler and more elegant derivation supported on the information inequality can be found in [23].

End of Example 2.8.2

Continuous Variables

For a continuous $\mathcal{S}$, the Kullback-Leibler divergence between two probability density functions $p_S(s)$ and $q_S(s)$ is defined as

$$D[p_S(s) \| q_S(s)] = \int_{\mathcal{S}} p_S(s) \log \left(\frac{p_S(s)}{q_S(s)} \right) ds. \qquad (2.99)$$

Let us briefly study the properties of this continuous version of the Kullback-Leibler divergence (also not symmetrical). Its key property is again its non-negativity, i.e., $D[p_S(s) \| q_S(s)] \geq 0$, with equality if and only if $p_S(s) =$

$q_S(s)$ almost everywhere. The proof of this property parallels that of the discrete case, with integrals replacing the summations.

If $q_S(s) = 1/|\mathcal{S}|$ is a uniform density over $\mathcal{S}$, where now $|\mathcal{S}|$ denotes the volume of $\mathcal{S}$, i.e., $|\mathcal{S}| = \int_{\mathcal{S}} ds$ (for now, assumed finite), the Kullback-Leibler divergence reduces to

$$D[p_S(s)\|q_S(s)] = \int_{\mathcal{S}} p_S(s) \log p_S(s)\, ds + \log|\mathcal{S}| = -h\left(S\right) + \log|\mathcal{S}|,$$

where $h\left(S\right)$ is called the *differential entropy* of the r.v. S whose probability density function is $p_S(s)$. Now, as in the discrete case, the *information inequality* implies that $h\left(S\right) \leq \log|\mathcal{S}|$, with equality if and only if $p_S(s) = \log|\mathcal{S}|$ (almost everywhere). Unlike the discrete entropy, however, its differential counterpart is not necessarily non-negative; for example, if $p_S(s) = 1/a$, the uniform density on the interval $[0, a]$, the differential entropy is $h(S) = \log a$, which may well be negative (if $a < 1$).

Example 2.8.3

An example of the use of the Kullback-Leibler divergence, relevant in the context of Bayesian inference, is in the asymptotic analysis of classification problems [14]. Let us consider a classification problem, where the set of possible states of nature is a discrete set $\mathcal{S}$. The class-conditional observation models are $\{f_{\mathbf{X}}(\mathbf{x}|s),\ s \in \mathcal{S}\}$, and let the true state of nature be $s_{\text{true}} \in \mathcal{S}$. If the observations are independent, the class-conditional densities can be factored as

$$f_{\mathbf{X}}(\mathbf{x}|s) = \prod_{i=1}^{n} f_X(x_i|s).$$

The posterior probability function, according to Bayes law, is

$$\begin{aligned} p_S(s|\mathbf{x}) &\propto p_S(s) \prod_{i=1}^{n} f_X(x_i|s) \\ &\propto p_S(s) \prod_{i=1}^{n} f_X(x_i|s) \left(\prod_{i=1}^{n} f_X(x_i|s_{\text{true}}) \right)^{-1} \\ &\propto \exp\left\{ \log p_S(s) - K(s) \right\} \end{aligned} \tag{2.100}$$

(notice that, with respect to s, $\prod_{i=1}^{n} f_X(x_i|s_{\text{true}})$ is a constant), where

$$K(s) = \sum_{i=1}^{n} \log \frac{f_X(x_i|s_{\text{true}})}{f_X(x_i|s)}.$$

Now, conditionally on s_{true}, the terms of the summation defining $K(s)$ are independent and identically distributed random variables, so, by the strong law of large numbers (see Appendix A), as n grows,

$$\lim_{n\to\infty} \frac{K(s)}{n} = \int f_X(x|s_{\text{true}}) \log \frac{f_X(x|s_{\text{true}})}{f_X(x|s)}\, dx = D\left[f_X(x|s_{\text{true}}) \| f_X(x|s)\right].$$

The crucial condition for the result we are seeking is

$$D\left[f_X(x|s_{\text{true}})\|f_X(x|s)\right] > 0, \quad \text{for } \; s \neq s_{\text{true}}; \tag{2.101}$$

if this is true, and since $D\left[f_X(x|s_{\text{true}})\|f_X(x|s_{\text{true}})\right] = 0$, then

$$\lim_{n\to\infty} K(s) = \begin{cases} 0, & \text{if } \; s = s_{\text{true}} \\ +\infty & \text{if } \; s \neq s_{\text{true}} \end{cases}$$

and, consequently,

$$\lim_{n\to\infty} p_S(s|\mathbf{x}) = \begin{cases} 1, & \text{if } \; s = s_{\text{true}} \\ 0 & \text{if } \; s \neq s_{\text{true}} \end{cases}$$

showing that the probability of any wrong decision goes to zero as the amount of observed data increases.

The condition in Eq. (2.101) can be seen as an indentifiability condition; of course, if one (non true) class has a conditional observation model which is indistinguishable with respect to the true class, it can not be guaranteed that the *a posteriori* probability function will asymptotically concentrate on the true class.

End of Example 2.8.3

2.8.3 Maximum Entropy Priors for Estimation Problems

Obtaining maximum entropy priors when the configuration set $\mathcal{S}$ is continuous is not as simple as in the discrete case because there is no such thing as the uncontroversial non-informative prior serving as global reference. To leave that issue aside, for now, let us consider that we have agreed on some non-informative prior $q_S(s)$. The available information is of the same type as expressed in Eq. (2.88), with the summations necessarily replaced by integrals,

$$\int_{\mathcal{S}} p_S(s) g_k(s)\, ds = \mu_k, \quad \text{for } \; k = 0, 1, ..., m. \tag{2.102}$$

The *maximum entropy* (ME) prior (or better, the *least informative prior*) becomes

$$p_S^{\text{ME}}(s) = q_S(s) \exp\{\lambda_0 + \sum_{k=1}^{m} \lambda_k g_k\,(s)\} \quad \text{for } \; s \in \mathcal{S}, \tag{2.103}$$

where the parameters λ_k are again obtained from the constraints; this result can (as in the discrete case) be derived by Lagrange multipliers or via the information inequality [23]. If there are no explicit constraints apart from normalization, i.e., if $m = 0$, then it is obvious that the ME prior coincides with the adopted non-informative density $q_S(s)$. The important conclusion is that, in the continuous case, everything becomes specified up to a reference prior $q_S(s)$ which has to be chosen *a priori*.

A technical issue may arise here if the non-informative reference prior is improper; for example, let $\mathcal{S} = I\!R$, and $q_S(s) = k$, where k is an arbitrary constant (since this prior is not normalizable, anyway). The Kullback-Leibler divergence becomes $D[p_S(s)\|k] = -h\,(S) + \log k$, which is only defined up to an arbitrary additive constant. Nevertheless, one can still look for the density that minimizes $D[p_S(s)\|k]$, as long as k is kept constant; the resulting *least informative prior* is really, in this case, a maximum (differential) entropy one.

Example 2.8.4

Let s be a location parameter (thus with uniform non-informative prior, say equal to 1) which is known to be non-negative, i.e. $\mathcal{S} = [0, +\infty)$; moreover, it is *a priori* known to have mean equal to μ_1 (since we have a constraint on the mean, $g_1(s) = s$). The ME prior is, according to Eq. (2.103), $p_S^{\mathrm{ME}}(s) = \exp\{\lambda_0 + \lambda_1 s\}$. The normalization constraint leads to $\exp\{\lambda_0\} = -\lambda_1$ (of course, as long as $\lambda_1 < 0$), while the constraint on the mean leads to $\lambda_1 = -1/\mu_1$; putting these two results together yields

$$p_S^{\mathrm{ME}}(s) = \frac{1}{\mu_1}\exp\left\{-\frac{s}{\mu_1}\right\}, \tag{2.104}$$

an exponential density.

End of Example 2.8.4

Example 2.8.5

Suppose that $s \in \mathcal{S} = I\!R$ is real valued location parameter (thus with uniform non-informative prior, say equal to 1). Consider again that the mean has to be equal to μ_1 (then $g_1(s) = s$); additionally, it is also now required that the variance be equal to σ^2. This means that we now also have $g_2(s) = (s - \mu_1)^2$ and $\mu_2 = \sigma^2$. The resulting ME prior, according to Eq. (2.103), is

$$p_{\mathrm{ME}}(s) = \exp\{\lambda_0 + \lambda_1 s \ + \lambda_2 (s - \mu_1)^2\}. \tag{2.105}$$

Invoking the constraints, some simple manipulations lead to $p_{\mathrm{ME}}(s) = \mathcal{N}(s|\mu_1, \mu_2)$; i.e., the least informative prior for a location parameter, with given mean and variance, is Gaussian; this is one of the several frequently invoked arguments for the ubiquity of Gaussian densities. The entropy of the resulting Gaussian density,

$$h(S) = \frac{1}{2}\log(2\pi e \sigma^2), \tag{2.106}$$

is, naturally, only a function of its variance (not of its mean).

End of Example 2.8.5

2.9 Mutual Information, Data Processing Inequality, and Jeffreys' Priors

2.9.1 Conditional Entropy and Mutual Information

The information theoretical concepts introduced in the previous section allow an alternative perspective on some aspects of Bayesian inference that we believe is enlightening. Let us start by defining two quantities which will be needed in the sequel: *conditional entropy* and *mutual information.*

Given two continuous random variables S and U, with joint probability density function $p_{S,U}(s,u)$, the conditional entropy of S, given U, is defined as

$$\begin{aligned} H(S|U) &= \int_{\mathcal{U}} H[S|u] p_U(u)\, du \\ &= -\int_{\mathcal{U}} \int_{\mathcal{S}} \left[p_S(s|u) \log p_S(s|u) \right]\, ds\, p_U(u)\, du \\ &= -\int_{\mathcal{U}} \int_{\mathcal{S}} p_{S,U}(s,u) \log p_S(s|u)\, ds\, du \end{aligned} \qquad (2.107)$$

and can be interpreted as the average uncertainty about S in the presence of U. Of course, a similar definition exists for discrete random variables with the integrations adequately replaced by summations. Letting $H(S,U)$ denote the joint entropy of S and U, which is simply the entropy associated with their joint probability density function, we can write an information theoretical equivalent to Bayes theorem,

$$H(S,U) = H(S|U) + H(U) = H(U|S) + H(S). \qquad (2.108)$$

The concept of *mutual information* between two random variables, say S and U, denoted $I[S;U]$, has a simple definition supported on the notion of conditional entropy:

$$I[S;U] = H(S) - H(S|U) = H(U) - H(U|S). \qquad (2.109)$$

As is clear from its definition, the mutual information between S and U can be thought of as the amount of information about S carried by U (or vice-versa). Some simple manipulation allows rewriting the mutual information as

$$\begin{aligned} I[S;U] &= \int_{\mathcal{U}} \int_{\mathcal{S}} p_{S,U}(s,u) \log \frac{p_{S,U}(s,u)}{p_U(u) p_S(s)}\, ds\, du \\ &= D\left[p_{S,U}(s,u) \| p_U(u) p_S(s) \right] \end{aligned} \qquad (2.110)$$

that is, it coincides with the Kullback-Leibler divergence between the joint density $p_{S,U}(s,u)$ and the product of the corresponding marginals. This fact also has a clear interpretation: the mutual information measures how far the two random variables involved are from being independent (recall that independence between S and U would be expressed by $p_{S,U}(s,u) = p_U(u)p_S(s)$). An immediate consequence of the information inequality (Eqs. (2.94) and (2.95)) is then $I[S;U] \geq 0$, with strict equality if only if U and S are independent. Notice that this, in turn, implies that $H(S) \geq H(S|U)$ and $H(U) \geq H(U|S)$ which can be read as "conditioning can not increase uncertainty".

Finally, before going into the main results we wish to present in this section, we need to introduce one more concept: that of *conditional mutual information*. This is a simple modification of above definition: the mutual information between S and U, given a third variable V, is

$$I[S;U|V] = H(S|V) - H(S|U,V). \tag{2.111}$$

Of course, the conditional mutual information also verifies the information inequality, *i.e.*, $I[S;U|V] \geq 0$, with strict equality if and only if S and U are conditionally independent, given V (that is, if $p_{S,U}(s,u|v) = p_S(s|v)p_U(u|v)$). The conditional mutual information also allows a decomposition that parallels the one for joint entropy expressed in Eq. (2.108). Specifically, the mutual information between a random variable S and a pair of random variables U and V, denoted $I[S;U,V]$, can be decomposed (invoking Eqs. (2.109) and (2.108)) in two ways

$$\begin{aligned} I[S;U,V] &= H(U,V) - H(U,V|S) \\ &= H(U|V) + H(V) - H(U|V,S) - H(V|S) \\ &= I[S;V] + I[S;U|V] \end{aligned} \tag{2.112}$$

and

$$\begin{aligned} I[S;U,V] &= H(U,V) - H(U,V|S) \\ &= H(V|U) + H(U) - H(V|U,S) - H(U|S) \\ &= I[S;U] + I[S;V|U]. \end{aligned} \tag{2.113}$$

2.9.2 The Data Processing Inequality

The data processing inequality formally captures a fact that can almost be considered as common sense: no processing can increase the amount of information contained in observed data. To be specific, let us assume that s is an unknown real quantity, with a prior $p_S(s)$, that is to be estimated from observations $\mathbf{x}$ obtained according to the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$. Now, in the presence of $\mathbf{x}$, we compute some quantity that is a (deterministic or random) function of $\mathbf{x}$; let the result of this *data processing* be denoted as $\mathbf{z}$. Of course, since $\mathbf{z}$ is exclusively a function of $\mathbf{x}$, not of s (which is unknown to this data processor), we can write

$$p_{\mathbf{Z}}(\mathbf{z}|\mathbf{x}, s) = p_{\mathbf{Z}}(\mathbf{z}|\mathbf{x}).$$

An immediate consequence of this is that S and $\mathbf{Z}$ are conditionally independent, given $\mathbf{X}$:

$$p_{S,\mathbf{Z}}(s,\mathbf{z}|\mathbf{x}) = \frac{p_{\mathbf{Z}}(\mathbf{z}|\mathbf{x})p_{S,\mathbf{X}}(s,\mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})} = p_{\mathbf{Z}}(\mathbf{z}|\mathbf{x})p_S(s|\mathbf{x}).$$

According to the results in the last section, the conditional mutual information between S and $\mathbf{Z}$, given $\mathbf{X}$, is then zero, $I[S,\mathbf{Z}|\mathbf{X}] = 0$. The two different decompositions of $I[S;\mathbf{X},\mathbf{Y}]$ expressed in Eqs. (2.112) and (2.113) allow writing

$$I[S;\mathbf{X}] + \underbrace{I[S;\mathbf{Z}|\mathbf{X}]}_{=0} = I[S;\mathbf{Z}] + \underbrace{I[S;\mathbf{X}|\mathbf{Z}]}_{\geq 0}, \tag{2.114}$$

leading to the *data processing inequality*

$$I[S;\mathbf{Z}] \leq I[S;\mathbf{X}]; \tag{2.115}$$

no matter how you process your observations, the result will have no more information about S than the observations themselves.

2.9.3 An Information Theoretic View of Jeffreys' Priors

Let us consider an unknown quantity s, about which prior information is expressed by $p_S(s)$, and data $\mathbf{x}$ obtained according to an observation model $f_{\mathbf{X}}(\mathbf{x}|s)$. Suppose we interpret s as a message that is transmitted through some communication channel whose output is $\mathbf{x}$, thus characterized by the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$.

A natural way to measure how much information about s is carried by the observations is the mutual information $I[S;\mathbf{X}]$. However, as is clear from its definition, $I[S;\mathbf{X}]$ not only depends on the likelihood $f_{\mathbf{X}}(\mathbf{x}|s)$, but also on the prior $p_S(s)$. The fundamental concept of *channel capacity* [23] is then defined as

$$C\left(f_{\mathbf{X}}(\mathbf{x}|s)\right) = \sup_{p_S(s)} I[S;\mathbf{X}],$$

the maximal attainable mutual information. Interestingly, the prior that achieves the channel capacity is, asymptotically, Jeffrey's prior [91]. In other words, we can say the Jeffreys' prior is the one that best adjusts itself to the observation model (channel). This is the type of path followed by Bernardo [14] to obtain the so-called *reference priors.* In the scalar case, and under regularity conditions, reference priors coincide with Jeffreys' priors. In the discrete case, reference priors are simply maximum entropy priors. For other information-theoretical aspects of Jeffreys' priors and Bayesian inference in general, see, *.e.g.*, [5] and references therein.

2.10 Minimum Description Length Priors

The *minimum description length* (MDL) principle is an information theoretical approach introduced by Rissanen (see [5], [88], [90], and also [28]). While MDL is not rooted on Bayesian considerations, it may be seen as a way to (objectively) obtain priors.

To present the MDL principle, let us consider a typical communication scenario. Imagine a *sender* whose role is to transmit some observed data $\mathbf{x}$ to a *receiver*, by using a channel that supports error-free transmission of binary digits. Both *sender* and *receiver* are aware of the fact that $\mathbf{x}$ is a sample of a random (possibly vector) variable characterized by a *parametric statistical model* (or likelihood function) $f_{\mathbf{X}}(\mathbf{x}|s)$; moreover, both are temporarily assumed to know the true s.

To use the binary channel, the sender has to encode the observations $\mathbf{x}$ into a sequence of binary digits. He does so by designing a *coder*, denoted $\mathcal{C}$, which is a (necessarily injective) mapping from $\mathcal{X}$ into the set of all finite binary strings (usually denoted as $\{0,1\}^*$), i.e. $\mathcal{C}:\mathcal{X}\to\{0,1\}^*$. The code-word for observation $\mathbf{x}$ is then denoted $\mathcal{C}(\mathbf{x})$.

Naturally, the goal of the code designer is that the binary messages produced are (on average) as short as possible, while being, of course, decodable by the receiver. However, in order for the *sender* to be able to use the same coder to transmit, not just one observation $\mathbf{x}$, but a sequence of observations $(\mathbf{x}_1,\mathbf{x}_2,\ldots)$, the code has to have the *prefix* property; this means that no codeword can be the prefix (initial segment) of another one. Such a code is also called *instantaneous*, because as soon as one particular code word has been fully received, the receiver recognizes it immediately without the need to look at any subsequent bits (see, e.g., [23]). Letting $L(\mathbf{x}) = l(\mathcal{C}(\mathbf{x}))$ denote the length (in binary digits, or bits) of the codeword associated with observation $\mathbf{x}$, it is well known that any code with the *prefix* property has to verify

$$\sum_{\mathbf{x}\in\mathcal{X}} 2^{-L(\mathbf{x})} \leq 1, \tag{2.116}$$

the Kraft-McMillan inequality [23]. Two observations are in order here. First, there is nothing particular about binary digits or binary channels, and Eq. (2.116) would be valid for code words on any other finite alphabet if the powers of 2 are replaced by the powers of the alphabet size, say m; moreover, all the base-2 logarithms to be used below, would have to be replaced by base-m ones. Secondly, if the observation space $\mathcal{X}$ is continuous, it has to be somehow discretized so that the set of possible messages becomes countable; it is easy to show that the effect of this discretization is (asymptotically) irrelevant to all the issues to be addressed in the sequel [91].

The requirement of having the shortest possible code length has to be imposed on its expected value, since the *sender* and the *receiver* do not know *a priori* which observation will have to be encoded (this brings a frequentist flavor to the obtained criterion). If the average code length

$$E_{\mathbf{X}}[L(\mathbf{x})|s] = \sum_{\mathbf{x}\in\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}|s)L(\mathbf{x}) \tag{2.117}$$

is minimized with respect to all possible coders which satisfy the Kraft-McMillan constraint (Eq. (2.116)), the result is as follows: optimality is achieved by any coder whose length function $L(\mathbf{x}) = l(\mathcal{C}(\mathbf{x}))$ verifies

$$L(\mathbf{x}) = L(\mathbf{x}|s) = -\log_2 f_{\mathbf{X}}(\mathbf{x}|s), \tag{2.118}$$

which are known as the Shannon code-lengths. In Eq. (2.118), the notation $L(\mathbf{x}|s)$ was introduced to stress the fact that this length function is under the assumption of a certain s. Of course, $-\log_2 f_{\mathbf{X}}(\mathbf{x}|s)$ may not be an integer and so this length function may not be achievable by a practical[8] code; this is a technical detail which will not be considered here because it is (asymptotically) irrelevant. Notice that these lengths (if we ignore the integer length requirement) clearly satisfy the Kraft-McMillan condition with equality. The average length achieved by this optimal code is

$$E_{\mathbf{X}}[L(\mathbf{x})|s] = -\sum_{\mathbf{x}\in\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}|s)\log_2\left(f(\mathbf{x}|s)\right) = H\left(f_{\mathbf{X}}(\mathbf{x}|s)\right), \tag{2.119}$$

i.e., precisely the entropy of $f_{\mathbf{X}}(\mathbf{x}|s)$ (recall Eq. (2.85)). It can also be shown that any instantaneous code (i.e., one which verifies Eq. (2.116)) has average length greater than or equal to the entropy [23]. It is worth pointing out that the optimal code-lengths in Eq. (2.118) can be easily derived by the *information inequality* [23].

To return to our inference problems, let us admit now that s is unknown; when some $\mathbf{x}$ is observed, the first step that has to be taken is to estimate s from this observation. If nothing else is known *a priori*, the natural coding-oriented criterion is to find that s that leads to a code in which that particular observation has the shortest possible code length. Or putting it in a different way, suppose that all possible coders (for all possible values of s) were previously designed and are available to the sender; in the presence of $\mathbf{x}$ which is to be communicated to the receiver via the channel, the obvious coder is the one yielding the shortest code-length for that $\mathbf{x}$. Formally, the best choice is then

$$\begin{aligned}\widehat{s} &= \arg\min_s L(\mathbf{x}|s) \;=\; \arg\min_s \{-\log_2 f_{\mathbf{X}}(\mathbf{x}|s)\}\\ &= \arg\max_s \{f_{\mathbf{X}}(\mathbf{x}|s)\} \;\equiv\; \delta_{\text{ML}}(\mathbf{x})\end{aligned} \tag{2.120}$$

[8] There are practical techniques to design codes whose expected length is close to the optimal, the best known being, by far, the Huffman procedure [23].

which coincides with the *maximum likelihood* criterion (see Eq. (2.5)). Once this estimate $\widehat{s}$ is obtained, the corresponding coder $\mathcal{C}(\cdot|\widehat{s})$ can be built, and $\mathbf{x}$ can be encoded into an $L(\mathbf{x}|\widehat{s})$ bits long code word, $\mathcal{C}(\mathbf{x}|s)$, which is then sent to the receiver. However, and this is the key observation behind the MDL principle, for the receiver to be able to decode $\mathcal{C}(\mathbf{x}|\widehat{s})$, it needs to know $\widehat{s}$ so that it can use the corresponding decoder. This requirement leads to the concept of a *two-part* code [89]: a *preamble*, carrying a particular s (whose length is denoted $L(s)$) based upon which the observation $\mathbf{x}$ was coded; and a *body*, which consists of $\mathcal{C}(\mathbf{x}|s)$. Notice that this scheme works regardless of the value of s; the only constraint is that the same s is used to define the coder and the decoder. The total length of this two-part code is then

$$L(\mathbf{x}, s) = L(\mathbf{x}|s) + L(s); \tag{2.121}$$

now, the natural (coding-oriented) goal is to minimize the (total) code length, and the resulting optimal value of s is

$$\begin{aligned} \widehat{s} &= \arg\min_{s} \{L(\mathbf{x}|s) + L(s)\} \\ &= \arg\min_{s} \{-\log_2 f(\mathbf{x}|s) + L(s)\} \equiv \delta_{\mathrm{MDL}}(\mathbf{x}) \end{aligned} \tag{2.122}$$

called the *minimum description length* (MDL) estimate. This is formally equivalent to a MAP criterion corresponding to the prior $p_S(s) \propto 2^{-L(s)}$ which is then called an *MDL prior.*

The MDL criterion may be seen as involving an alternative interpretation of prior probabilities; rather than *subjective* (or *personal*) *degrees of belief*, we now have what can be called a *coding approach to probability* [28]. In essence, it may be formalized as follows: given a sample space $\mathcal{X}$ and a coder $\mathcal{C} : \mathcal{X} \to \{0,1\}^*$ whose length function $L(\mathbf{x})$ meets the Kraft-McMillan inequality (i.e., it generates a prefix code), this code defines a probability function over $\mathcal{X}$ given by

$$f_{\mathbf{X}}^{\mathcal{C}}(\mathbf{x}) = \frac{2^{-L(\mathbf{x})}}{\sum_{\mathbf{x}\in\mathcal{X}} 2^{-L(\mathbf{x})}}. \tag{2.123}$$

In this framework, code lengths rather than probabilities are the fundamental concepts. Another implication of this approach is regarding the choice of the loss function; when dealing with code lengths, there is only one natural optimality criterion: minimize code lengths. In conclusion, MDL can be seen as an alternative (though a closely related one) to the Bayesian approach: it still needs a *parametric probabilistic model* $f_{\mathbf{X}}(\mathbf{x}|s)$ for the observations, although it reinterprets it in terms of code lengths; it requires a coding scheme for the unknown s which plays the role of prior; it adopts a unique optimality criterion, which is to minimize the total code length.

Example 2.10.1

Suppose $\mathcal{S} = \{0, 1, 2, ...\}$ is the set of positive integers. An optimal way of encoding any s, devised by Elias [40], proceeds as follows: Take some integer s; its binary expansion is of order $\lceil \log_2(s+1) \rceil$, where $\lceil x \rceil$ denotes "the smallest integer not smaller than x"; let us use the shorter notation $\lambda(x) \equiv \lceil \log_2(x+1) \rceil$. For example, $s = 1998$ is written as the 11 bit binary sequence 11111001110, and $\lambda(s) = 11$. To detect this sequence, possibly immersed in a longer one, the receiver must be told in advance how many bits to expect; this can be done by including as prefix the length information 1011 (eleven, in binary), yielding 101111111001110 which is $15 = \lambda(s) + \lambda(\lambda(s))$ bits long. But now the same problem reemerges: how to distinguish the two parts of the codeword? Let us include yet another preamble stating that the next 4 bits encode the length of the following bit sequence; this new preamble is 100 (4, in binary) which has length $3 = \lambda(4) = \lambda(\lambda(11)) = \lambda(\lambda(\lambda(1998)))$. Of course, the procedure is reiterated until we have only 2 bits; which for $s = 1998$ happens in the next step; i.e., 3 in binary notation is 11. Noticing that all the binary numbers referred above (101111111001110, 1011, 100, 11) start with a one, provides a way of signaling when a sequence is the final one and not the length of the next: just append a zero to the end of the message. The final complete codeword would be $11, 100, 1011, 11111001110, 0$ (of course, without the commas). The total number of bits required to encode an arbitrary integer s is then $1 + \lambda^*(s) = \lambda(s) + \lambda(\lambda(s)) + \cdots + \lambda(\lambda(\cdots \lambda(s)))$, where the recursion is only carried out until the the result first equals 2. Noting that $\lambda(s) \simeq \log_2 s$, which in a relative sense is asymptotically true, leads to $L(s) \simeq \log_2^* s$ where $log_2^* s = \log_2 s + log2 \log_2 s + \cdots + \log_2 \log_2 \cdots \log_2 s$, where only positive terms are kept. Table 2.10.1 shows some examples of codes for integers obtained by this procedure and the corresponding $1 + \lambda^*$ and $\log_2^*$ functions. Since this is clearly a prefix code, it satisfies the Kraft-McMillan inequality and Rissanen used it to define a probability mass function over $\mathcal{S}$,

$$p_S(s) \propto 2^{-\log_2^* s} \tag{2.124}$$

which he called the *universal prior for integers* [88]. It is a proper probability mass function, although it has infinite entropy.

End of Example 2.10.1

When $\mathcal{S}$ is a continuous set, designing finite codes requires $\mathcal{S}$ to be discretized; let $s' \in \mathcal{S}'$ denote the discretized values. A delicate trade-off is at play: if the discretization is very fine, $L(s')$ will be large while $-\log_2 f_{\mathbf{X}}(\mathbf{x}|s')$ will be small because s' will be close to the optimal s; a coarse discretization will imply a smaller $L(s')$ but a larger $-\log_2 f(\mathbf{x}|s')$ since s' can be far from the optimal value of s. The best compromise is not easy to obtain in a closed form; a now well known expression, valid when s is a real parameter and asymptotically in n (the dimension of $\mathbf{x}$) is (in

s	$\mathcal{C}(s)$	$1+\lambda_2^*(s)$	$\log_2^* s$
0	0	1	$-\infty$
1	1.0	2	0
2	10.0	3	1
4	11.100.0	6	3
500	11.100.1001.111110100.0	19	14.5
1998	11.100.1011.11111001110.0	20	16.21
2^{16}	11.101.10001.1000000000000000.0	28	23
10^6	11.101.10100.11110100001001000000.0	46	27.54
2^{128}	11.100.1000.10000001.1(128 zeros).0	147	$\simeq 142$
$2^{2^{16}}$	11.101.10001.1(15 zeros)1.1(2^{16} zeros).0	$\simeq 2^{16}$	$\simeq 2^{16}$

TABLE 2.1. Some integers, their codes, true code lengths, and approximate code lengths given by the $\log_2^*$ function.

bits)

$$L(s) = \frac{1}{2}\log_2 n \tag{2.125}$$

(more accurate and recent results can be found in [5] and [91]). This being a constant with respect to s, we can conclude that if s is an unknown real quantity, the MDL and ML criteria will coincide. This seems to be a disappointing result, and in fact it is; the MDL criterion will only reveal all its usefulness in situations where s is a vector of unknown dimension.

2.11 Sufficient Statistics

There is clearly a common feature to all the examples considered above: inference procedures are based on a function of the observed data, rather than on the data itself. Let us denote this function as $t(\mathbf{x})$.

In Examples 1.5.2 and 2.3.1, where the problem was that of estimating the mean of a Gaussian from a set of independent observations $\{x_1, x_2, ..., x_n\}$, the relevant function was $t(\mathbf{x}) = x_1+x_2+...+x_n$. When the samples are not independent, recall Example 1.5.3, the function of the data on which the estimates depend is $t(\mathbf{x}) = \mathbf{x}\mathbf{C}^{-1}\mathbf{u}$. To estimate the probability of success (heads outcomes, when tossing coins as in Example 2.3.2) from a set of n Bernoulli trials, all that matters is $t(\mathbf{x}) = n_h(\mathbf{x}) =$ "number of successes". Estimates of the variance of a Gaussian density, from a set of n independent samples, depend on $t(\mathbf{x}) = x_1^2 + x_2^2 + ... + x_n^2$. In the Poisson observation model, the appropriate function is again $t(\mathbf{x}) = x_1 + x_2 + ... + x_n$. Finally, even in the uniform distribution case from Example 2.5.1, any inference procedure depends on a simple function, $t(\mathbf{x}) = max\{x_1, x_2, ..., x_n\}$.

The general concept behind all these examples is that of *sufficient statistic*, introduced by Fisher [42]. Formally, let $\mathbf{t}(\mathbf{x})$ be a (possibly vector) function of the observations; such a function is called a *statistic.* If $\mathbf{t}(\mathbf{x})$ is such that

$$f_{\mathbf{X}}(\mathbf{x}|\mathbf{t}(\mathbf{x}), s) = f_{\mathbf{X}}(\mathbf{x}|\mathbf{t}(\mathbf{x})), \tag{2.126}$$

i.e., if the likelihood function, when conditioned on some particular value of the statistic, does not depend on s, then $\mathbf{t}(\mathbf{x})$ is called a *sufficient statistic.*

The key implication of this definition is that $\mathbf{t}(\mathbf{x})$ preserves all the information carried by observed data $\mathbf{x}$, with respect to the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$. To see why this is true, imagine some particular observation $\mathbf{x}$; after computing $\mathbf{t}(\mathbf{x})$, we obtain another sample $\mathbf{x}'$ but now under the constraint that $\mathbf{t}(\mathbf{x}') = \mathbf{t}(\mathbf{x})$. According to Eq. (2.126),

$$f_{\mathbf{X}}(\mathbf{x}'|\mathbf{t}(\mathbf{x}') = \mathbf{t}(\mathbf{x}), s) = f_{\mathbf{X}}(\mathbf{x}'|\mathbf{t}(\mathbf{x}') = \mathbf{t}(\mathbf{x})) \tag{2.127}$$

which means that this new sample is independent of s and so carries no further information about it. This fact is formalized by the *sufficiency principle.*

2.11.1 The Sufficiency Principle

The sufficiency principle states that if $\mathbf{t}(\mathbf{x})$ is a sufficient statistic with respect to the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are two observations such that $\mathbf{t}(\mathbf{x}_1) = \mathbf{t}(\mathbf{x}_2)$, then both the observations must yield the same decision.

It is usually difficult to directly obtain $f_{\mathbf{X}}(\mathbf{x}|\mathbf{t}(\mathbf{x}))$. In many situations, such as the ones referred above, it is easy to identify sufficient statistics by a simple inspection of the likelihood function. The common alternative is to rely on the following factorization-based criterion: if we can write the likelihood function as

$$f_{\mathbf{X}}(\mathbf{x}|s) = \phi(\mathbf{x})\, g_{\mathbf{T}}(\mathbf{t}(\mathbf{x})|s) \tag{2.128}$$

where $\phi(\mathbf{x})$ is a function that does not depend on s, and $g_{\mathbf{T}}(\mathbf{t}(\mathbf{x})|s)$ is a probability (density or mass) function, then $\mathbf{t}(\mathbf{x})$ is a sufficient statistic. The reciprocal is also true; that is, if $\mathbf{t}(\mathbf{x})$ is a sufficient statistic, then the original likelihood can be factored as in Eq. (2.128). There are, of course, many trivial sufficient statistics; e.g., the complete data itself, $\mathbf{t}(\mathbf{x}) = \mathbf{x}$, is obviously a sufficient statistic. Useful sufficient statistics have smaller dimension than the data (as is the case in the examples mentioned in the second paragraph of this section); this requirement is captured by the concept of *minimal sufficient statistic*: let $f_{\mathbf{X}}(\mathbf{x}|s)$ be a likelihood function; a *minimal sufficient statistic* is any sufficient statistic $\mathbf{t}(\mathbf{x}) \in I\!R^n$ such that there is no $m < n$ for which it is possible to find a sufficient statistic $\mathbf{t}'(\mathbf{x}) \in I\!R^m$. Furthermore, notice that any invertible function of a sufficient statistic is also a sufficient statistic.

We conclude this introduction to the concept of sufficiency with a simple example. The other examples referred at the beginning of this section (Bernoulli, Gaussian, Poisson) will be considered in the next section, which will be devoted to the so-called *exponential families*.

Example 2.11.1

Consider the observation model of Example 2.5.1, expressed in Eq. (2.30). Notice that it can be rewritten as

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \begin{cases} \dfrac{1}{\theta^n}, & \max\{x_1, x_2, ..., x_n\} \leq \theta \\ 0, & \text{otherwise} \end{cases} \tag{2.129}$$

revealing that $t(\mathbf{x}) = \max\{x_1, x_2, ..., x_n\}$ is in fact a sufficient statistic. Concerning the factorization in Eq. (2.128), we simply have $\phi(\mathbf{x}) = 1$.

End of Example 2.11.1

2.11.2 An Information Theoretic Perspective

The *data processing inequality* (see Section 2.9.2) allows looking at sufficient statistics from an information theoretic point of view. Like in Section 2.9.2, let us assume that s, characterized by a prior $p_S(s)$, is to be estimated from observations $\mathbf{x}$ obtained according to the likelihood function $f_{\mathbf{X}}(\mathbf{x}|s)$. Now, in the presence of each $\mathbf{x}$, we compute a statistic $\mathbf{t}(\mathbf{x})$; this, of course, defines a new random variable $\mathbf{T} = \mathbf{t}(\mathbf{X})$ which is a function of $\mathbf{X}$. Let us recall Eq. (2.114) that leads to the data processing inequality:

$$I[S; \mathbf{X}] + \underbrace{I[S; \mathbf{T}|\mathbf{X}]}_{=0} = I[S; \mathbf{T}] + I[S; \mathbf{X}|\mathbf{T}]; \tag{2.130}$$

if all we can state is that $I[S; \mathbf{X}|\mathbf{Z}] \geq 0$, we obtain the data processing inequality. When $\mathbf{T} = \mathbf{t}(\mathbf{X})$ is a sufficient statistic, then

$$I[S; \mathbf{X}|\mathbf{T}] = H(\mathbf{X}|\mathbf{T}) - H(\mathbf{X}|S, \mathbf{X}) = 0 \tag{2.131}$$

because of the equality expressed by Eq. (2.126). Consequently, the mutual informations between S and $\mathbf{T}$, and S and $\mathbf{X}$, satisfy

$$I[S; \mathbf{T}] = I[S; \mathbf{X}], \tag{2.132}$$

meaning that the sufficient statistic $\mathbf{T}$ and the observed data $\mathbf{X}$ contain the same amount of information about the unknown S.

2.12 Exponential Families

2.12.1 Fundamental Concepts

Some probability (density or mass) functions allow a factorization of the type expressed in Eq. (2.128) of a more particular kind

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \phi(\mathbf{x})\,\psi(\theta)\,\exp\{\boldsymbol{\xi}(\theta)^T\mathbf{t}(\mathbf{x})\} \\ &= \phi(\mathbf{x})\,\psi(\theta)\,\exp\left\{\sum_{j=1}^{k}\xi_j(\theta)t_j(\mathbf{x})\right\}, \end{aligned} \tag{2.133}$$

where $\mathbf{t}(\mathbf{x})$ is a k-dimensional statistic and $\boldsymbol{\xi}(\theta)$ is a (also k-dimensional) function of the parameter θ. A family of probability functions which can be written under this form is called an *exponential family of dimension-k* [19], [93]. Usually, when θ is a single parameter (a scalar), both $\boldsymbol{\xi}(\theta)$ and $\mathbf{t}(\mathbf{x})$ are scalar (i.e., $k = 1$); however, there are situations where there is only one unknown parameter, but still $k > 1$ (see Example 2.12.5 below). Notice that $\psi(\theta)$ must guarantee that each member of the family is normalized to one (see Eq. (2.139), below). The fact that Eq. (2.133) is a particular case of Eq. (2.128) (with $g_{\mathbf{T}}(\mathbf{t}(\mathbf{x})|\theta) = \psi(\theta)\,\exp\{\boldsymbol{\xi}(\theta)^T\mathbf{t}(\mathbf{x})\}$) immediately shows that $\mathbf{t}(\mathbf{x})$ is in fact a sufficient statistic. The $\boldsymbol{\xi}(\theta)$ and $\mathbf{t}(\mathbf{x})$ that allow writing the likelihood function under the exponential family form are called the *natural* (or *canonical*) parameter and sufficient statistic, respectively. It is always possible to use the change of variables $\mathbf{t} = \mathbf{t}(\mathbf{x})$ and the reparameterization $\boldsymbol{\xi} = \boldsymbol{\xi}(\theta)$ to cast an exponential family into its so-called *natural form*

$$f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\xi}) = \phi(\mathbf{t})\,\psi(\boldsymbol{\xi})\,\exp\{\boldsymbol{\xi}^T\mathbf{t}\}, \tag{2.134}$$

where these $\phi(\cdot)$ and $\psi(\cdot)$ functions may be different from the ones in (2.133). Finally, it is important to be aware that many of the most common probabilistic models, continuous and discrete, do belong to exponential families, which makes their manipulation particularly convenient; e.g., Gaussian, Poisson, Bernoulli, binomial, exponential, gamma, and beta.

Consider a set of observations $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ that are independent and identically distributed according to an exponential family (see Eq. (2.133)); then, the resulting joint likelihood belongs to an exponential family with the same natural parameter. In fact,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{i=1}^{n}\phi(\mathbf{x}_i)\,\psi(\theta)\,\exp\{\boldsymbol{\xi}(\theta)^T\mathbf{t}(\mathbf{x}_i)\} \\ &= \psi(\theta)^n\left(\prod_{i=1}^{n}\phi(\mathbf{x}_i)\right)\exp\{\boldsymbol{\xi}(\theta)^T\mathbf{t}(\mathbf{x})\}, \end{aligned} \tag{2.135}$$

automatically defining the (joint) natural sufficient statistic as the summation of the individual ones,

$$\mathbf{t}(\mathbf{x}) = \mathbf{t}(\mathbf{x}_1) + \mathbf{t}(\mathbf{x}_2) + ... + \mathbf{t}(\mathbf{x}_n), \tag{2.136}$$

which is still k-dimensional.

The most important aspect of this property is the fact that the sufficient statistic is of constant dimension, regardless of the sample size. A reciprocal of this result (with an additional constraint) has also been shown to be true and is known as the Pitman-Koopman theorem; basically it states that if a family of probability functions $f_{\mathbf{X}}(\mathbf{x}|\theta)$ has a sufficient statistic of constant dimension, and if the support of $f_{\mathbf{X}}(\mathbf{x}|\theta)$ (i.e., the set $\{\mathbf{x} : f_{\mathbf{X}}(\mathbf{x}|\theta) > 0\}$) does not depend on θ, then this is necessarily an exponential family [93]. As a counter-example (showing why it is necessary to include the additional condition on the support of the likelihood), consider the uniform likelihood model from Example 2.5.1; although there is a constant dimension sufficient statistic $t(\mathbf{x}) = \max\{x_1, x_2, ..., x_n\}$, the fact that the support of the likelihood does depend on θ prevents the Pitman-Koopman theorem from classifying it as an exponential family (which in fact it is not).

Finally, notice that the definition of exponential families in Eq. (2.133) makes clear that the maximum entropy priors in Eqs. (2.89) and (2.103) have the exact same form. That is, maximum entropy probability distributions constitute exponential families. The $g_k(\cdot)$ functions used to express the constraints for the maximum entropy distributions, appear as statistics in the resulting exponential families.

2.12.2 Partition Function, Free Energy, and Entropy

It is possible to express an exponential family (in natural form, see Eq. (2.134)) as

$$\begin{aligned} f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\xi}) &= \frac{1}{Z(\boldsymbol{\xi})}\,\phi(\mathbf{t})\exp\left\{\boldsymbol{\xi}^T\mathbf{t}\right\} && (2.137)\\ &= \phi(\mathbf{t})\exp\left\{\boldsymbol{\xi}^T\mathbf{t} - \log Z(\boldsymbol{\xi})\right\} && (2.138)\end{aligned}$$

where the logarithm is natural, and $Z(\boldsymbol{\xi}) = 1/\psi(\boldsymbol{\xi})$ is the normalizing constant given by

$$Z(\boldsymbol{\xi}) = \frac{1}{\psi(\boldsymbol{\xi})} = \int \phi(\mathbf{t})\exp\{\boldsymbol{\xi}^T\mathbf{t}\}\,d\mathbf{t}. \tag{2.139}$$

The common name for this constant is imported from statistical physics where it is called the *partition function* (see, *e.g.*, [22], [81]). If we put in evidence a common factor β in the canonical parameter, *i.e.*, writing

$\boldsymbol{\xi} = \beta\boldsymbol{\xi}'$, then, Eq. (2.138) can be written as

$$f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\xi}', \beta) = \phi(\mathbf{t}) \exp\left\{\beta\left[\boldsymbol{\xi}'^T\mathbf{t} - \frac{1}{\beta}\log Z(\boldsymbol{\xi})\right]\right\}. \tag{2.140}$$

Again importing designations from statistical physics, the quantity

$$F(\boldsymbol{\xi}) \equiv -\frac{1}{\beta}\log Z(\boldsymbol{\xi}) \tag{2.141}$$

is known as the *Gibbs free energy*, while $1/\beta$ is called the *temperature.*

Means, variances, covariances, or any other moments of the natural statistics can be computed directly from the derivatives of $\log Z(\boldsymbol{\xi})$. Technically, this is made possible by the fact (which is proved, for example, in [71]) that any integral of the form

$$\int g(\mathbf{x}) \exp\left\{\boldsymbol{\xi}^T\mathbf{t}(\mathbf{x})\right\} \phi(\mathbf{t}(\mathbf{x}))\, d\mathbf{x}, \tag{2.142}$$

where $g(\mathbf{x})$ is integrable, is a continuous function of $\boldsymbol{\xi}$, with derivatives of all orders which can be obtained under the integral. Using this fact,

$$\begin{aligned}\frac{\partial \log Z(\boldsymbol{\xi})}{\partial \xi_i} &= \frac{\partial Z(\boldsymbol{\xi})}{\partial \xi_i}\frac{1}{Z(\boldsymbol{\xi})}\\ &= \frac{\displaystyle\int t_i\, \phi(\mathbf{t}) \exp\{\boldsymbol{\xi}^T\mathbf{t}\}\, d\mathbf{t}}{\displaystyle\int \phi(\mathbf{t}) \exp\{\boldsymbol{\xi}^T\mathbf{t}\}\, d\mathbf{t}} \equiv E_{\mathbf{T}}\left[t_i|\, \boldsymbol{\xi}\right]\end{aligned} \tag{2.143}$$

and, after some additional work,

$$\frac{\partial^2 \log Z(\boldsymbol{\xi})}{\partial \xi_j\, \partial \xi_i} = E_{\mathbf{T}}\left[t_j\, t_i|\, \boldsymbol{\xi}\right] - E_{\mathbf{T}}\left[t_j|\, \boldsymbol{\xi}\right] E_{\mathbf{T}}\left[t_i|\boldsymbol{\xi}\right] \equiv \text{Cov}_{\mathbf{T}}\left[t_j, t_i|\, \boldsymbol{\xi}\right], \tag{2.144}$$

which is the conditional covariance (given $\boldsymbol{\xi}$) between random variables T_i and T_j. An important implication of these equalities, together with Eq. (2.135), is that for a set of n independent and identically distributed observations, any moment of a natural sufficient statistic is simply n times what it is for a single observation.

Finally, notice that for exponential families it is also easy to compute the entropy. Writing $H(\boldsymbol{\xi})$ to stress the fact that the entropy is a function of the canonical parameter. we have

$$H(\boldsymbol{\xi}) = E\left[-\log f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\xi})\right] = \beta\boldsymbol{\xi}^T E[\mathbf{t}|\boldsymbol{\xi}] + \log Z(\boldsymbol{\xi}) + E[\log\phi(\mathbf{t})|\boldsymbol{\xi}].$$

If $\phi(\mathbf{t}) = 1$, this leads to a expression (well known in statistical physics) relating the entropy with the free energy,

$$\boldsymbol{\xi}^T E[\mathbf{t}|\boldsymbol{\xi}] = F(\boldsymbol{\xi}) + TH(\boldsymbol{\xi}),$$

where $T = 1/\beta$ is the temperature.

Example 2.12.1

Consider the coin-tossing example, characterized by the Bernoulli model in Eq. (2.8), $f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})}$. This can be rewritten in the exponential family form as

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \underbrace{\exp\{n\log(1-\theta)\}}_{\psi(\theta)^n} \exp\left\{n_h(\mathbf{x})\log\frac{\theta}{1-\theta}\right\}, \tag{2.145}$$

where the natural sufficient statistic is $t(\mathbf{x}) = n_h(\mathbf{x})$ and the natural parameter is $\xi(\theta) = \log[\theta/(1-\theta)]$; the inverse reparameterization is $\theta = \exp(\xi)/(1+\exp(\xi))$. In natural exponential form,

$$f_{N_h}(n_h|\xi) = \frac{1}{(1+\exp\xi)^n}\exp\{n_h\,\xi\} \tag{2.146}$$

revealing that $Z(\xi) = (1+\exp\xi)^n$ and, as expected,

$$E_{N_h}[n_h|\xi] = \frac{\partial n\log(1+\exp\xi)}{\partial\xi} = n\frac{\exp\xi}{1+\exp\xi} = n\theta. \tag{2.147}$$

End of Example 2.12.1

Example 2.12.2

When estimating the mean, say μ, from independent Gaussian observations of known variance σ^2, the likelihood function $f_X(x|\mu) = \mathcal{N}(x|\mu,\sigma^2)$ for one observation can be put in the exponential family form as

$$f_X(x|\mu) = \underbrace{\exp\left\{-\frac{x^2}{2\sigma^2}\right\}}_{\phi(x)} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}_{\psi(\mu)} \exp\left\{\frac{x\mu}{\sigma^2}\right\}, \tag{2.148}$$

revealing that $\xi(\mu) = \mu$ is the canonical parameter, and $t(x) = x/\sigma^2$ is the natural sufficient statistic (of course, we could instead have associated the $1/\sigma^2$ factor with the parameter and chosen $\xi(\mu) = \mu/\sigma^2$ and $t(x) = x$). Then, for n independent observations, the natural sufficient statistic is simply $t(\mathbf{x}) = (x_1 + x_2 + ... + x_n)/\sigma^2$. The partition function is $Z(\mu) = \sqrt{2\pi\sigma^2}\exp\{\mu^2/(2\sigma^2)\}$, which leads to

$$E_T[t|\mu] = \frac{\partial\left(\mu^2/(2\sigma^2)\right)}{\partial\mu} = \frac{\mu}{\sigma^2}, \tag{2.149}$$

for one observation, or $n\mu/\sigma^2$ for n observations.

End of Example 2.12.2

Example 2.12.3

Concerning the Poisson likelihood, for one single observation interval of duration T, the probability function for a count x is $f_X(x|\theta) = e^{-\theta T}\,(\theta T)^x/x!$. This can be cast into exponential family form according to

$$f_X(x|\theta) = \underbrace{\frac{1}{x!}}_{\phi(x)}\;\underbrace{\exp\{-\theta\, T\}}_{\psi(\theta)}\;\exp\{x\log(\theta T)\} \tag{2.150}$$

showing that the natural parameter is $\xi = \log(\theta T)$, and the natural statistic is x itself. From this, it can immediately be concluded that the sufficient statistic for a set of n counts is simply their sum $t(\mathbf{x}) = x_1 + x_2 + ... + x_n$. The partition function is $Z(\xi) = \exp(\exp(\xi))$, from which the mean can be obtained as

$$E_X[x|\xi] = \frac{\partial \log(\exp(\exp(\xi)))}{\partial \xi} = \frac{\partial \exp(\xi)}{\partial \xi} = \exp(\xi) = \theta T. \tag{2.151}$$

End of Example 2.12.3

Example 2.12.4

Let us look at the problem of estimating the unknown variance σ^2 of zero mean Gaussian observations. The likelihood function for one observation is

$$f_X(x|\sigma^2) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\psi(\sigma^2)}\exp\left\{-\frac{x^2}{2\sigma^2}\right\} \tag{2.152}$$

(with $\phi(x) = 1$), which shows that the natural parameter is $\xi(\sigma^2) = -1/(2\sigma^2)$ and the natural statistic is $t(x) = x^2$. Accordingly, the natural sufficient statistic for a set of observations is $t(\mathbf{x}) = x_1^2 + x_2^2 + ... + x_n^2$, as expected. Notice that the 1/2 factor in the exponent can be associated either with $t(\mathbf{x})$ or with $\xi(\sigma^2)$. Again, it is easy to check via the derivative of the logarithm of the partition function that the expected value of the natural statistic is σ^2, for one observation, or $n\sigma^2$ for n observations.

End of Example 2.12.4

Example 2.12.5

Finally, we consider an example where we are still estimating a scalar parameter, but $\boldsymbol{\xi}(\theta)$ and $\mathbf{t}(\mathbf{x})$ are vector valued. Suppose we have independent and identically distributed observations of a Gaussian likelihood of mean θ and variance θ^2. This is sometimes known as the multiplicative model because it can arise from the following observation model: given θ, each observation is obtained by multiplying it by a sample of a unit mean and unit variance Gaussian random variable. In exponential form, each observation is characterized by

$$f_X(x|\theta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left\{-\frac{(x-\theta)^2}{2\theta^2}\right\} = \underbrace{\frac{1}{\sqrt{2\pi\theta^2}} \exp\left\{-\frac{1}{2}\right\}}_{\psi(\theta)} \exp\left\{-\frac{x^2}{2\theta^2} + \frac{x}{\theta}\right\},$$

which shows that the natural (vector) parameter is $\boldsymbol{\xi}(\theta) = [1/\theta^2 \quad 1/\theta]^T$ and the natural sufficient statistic is $\mathbf{t}(x) = [-x^2/2 \quad x]$. Accordingly, the natural sufficient statistic for a set of n observations is

$$\mathbf{t}(\mathbf{x}) = \left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2 \quad \sum_{i=1}^{n} x_i\right]^T. \tag{2.153}$$

End of Example 2.12.5

2.12.3 Conjugate Priors

Having a likelihood written in the exponential family form brings an additional important advantage: it makes it very easy to write conjugate priors. For a likelihood in exponential form as in Eq. (2.135), any family of priors of the form

$$\mathcal{P} = \left\{p_\Theta(\theta) = \psi(\theta)^\nu \exp\left\{\boldsymbol{\xi}(\theta)^T \boldsymbol{\gamma}\right\}\right\} \tag{2.154}$$

is a conjugate one. Simply notice that the *a posteriori* probability function resulting from multiplying any element of $\mathcal{P}$ by a likelihood of the form in Eq. (2.135) results in

$$p_\Theta(\theta|\mathbf{x}) \propto \psi(\theta)^{(\nu+n)} \exp\left\{\boldsymbol{\xi}(\theta)^T(\boldsymbol{\gamma} + \mathbf{t}(\mathbf{x}))\right\} \tag{2.155}$$

which clearly still belongs to $\mathcal{P}$. From the simple parametric form of the priors in $\mathcal{P}$, the *a posteriori* probability functions are obtained by simple parameter updates: ν becomes $\nu+n$ and $\boldsymbol{\gamma}$ is updated to $\boldsymbol{\gamma}+\mathbf{t}(\mathbf{x})$. Another consequence of this form of conjugate prior is that its effect is similar to observing additional independent data: Eq. (2.155) is similar to a likelihood (Eq. (2.135)) in which ν more data points were observed and that data produced the sufficient statistic $\boldsymbol{\gamma}$. We will now show how this approach could have been used to obtain previously considered conjugate priors. It is important to keep in mind that the notion of conjugacy is not limited to likelihoods in the exponential class; however, for non-exponential families, the conjugate priors are often difficult to obtain and/or impractical to use.

Example 2.12.6

As we saw in Example 2.12.1, the Bernoulli probability function can be put in exponential form using $\psi(\theta) = (1-\theta)$ and $\xi(\theta) = \log(\theta/(1-\theta))$.

According to Eq. (2.154), the resulting conjugate priors have the form

$$p_\Theta(\theta) \propto \exp\{\nu \log(1-\theta)\} \exp\left\{\gamma \log \frac{\theta}{1-\theta}\right\} = \theta^\gamma \, (1-\theta)^{(\nu-\gamma)}, \quad (2.156)$$

which are, in fact, Beta densities $\text{Be}(\theta|\gamma+1, \nu-\gamma+1)$ (recall Example 2.3.2).

End of Example 2.12.6

Example 2.12.7

For the Poisson likelihood (see Example 2.12.3), the natural parameter is $\xi(\theta) = \log(\theta T)$, while $\psi(\theta) = \exp\{-\theta T\}$. Then, conjugate priors have the form

$$p_\Theta(\theta) \propto \exp\{-\nu\,\theta\,T\} \; \exp\left\{\gamma \log(\theta T)\right\} = \exp\{-\nu\,T\,\theta\}(\theta\,T)^\gamma \qquad (2.157)$$

which is, as expected (see Example 2.3.4), a Gamma prior $\text{Ga}(\theta|\gamma,\nu)$ with respect to the normalized parameter θT.

End of Example 2.12.7

Example 2.12.8

As a last example, we return to the problem of estimating the unknown variance σ^2 of a zero mean Gaussian density. As seen in Example 2.12.4, the natural parameter can be $\xi(\sigma^2) = 1/\sigma^2$ while $\psi(\sigma^2) = (2\pi\sigma^2)^{-1/2} = (\xi/2\pi)^{1/2}$. The resulting conjugate prior (in terms of the natural parameter ξ) is then

$$p_\Xi(\xi) \propto (\xi)^{\nu/2} \exp\{\gamma\xi\} \qquad (2.158)$$

which is a $\text{Ga}(\xi|\nu/2+1, -\gamma)$ density (only proper for $\gamma < 0$).

End of Example 2.12.8

2.12.4 Fisher Information and Jeffreys' Priors

For exponential families in canonical form, the Fisher information (and, consequently, the Jeffreys' prior) has a particularly simple form. With a likelihood of the form of Eq. (2.137) inserted in the definition of Fisher information (Eq. (2.55)), considering a scalar parameter ξ,

$$\begin{aligned} \mathcal{I}(\xi) &= E\left[\left(\frac{\partial \log \phi(t)}{\partial \xi} + t - \frac{\partial \log Z(\xi)}{\partial \xi}\right)^2 \middle| \xi\right] \\ &= E[t^2|\xi] - E[t|\xi]^2 \;=\; \text{var}[t|\xi]; \qquad (2.159) \end{aligned}$$

that is, the Fisher information equals the variance of the natural statistic. An alternative form, from Eq. (2.144), is

$$I(\xi) = \frac{\partial^2 \log Z(\xi)}{\partial \xi^2}. \qquad (2.160)$$

Finally, Jeffreys' priors for canonical parameters of exponential families are given by

$$p_{\Xi}(\xi) = \sqrt{\mathrm{var}[t|\xi]} = \sqrt{\frac{\partial^2 \log Z(\xi)}{\partial \xi^2}}, \tag{2.161}$$

the standard deviation of the corresponding canonical statistic.

Example 2.12.9

Consider the the Bernoulli model $f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{n_h(\mathbf{x})}\,(1-\theta)^{n-n_h(\mathbf{x})}$. We saw in Example 2.12.1 that the natural statistic is $t(\mathbf{x}) = n_h(\mathbf{x})$ and the natural parameter is $\xi(\theta) = \log[\theta/(1-\theta)]$. From the results in this section,

$$\mathcal{I}(\xi) = \mathrm{var}[n_h(\mathbf{x})|\xi] = n\theta(\xi)(1-\theta(\xi)) \tag{2.162}$$

where $\theta(\xi) = \exp(\xi)/(1+\exp(\xi))$, and the variance of $n_h(\mathbf{x})$ is the corresponding Binomial distribution. To obtain the Fisher information with respect to the original parameter θ, we use the rule for transforming Fisher informations (see Eq. (2.76)) and obtain, as expected, $\mathcal{I}(\theta) = n/(\theta(1-\theta))$.

End of Example 2.12.9

Example 2.12.10

For the mean μ of a Gaussian observation of known variance σ^2, $\xi(\mu) = \mu$ is a canonical parameter, and $t(x) = x/\sigma^2$ is the corresponding sufficient statistic (see Example 2.12.2). Then

$$\mathcal{I}(\mu) = \mathrm{var}[x/\sigma^2|\mu] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}, \tag{2.163}$$

as we have seen in Example 2.7.1.

End of Example 2.12.10

Example 2.12.11

For the Poisson likelihood (see Example 2.12.3), for one single observation interval of duration T, the natural parameter is $\xi = \log(\theta T)$, and the natural statistic is x itself. The Fisher information is

$$\mathcal{I}(\xi) = \mathrm{var}[x|\xi] = \theta(\xi), \tag{2.164}$$

where $\theta(\xi) = \exp\{\xi\}/T$. Using Eq. (2.76), we obtain the Fisher information with respect to θ, $\mathcal{I}(\theta) = 1/\theta$.

End of Example 2.12.11

Example 2.12.12

Finally, consider the unknown variance σ^2 of a zero mean Gaussian observation x. The natural parameter is $\xi(\sigma^2) = -1/(2\sigma^2)$ and the natural statistic is $t(x) = x^2$ (see Example 2.12.4). The Fisher information is

$$\mathcal{I}(\xi) = \mathrm{var}[t|\xi] = \mathrm{var}[x^2|\xi] = 2\sigma^4(\xi), \tag{2.165}$$

where $\sigma^4(\xi) = 1/(4\xi^2)$. As in the previous examples, invoking Eq. (2.76), we obtain the Fisher information with respect to σ^2, $\mathcal{I}(\sigma^2) = 1/(2\sigma^2)$.

End of Example 2.12.12

2.13 Intrinsic Loss Functions and Density Estimation Problems

The problem we have been addressing is that of obtaining an estimate $\widehat{s}$ of an unknown parameter s, from observations $\mathbf{x}$ obtained according to a likelihood $f_{\mathbf{X}}(\mathbf{x}|s)$, in the presence of a prior $p_S(s)$ and a specified loss function $L(s,\widehat{s})$. As we have seen, both the prior and the loss function determine the optimal Bayes' decision rule. In an attempt to remove the subjectiveness involved in creating priors, several approaches have been proposed to automatically "extract" non-informative priors from the likelihood functions (see Section 2.6). However, only recently has the same goal been pursued concerning loss functions [93], [92].

It may make sense in many situations to evaluate the performance of some estimate $\widehat{s}$ of s by measuring how "close" $f_{\mathbf{X}}(\cdot|\widehat{s})$ is to $f_{\mathbf{X}}(\cdot|s)$ (where the argument is omitted to stress that the evaluation is with respect to the whole function and not for a particular point). In the situations where these performance measures make sense, there is usually a different perspective: rather than producing an estimate $\widehat{s}$ of s, the (equivalent) goal is to come up with an estimate $f_{\mathbf{X}}(\cdot|\widehat{s})$ of the unknown (probability) density (function) $f_{\mathbf{X}}(\cdot|s)$ that is assumed to have generated the data. These are called (parametric) *density estimation* problems (see, e.g., [33], [101]). Usually, in density estimation problems, the observed data consists of a set of, say n, independent observations $\mathbf{x} = \{\mathbf{x}_i, i = 1,2,...,n\}$, which are identically distributed according to an unknown density to be estimated; then

$$f_{\mathbf{X}}(\mathbf{x}|s) = \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i|s);$$

there is, however, no conceptual constraint stopping us from considering any other observation models. Such a *density estimation* perspective for Bayesian parameter estimation problems clearly demands loss functions that directly compare the involved likelihoods $f_{\mathbf{X}}(\cdot|\widehat{s})$ and $f_{\mathbf{X}}(\cdot|s)$; this provides a natural solution to the above mentioned desideratum of having loss functions objectively derived from the likelihoods, which may then be called *intrinsic* [92].

Although other choices are possible (specially in non-parametric settings [101]), [33]), we will consider here the two proposed in [92]: the Kullback-Leibler divergence (recall its definition in Section 2.8.2)

$$L_{\text{K-L}}(s,\widehat{s}) \equiv D\left[f_{\mathbf{X}}(\cdot|s) \parallel f_{\mathbf{X}}(\cdot|\widehat{s})\right] \qquad (2.166)$$

and the (squared) Hellinger-2 distance

$$L_{\mathrm{H}}(s,\widehat{s}) \equiv H_2^2\left(f_{\mathbf{X}}(\mathbf{x}|s), f_{\mathbf{X}}(\mathbf{x}|\widehat{s})\right) \equiv \frac{1}{2}E_X\left[\left(\sqrt{\frac{f_{\mathbf{X}}(\mathbf{x}|\widehat{s})}{f_{\mathbf{X}}(\mathbf{x}|s)}}-1\right)^2\middle| s\right]. \tag{2.167}$$

The main feature of these likelihood related loss functions is that (as non-informative priors) they are independent of the parameterization chosen. If the likelihood functions are somehow reparameterized, say into $\eta(s)$, these loss functions retain their behavior. The same is, of course, not true about loss functions which explicitly involve the parameters and parameter estimates themselves. The unfortunate disadvantage of this type of loss functions is that they usually do not allow deriving closed-form Bayes' rules; exceptions occur for some special likelihood models (in fact for exponential families, see [92], for details), as illustrated in the following examples.

Example 2.13.1

Consider a Gaussian likelihood function, i.e., $f_X(x|s) = \mathcal{N}(x|s,\sigma^2)$, with known variance σ^2. The Kullback-Leibler divergence loss function is given by

$$\begin{aligned} L(s,\widehat{s}) &= E_X\left[\log\left(\frac{\exp\left\{-\dfrac{(x-s)^2}{2\sigma^2}\right\}}{\exp\left\{-\dfrac{(x-\widehat{s})^2}{2\sigma^2}\right\}}\right)\middle| s\right] \\ &= \frac{1}{2\sigma^2}E_X\left[(x-\widehat{s})^2-(x-s)^2\middle| s\right] = \frac{1}{2\sigma^2}(\widehat{s}-s)^2, \end{aligned} \tag{2.168}$$

which coincides with a quadratic loss function and thus leads to the same Bayesian rules.

Example 2.13.2

For a similar Gaussian likelihood $f_X(x|s) = \mathcal{N}(x|s,\sigma^2)$, of known variance σ^2, the Hellinger distance loss function is

$$L(s,d) = 1-\exp\left\{-\frac{(\widehat{s}-s)^2}{8\sigma^2}\right\} \tag{2.169}$$

which has the aspect shown in Figure 2.10. Notice how it can be interpreted as a smoothed version of the "0/1" loss function. In fact, this loss function converges to the "0/1" when σ^2 goes to zero; an intuitive interpretation of this fact is that as the observation variance approaches zero, the loss function becomes less forgiving.

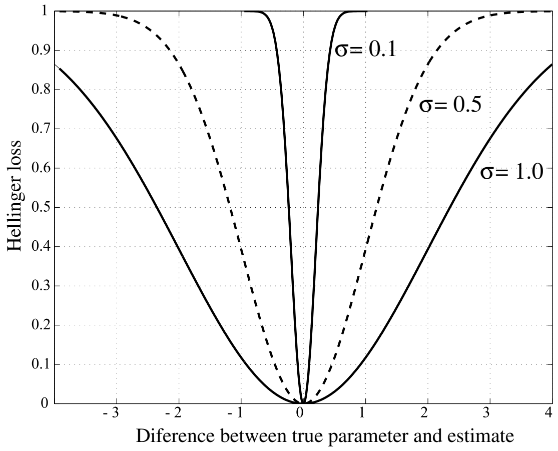

FIGURE 2.10. Hellinger loss function for three different likelihood variances.

2.14 Foundational Aspects: The Likelihood, Conditionality, and Sufficiency Principles

The fact that all Bayesian rules are based on the posterior probability function, which in turn is computed via Bayes' theorem, has an important consequence: they will automatically obey the so-called *likelihood principle.* This principle, which is due to Fisher [43] and was formalized by Birnbaum [15], can be stated as follows:

Likelihood principle (first part): *The information about the state of nature s contained in the observation* $\mathbf{x}$ *can only be carried via the likelihood function* $f_{\mathbf{X}}(\mathbf{x}|s)$, for that observed $\mathbf{x}$.

In fact, it is clear from Eqs. (1.9) and (1.10) that the only dependence of the posterior $p_S(s|\mathbf{x})$ on the observation $\mathbf{x}$ is, in fact, mediated by the likelihood $f_{\mathbf{X}}(\mathbf{x}|s)$, as required by the likelihood principle. Notice that once plugged into Bayes' theorem, $f_{\mathbf{X}}(\mathbf{x}|s)$ stops being seen as a function of $\mathbf{x}$ and begins playing its role as a function of s. This observation underlies the *second part* of the likelihood principle:

Likelihood principle (second part): *If two observations* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *are obtained according to two likelihood functions* $f_{\mathbf{X}}(\mathbf{x}_1|s)$ *and* $g_{\mathbf{X}}(\mathbf{x}_2|s)$ *such that there exists a function* $\psi(\mathbf{x}_1, \mathbf{x}_2)$ *(not a function of s) verifying*

$$g_{\mathbf{X}}(\mathbf{x}_2|s) = \psi(\mathbf{x}_1, \mathbf{x}_2)\, f_{\mathbf{X}}(\mathbf{x}_1|s), \tag{2.170}$$

then, both observations should necessarily lead to the same decision.

Notice that this simply formalizes the fact that all that matters is the "shape" of the likelihood, as a function of s. Again, this property is automatically verified by the *a posteriori* probability function obtained via

Bayes' law,

$$\begin{aligned} p_S(s|\mathbf{x}_2) &= \frac{g_{\mathbf{X}}(\mathbf{x}_2|s)p_S(s)}{\displaystyle\int_{\mathcal{S}} g_{\mathbf{X}}(\mathbf{x}_2|s)p_S(s)\,ds} \\ &= \frac{\psi(\mathbf{x}_1,\mathbf{x}_2)f_{\mathbf{X}}(\mathbf{x}_1|s)p_S(s)}{\displaystyle\int_{\mathcal{S}} \psi(\mathbf{x}_1,\mathbf{x}_2)f_{\mathbf{X}}(\mathbf{x}_1|s)p_S(s)\,ds} \\ &= \frac{f_{\mathbf{X}}(\mathbf{x}_1|s)p_S(s)}{\displaystyle\int_{\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}_1|s)p_S(s)\,ds} = p_S(s|\mathbf{x}_1), \end{aligned} \tag{2.171}$$

and, consequently, by any Bayesian decision rule.

Example 2.14.1

Let us look back at Example 2.3.2; imagine that rather than knowing the particular sequence $\mathbf{x}$ of outcomes observed, we are only informed about the total number of heads, $n_h(\mathbf{x})$; recall from Section 2.11 that this is a sufficient statistic. In this case, we should have to redefine the observation model to be a binomial distribution

$$f_{N_h}(n_h(\mathbf{x})|\theta) = \binom{n}{n_h(\mathbf{x})}\theta^{n_h(\mathbf{x})}(1-\theta)^{n-n_h(\mathbf{x})}. \tag{2.172}$$

According to the likelihood principle, this is an irrelevant change because the Bernoulli and the binomial distributions only differ by a multiplicative factor (in this case the binomial coefficients), thus having the same shape as functions of θ.

End of Example 2.14.1

The likelihood principle, although intuitively satisfactory, can be shown to be equivalent to the conjunction of two other (arguably more universally accepted and self-evident) principles: the *sufficiency principle* and the *conditionality principle*. The proof is beyond the scope of this text and any interested reader should consult [10] or [93]. The *sufficiency principle* (together with the concept of *sufficient statistic*) was addressed in Section 2.11. The *conditionality principle* can be understood by imagining the following situation. In some decision problem, there are two observation mechanisms available, characterized by two different likelihood functions $f_{\mathbf{X}}(\mathbf{x}|s)$ and $g_{\mathbf{X}}(\mathbf{x}|s)$; before acquiring any data, one of these is chosen at random, with probabilities p_f and $p_g = 1 - p_f$, respectively. The conditionality principle states that:

Conditionality Principle: *Any inference concerning s must only depend on the particular observation mechanism that was used, and not on any other one that could have been, but was not, used.*

Notice that frequentist criteria clearly violate this principle since they involve averaging (see Eq. (1.2)) over all possible experimental outcomes, i.e., with respect to the global likelihood function which is a mixture $h_{\mathbf{x}}(\mathbf{x}|s) = p_f\, f_{\mathbf{X}}(\mathbf{x}|s) + p_g\, g_{\mathbf{X}}(\mathbf{x}|s)$.

The equivalence between the conjunction of the sufficiency and conditionality principles and the likelihood principle has extremely important implications: it means that rejecting the likelihood principle implies denying either the sufficiency principle (which is equally accepted by Bayesian and non-Bayesian statisticians) and/or the conditionality principle, which is so self-evident. We conclude by presenting an example (adapted from [8]) of how not following the likelihood principle (or the conditionality principle) may lead to some strange situations.

Example 2.14.2

An engineer wishes to study the stability of some new transistor fabrication process. To do so, gain measurements from a randomly chosen sample are taken, using a high precision meter, and reported to the company's statistician. The statistician performs his classical (frequentist) analysis whose output is a confidence interval for the mean gain. Later that day, the engineer and the statistician meet for coffee, and the engineer says: "I was lucky that none of the measured transistors showed a gain above 1000; the meter I was using only measures gains up to 1000." The statistician suddenly looks upset and says "but that is bad news! The fact that your meter's range is only up to 1000 means that your data is *censored.* I'll have to repeat my analysis taking that into account". But the engineer replies: "well, but I have another equally precise meter that measures gains up to 10000, which I would have used if any gain had been greater than 1000". "Oh! So your data was not censored, after all, and my analysis is still valid" says the statistician. One week later the engineer calls the statistician: "Remember that gain meter that measures up to 10000? It was being repaired, the day I took the measurements; after all, I could not have used it". When the statistician tells the engineer that he will have to repeat his analysis, the engineers replies: "but the data I gave you is exactly the same as if the meter had been working; how can it make any difference? What if I now tell you that although our high-range meter was being repaired, I could have borrowed one from another laboratory, if necessary? Would you change your mind once more and tell me that after all the initial conclusions are again valid?"

End of Example 2.14.2

2.15 Summary

In this Chapter we have considered various aspects surrounding Bayesian inference. A considerable part of it was devoted to the issue of how to specify priors. We started by looking at the concept of improper priors and how it allows recovering the *maximum likelihood* criterion as a particular Bayesian decision rule. Conjugate priors, i.e., priors designed to match the likelihood function in the sense that the resulting *a posteriori* probability function is easily obtained, were considered next. Non-informative priors are conceived in an attempt to remove some of the subjectiveness from Bayesian inference; we saw, in particular, how invariance arguments come into play and how they lead to Jeffreys' priors. Finally, we concluded our study of priors with two design criteria rooted in information-theoretic concepts: maximum entropy and minimum description length; in passing, we introduced and studied the important concepts of entropy, Kullback-Leibler divergence, and mutual information. These concepts were used to provide an information theoretical view of some issues addressed (namely, Jeffreys' priors and sufficient statistics).

Sections 2.11 and 2.12 were devoted to two highly interrelated issues: sufficient statistics, and exponential families. In particular, it was seen how exponential families allow a very convenient manipulation including how conjugate priors can be easily obtained in that case.

Section 2.13 was devoted to intrinsic loss functions (in the sense that they are derived from the likelihood, rather than chosen arbitrarily), which are the counterpart of non-informative priors. A density estimation perspective was seen to be adequate for this purpose.

The last section focused on some foundational aspects of Bayesian inference: the likelihood, conditionality, and sufficiency principles.

3

Compound Decision Theory and Contextual Decision Making

3.1 Introduction

The previous Chapters have not dealt with problems (either of estimation or classification) where the unknown quantities are vectors, i.e., have several elements. Although the Bayesian approaches described are very general and by no means restricted to scalar problems, some issues deserve special attention. Problems where a set of (possibly dependent) decisions (rather than a single one) are to be taken fall in the general category of *compound decision theory.* Another common term is *contextual decision* making, which emphasizes the idea that each decision should take into account its contextual information, i.e. the other decisions that can influence the current decision making task. Contextual information is extremely important in many application areas such as pattern recognition, and image analysis; context often allows us to reduce or eliminate ambiguities or errors, and to recover missing information, as pointed out by Haralick [51].

To formalize compound decision theory, it is required that the states of nature be described by d-dimensional vectors[1] $\mathbf{s} = [s_1, s_2, \ldots, s_d]^T \in \mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_d$, with × denoting Cartesian product. In many problems, all the elements of $\mathbf{s}$ belong to the same configuration space and thus the configuration space is simply a Cartesian power, $\mathcal{S} = \mathcal{S}_0^d$. When the vector

[1] The notation s_i used here should not be confused with the one used in previous Chapters to enumerate all the elements of set $\mathcal{S}$.

nature of $\mathbf{s}$ is not relevant, we can simply write $\mathbf{s} = (s_1, s_2, ..., s_d)$ and view it simply as an ordered set.

For compound problems, the role of the prior $p_{\mathbf{S}}(\mathbf{s})$ is now, not only to express prior knowledge concerning each s_i, but also about their (probabilistically modeled) mutual interdependencies; this is arguably the key point in compound/contextual inference since it is this joint prior that formalizes how each s_i is expected to be influenced and to influence the other states (i.e., its context).

Let us consider a Bayesian inference problem with $\mathcal{A} = \mathcal{S}$, where the goal is to design a decision rule $\boldsymbol{\delta}(\mathbf{x}) : \mathcal{X} \to \mathcal{S}$ by minimizing the *a posteriori expected loss* associated with a given loss function $L(\boldsymbol{s}, \mathbf{a})$. As in the non-compound case, observed data $\mathbf{x}$ is obtained according to the likelihood function $f_{\mathbf{X}}(\mathbf{x}|\mathbf{s})$ which is considered known.

Sometimes, the observation model has a particular feature that allows simplifying some of the results ahead; this is called *conditional independence* and applies to problems where the likelihood function can be factored according to

$$f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}) = \prod_{i=1}^{d} f_{\mathbf{X}_{(i)}}(\mathbf{x}_{(i)}|s_i), \tag{3.1}$$

where $\mathbf{x}_{(i)}$ stands for a subset of data exclusively associated with the component s_i of $\mathbf{s}$, which is assumed conditionally independent (conditioned on $\mathbf{s}$) from the other subsets. Each $\mathbf{x}_{(i)}$ can be interpreted as *an observation of* s_i.

It is convenient to separately treat loss functions that can be decomposed according to

$$L(\boldsymbol{s}, \mathbf{a}) = \sum_{i=1}^{d} L_i(s_i, a_i), \tag{3.2}$$

which we will call *additive*, from those that do not allow such a decomposition, named *non-additive*. Before proceeding, we conclude this introduction by pointing out that, in the compound case, the definition of the marginal $f_{\mathbf{X}}(\mathbf{x})$ (see Eq. (1.10)) now involves multiple integrals or summations. In mixed situations, where some of the $\mathcal{S}_i$'s may be discrete while others are continuous, computing $f_{\mathbf{X}}(\mathbf{x})$ involves nested summations and integrals.

3.2 Non-additive Loss Functions

Non additive loss-functions are those that can not be written as Eq. (3.2). This is the most general situation and naturally leads to decision rules formally equivalent to those obtained for non-compound problems. Next, we will look at some instances of non-additive loss functions.

3.2.1 The "0/1" Loss Function

The non-additive "0/1" loss function for compound classification problems is the natural extension of the non-compound one in Eq. (1.22):

$$L(\boldsymbol{s}, \mathbf{a}) = \begin{cases} 1 & \Leftarrow \quad \boldsymbol{s} \neq \mathbf{a} \\ 0 & \Leftarrow \quad \boldsymbol{s} = \mathbf{a}. \end{cases} \tag{3.3}$$

Clearly, this loss does not allow a decomposition as the one in Eq. (3.2). By observing that there is absolutely nothing in Eqs. (1.22) to (1.25) that makes them specific to univariate problems, we conclude that the MAP decision rule

$$\delta(\mathbf{x}) = \arg\max_{\mathbf{s}\in\mathcal{S}} p_{\mathbf{S}}(\mathbf{s}|\mathbf{x}) = \arg\max_{\mathbf{s}\in\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}) p_{\mathbf{S}}(\mathbf{s}), \tag{3.4}$$

is still the optimal Bayes' criterion for classification problems under this loss function. This (joint) MAP rule looks for the configuration $(\widehat{s}_1, \widehat{s}_2, \ldots, \widehat{s}_d)_{\text{MAP}}$ that jointly globally maximizes the *a posteriori* (joint) probability function $p_{\mathbf{S}}(\mathbf{s}|\mathbf{x})$. As in the non-compound case, if the prior is uniform, the resulting rule is called the *maximum likelihood* (ML) classifier.

For estimation problems, the "0/1" loss function is also similar to Eq. (1.79). Again, Eqs. (1.79) to (1.81) apply equally well to compound situations, leading to a compound MAP estimation rule formally similar to Eq. (3.4), except for the fact that $p_{\mathbf{S}}(\mathbf{s}|\mathbf{x})$ and $p_{\mathbf{S}}(\mathbf{s})$ are now (multivariate) probability density functions rather than probability mass functions. Since the MAP classifier and estimator are formally similar, the same expression also applies to mixed problems.

A special situation allows simplifying the general MAP rule of Eq. (3.4): if conditional independence (see Eq. (3.1)) is assumed for the observation model and, in addition, the prior can be factored into

$$p_{\mathbf{S}}(\mathbf{s}) = \prod_{i=1}^{d} p_{S_i}(s_i) \tag{3.5}$$

i.e., it models the components of $\mathbf{S}$ as *a priori* independent, then the compound MAP rule reduces to a non-interacting set of scalar MAP rules

$$\begin{aligned} \boldsymbol{\delta}_{\text{MAP}}(\mathbf{x}) &= \arg\max_{(s_1,\ldots,s_d)} \left\{ \left(\prod_{i=1}^{d} f_{\mathbf{X}_{(i)}}(\mathbf{x}_i|s_i) \right) \left(\prod_{i=1}^{d} p_{S_i}(s_i) \right) \right\} \\ &= \arg\max_{(s_1,\ldots,s_d)} \left\{ \prod_{i=1}^{d} f_{\mathbf{X}_{(i)}}(\mathbf{x}_i|s_i) p_{S_i}(s_i) \right\} \\ &= \left[\arg\max_{s_1} \{ p_{S_1}(s_1|\mathbf{x}_{(1)}) \}, \ldots, \arg\max_{s_d} \{ p_{S_d}(s_d|\mathbf{x}_{(d)}) \} \right]^T . \end{aligned}$$

Finally, it is worth mentioning that Eq. (3.4) is probably, as we shall see, the most widely used criterion in Bayesian image analysis applications, such as restoration, reconstruction, segmentation, and classification.

3.2.2 Quadratic Loss Function

A general quadratic loss function for vectors ($\mathcal{A} = \mathcal{S} = I\!R^n$) is

$$L(\boldsymbol{s}, \mathbf{a}) = (\boldsymbol{s} - \mathbf{a})^T \mathbf{Q} (\boldsymbol{s} - \mathbf{a}); \tag{3.6}$$

only in the particular case of $\mathbf{Q}$ being diagonal, say $\mathbf{Q} = \mathrm{diag}(q_1, q_2, ..., q_d)$ do we get

$$L(\boldsymbol{s}, \mathbf{a}) = \sum_{i}^{d} q_i (s_i - a_i)^2 \tag{3.7}$$

which is an additive loss function. However, a general result which covers both cases can be obtained: for any symmetric positive-definite $\mathbf{Q}$, the optimal Bayes' estimator is always the posterior mean [93]. To see that this is true, notice that the *a posteriori expected loss* can be written as

$$\begin{aligned} \rho\left(p_{\mathbf{S}}(\mathbf{s}), \boldsymbol{\delta}(\mathbf{x})|\mathbf{x}\right) &= E\left[(\mathbf{s} - \boldsymbol{\delta}(\mathbf{x}))^T \mathbf{Q}(\mathbf{s} - \boldsymbol{\delta}(\mathbf{x}))|\mathbf{x}\right] \\ &= E\left[\mathbf{s}^T\mathbf{Q}\mathbf{s}|\mathbf{x}\right] + \boldsymbol{\delta}(\mathbf{x})^T\mathbf{Q}\boldsymbol{\delta}(\mathbf{x}) - 2\boldsymbol{\delta}(\mathbf{x})^T\mathbf{Q}E\left[\mathbf{s}|\mathbf{x}\right], \end{aligned}$$

because $\mathbf{Q}$ is symmetric. Now, since $\mathbf{Q}$ is positive-definite this is a convex function of $\boldsymbol{\delta}(\mathbf{x})$ and so we can find its minimum by computing the gradient and equating it to zero; this yields

$$\boldsymbol{\delta}_{\mathrm{PM}}(\mathbf{x}) = E_{\mathbf{S}}\left[\mathbf{s}|\mathbf{x}\right]. \tag{3.8}$$

Finally, notice that it is quite natural to demand that $\mathbf{Q}$ be positive-definite. In fact, if $\mathbf{Q}$ is not positive-definite it is possible to have cases where $\|\mathbf{s} - \mathbf{a}\|^2$ is arbitrarily large, but $(\mathbf{s} - \mathbf{a})^T\mathbf{Q}(\mathbf{s} - \mathbf{a}) = 0$, which does not (in general) seem reasonable for a loss function.

3.2.3 Likelihood Related Loss Functions

The likelihood related loss functions considered in Section 2.13 are non-additive, and, in general, no explicit expressions can be found for the resulting Bayes' criteria. Worth mentioning is the particular case of the Kullback-Leibler divergence loss function in the presence of a likelihood function that has the *conditional independence* property (see Eq. (3.1)); in that case, because of the linearity of the expected value,

$$\begin{aligned} L(\mathbf{s}, \mathbf{a}) &= E_{\mathbf{X}}\left[\log \frac{f_{\mathbf{X}}(\mathbf{x}|\mathbf{s})}{f_{\mathbf{X}}(\mathbf{x}|\mathbf{a})}\middle| \mathbf{s}\right] \\ &= E_{\mathbf{X}}\left[\sum_{i=1}^{d} \log \frac{f_{\mathbf{X}_{(i)}}(\mathbf{x}_{(i)}|s_i)}{f_{\mathbf{X}_{(i)}}(\mathbf{x}_{(i)}|a_i)}\middle| \mathbf{s}\right] \\ &= \sum_{i=1}^{d} E_{\mathbf{X}}\left[\log \frac{f_{\mathbf{X}_{(i)}}(\mathbf{x}_{(i)}|s_i)}{f_{\mathbf{X}_{(i)}}(\mathbf{x}_{(i)}|a_i)}\middle| s_i\right] \end{aligned} \tag{3.9}$$

which is clearly an additive loss function.

3.3 Linear-Gaussian Observations and Gaussian Priors

One of the most simple and important classes of compound estimation problems is that where the likelihood is multivariate Gaussian; despite its simplicity, it provides an extremely rich and versatile family of models which is widely used in many image and signal analysis applications, not to mention its many other uses in statistical analysis.

Formally, let the unknown be a d-dimensional vector be denoted as $\mathbf{s}$. Let $\mathbf{x}$ be the n-dimensional observed vector which is linearly related to $\mathbf{s}$ through the observation equation

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{n} \tag{3.10}$$

where $\mathbf{H}$ is a $n \times d$ (possibly non-invertible) matrix, and $\mathbf{n}$, usually called *noise*, is a sample of a zero mean Gaussian random vector $\mathbf{N}$, independent of $\mathbf{S}$, with covariance matrix $\mathbf{C}$. The resulting likelihood function is

$$f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}) = \mathcal{N}(\mathbf{x}|\mathbf{H}\mathbf{s}, \mathbf{C}) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathbf{C})}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{H}\mathbf{s})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{H}\mathbf{s}) \right\}. \tag{3.11}$$

This is, for example, the standard observation model in image restoration/reconstruction problems; in that case, $\mathbf{H}$ models the degradation or observation process (*e.g.*, blur or tomographic projections) and the goal is to estimate the uncorrupted image $\mathbf{s}$.

3.3.1 Regression Models

In the statistics literature, Eq. (3.10) is called a *linear regression model* and is one of the most common and studied statistical analysis tools; see recent specialized accounts in [86] and [99], and Bayesian perspectives in [18], [46], and [93]. In the regression context, it is assumed that some variable[2] x, called the *response* (here assumed real-valued) can be *explained* as a function (usually deterministic, but not necessarily so) of the so-called *explanatory* variables $\mathbf{y} = [y_1, y_2, ..., y_d]$. There are, of course, many possible interpretations for the word *explain*; the simplest one, leading to standard linear regression models is that, given $\mathbf{y}$, x is a linear combination of the elements of $\mathbf{y}$

$$x = \mathbf{y}^T \mathbf{s} = y_1\, s_1 + y_2\, s_2 + ... + y_d\, s_d. \tag{3.12}$$

The problem here is to estimate the weights of this linear combination, after having observed a set of, say n, *examples* or *subjects* $\{(x_i, \mathbf{y}_i),\quad i =$

[2] We are not using standard regression notation to stay compatible with Eq. (3.10).

$1, 2, ..., n\}$, with $\mathbf{y}_i = [y_{i1}, ..., y_{id}]^T$, where y_{ij} is the j-th explanatory variable of the i-th subject. Moreover, it is assumed that the x_i's are observed with noise. The joint observation model for the set of examples is

$$\mathbf{x} = \mathbf{Hs} + \mathbf{n} \tag{3.13}$$

(see Eq. (3.10)) where $\mathbf{x} = [x_1, ..., x_n]$, matrix $\mathbf{H}$ (often called the *design matrix*) collects the explanatory variables, that is $H_{ij} = y_{ij}$, and $\mathbf{n}$ is a sample of a zero mean Gaussian random vector with known covariance matrix $\mathbf{C}$. An alternative view, leading to the exact same equations, is that the underlying function is itself random; for example, given $\mathbf{y}$, x is a sample of a Gaussian random variable whose mean is a linear combination $\mathbf{y}^T\mathbf{s}$. If the noise in the response variable of each subject has a common variance σ^2 and is independent of the other subjects, we have $\mathbf{C} = \sigma^2\mathbf{I}$; in the regression literature this is called a *homoscedastic* model. In a *heteroscedastic* model, $\mathbf{C}$ is also diagonal, but its elements are not equal.

What about when the underlying function (whose parameters we seek) is non-linear? The answer is that, even in this case, it is possible to end up with a linear regression problem; all that is necessary is that we look for a representation of that function in terms of a fixed basis (not necessarily orthogonal); this may be called *generalized linear regression.* To be more specific, let the underlying function to be estimated be written as $x = \xi(\mathbf{y}; \mathbf{s})$, where $\mathbf{s}$ are the coefficients of its representation with respect to some (fixed) set of (linear independent) functions

$$\xi(\mathbf{y}; \mathbf{s}) = \sum_{j=1}^{d} s_j \; \phi_j(\mathbf{y}). \tag{3.14}$$

Classical examples of such representations are Fourier-type (complex exponentials, cosines, sines) series, B-spline bases (see [41], [29], [36], and Example 3.3.1 below), polynomials, wavelet bases [73], or radial basis functions (RBF) (see [83], [85], references in [52] and [87], and Example 3.3.2). Notice that the linear regression model in Eq. (3.12) is a particular case of Eq. (3.14) with $\phi_j(\mathbf{y}) = y_j$. Consider again a set of n observations, $\{(x_i, \mathbf{y}_i), \quad i = 1, 2, ..., n\}$, such that x_i is a noisy observation of $\xi(\mathbf{y}_i; \mathbf{s})$. The observation model for this set of data can again be written as in Eq. (3.13), where the elements of matrix $\mathbf{H}$ are now given by $H_{ij} = \phi_j(\mathbf{y}_i)$.

Example 3.3.1

Splines are a very powerful and popular function approximation tool; in particular, they have been used in computer graphics and, more recently, in computer vision and image analysis (see [41] and references therein). In this example, we will briefly see how fitting univariate spline models to observed data leads to linear regression problems. For more detailed accounts on splines see [29], [36], and [106].

Let $\{y_0 < y_1 < y_2 < ... < y_k\} \subset [y_0, y_k] \subset \mathbb{R}$ be a set of points, called *knots* [3]. By definition, spline functions are polynomial inside each interval $[t_{i-1}, t_i]$ and exhibit a certain degree of continuity at the knots. The set of all splines on $[t_m,\, t_{k-m}]$ with C^{m-1} continuity at the knots (that is, that have $m-1$ continuous derivatives) is a linear space of dimension $(k-m)$. The set of B-splines, denoted $\{\mathcal{B}_k^m(t), k\!=\!0,..,k\!-\!m\!-\!1\}$, and which can be defined by the Cox-deBoor recursion,

$$\mathcal{B}_i^m(y) = \frac{(y-y_i)\mathcal{B}_i^{m-1}(y)}{y_{i+m}-y_i} + \frac{(y_{i+m+1}-y)\mathcal{B}_{i+1}^{m-1}(y)}{y_{i+m+1}-y_{i+1}},$$

with

$$\mathcal{B}_i^0(y) = \begin{cases} 1, & y_i \le y < y_{i+1} \\ 0, & \text{otherwise,} \end{cases}$$

constitute a basis of this space, though a non-orthogonal one. Accordingly, each spline function $\xi(y)$ in this space has a unique representation as a linear combination with weights $\mathbf{s} = \{s_0, s_1, ..., s_{k-m-1}\}$:

$$\xi(y) = \sum_{j=0}^{k-m-1} s_j\, \mathcal{B}_j^m(y), \quad y \in [y_m,\, y_{k-m}]. \tag{3.15}$$

Now, given m and a set of knots $\{y_0 < y_1 < ... < y_k\}$ (which in turn define a fixed B-spline basis), consider the problem of estimating the spline function that best fits a set of n (possibly noisy) samples $\{(x_i, y_i),\ i = 0, ..., N-1\}$. The resulting observation model is similar to Eq. (3.13), with the elements of matrix $\mathbf{H}$ now given by $H_{ij} = \mathcal{B}_j^m(y_i)$. If the knots are unknown, thus also having to be inferred from the data, we no longer have a linear problem; that more general case is called the *free-knot* problem and it is considerably more difficult.

End of Example 3.3.1

Example 3.3.2

Another approach which has been widely advocated to represent functions (and arguably better suited than splines to high dimensional cases) is based on radial basis functions (RBFs) (see [83], [85], and references in [52] and [87]). A radial basis function is any function that is characterized by a *center* $\mathbf{c}$ and a *width* h in such way that

$$\psi(\mathbf{y}, \mathbf{c}, h) = \psi\left(\frac{\parallel \mathbf{y}-\mathbf{c} \parallel}{h}\right),$$

[3] For simplicity, we exclude here the possibility of multiple knots, *i.e.*, one or more knots at the same location; see, e.g., [36] for the consequences of this option.

where $\| \cdot \|$ denotes (usually but not necessarily) the Euclidean norm. Examples of RBFs include Gaussian functions, $\psi(r) = \exp(-r^2)$, multiquadratic $\psi(r) = (r^2 + a^2)^{1/2}$, and the so-called *thin-plate-spline* function, $\psi(r) = r^2 \log r$ [87]. Let a set of basis functions with fixed centers $\{\mathbf{c}_j,\ j = 1, ..., d\}$ and widths $\{h_j,\ j = 1, ..., d\}$ be given, $\{\phi_j(\mathbf{y}), j = 1, ..., d\}$, where $\phi_j(\mathbf{y}) = \psi(\mathbf{y}, \mathbf{c}_j, h_j)$. In this case, we fall exactly in a generalized linear regression problem, with the elements of matrix $\mathbf{H}$ again given by $H_{ij} = \phi_j(y_i)$. Of course, if the centers and widths are also unknown, the problem looses its linear regression nature.

End of Example 3.3.2

3.3.2 Gaussian Prior: General Solution

Consider a Gaussian prior of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}$, i.e.,

$$p_{\mathbf{S}}(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}, \mathbf{A}) = \frac{(2\pi)^{-d/2}}{\sqrt{\det(\mathbf{A})}} \exp\left\{ -\frac{1}{2} (\mathbf{s} - \boldsymbol{\mu})^T \mathbf{A}^{-1} (\mathbf{s} - \boldsymbol{\mu}) \right\}. \quad (3.16)$$

The *a posteriori* probability density function can be obtained via Bayes' law, $p_{\mathbf{S}}(\mathbf{s}|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}) p_{\mathbf{S}}(\mathbf{s}) / f_{\mathbf{X}}(\mathbf{x})$. Somewhat lengthy but straightforward manipulations eventually lead to a Gaussian *a posteriori* density

$$p_{\mathbf{S}}(\mathbf{s}|\mathbf{x}) = \mathcal{N}(\mathbf{x}|\widehat{\mathbf{s}}, \mathbf{P}), \quad (3.17)$$

showing that multivariate Gaussians are conjugate priors for Gaussian likelihoods. The mean and covariance matrix are given, respectively, by

$$\widehat{\mathbf{s}} = \boldsymbol{\mu} + \mathbf{A}\mathbf{H}^T \left(\mathbf{C} + \mathbf{H}\mathbf{A}\mathbf{H}^T \right)^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\mu}), \quad (3.18)$$

and

$$\mathbf{P} = \mathbf{A} - \mathbf{A}\mathbf{H}^T \left(\mathbf{C} + \mathbf{H}\mathbf{A}\mathbf{H}^T \right)^{-1} \mathbf{H}\mathbf{A} = \left[\mathbf{A}^{-1} + \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H} \right]^{-1}. \quad (3.19)$$

Equality between the two alternative forms for $\mathbf{P}$ springs directly from the *matrix inversion lemma* (see, for example, [98]). Of course, since the *a posteriori* density is Gaussian, both the MAP and PM estimates are equal to $\widehat{\mathbf{s}}$ (this being why we called it $\widehat{\mathbf{s}}$, in the first place). This result appears under several names and has many disguised equivalents; the best known of these, at least in the signal processing literature, is the *Gauss-Markov* theorem [98].

If the *a posteriori* covariance matrix is not of interest, $\widehat{\mathbf{s}}$ can be more directly obtained as a MAP estimate from Eq. (1.83); after dropping additive constants, we are left with

$$\widehat{\mathbf{s}} = \arg\min_{\mathbf{s}} \left\{ \mathbf{s}^T \left[\mathbf{A}^{-1} + \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H} \right] \mathbf{s} - 2\mathbf{s}^T \left(\mathbf{A}^{-1} \boldsymbol{\mu} + \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} \right) \right\}. \quad (3.20)$$

Then, since the function being minimized is convex (in general[4]) with respect to $\mathbf{s}$, $\widehat{\mathbf{s}}$ is found by looking for the zero gradient configuration, leading to

$$\widehat{\mathbf{s}} = \left[\mathbf{A}^{-1} + \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right]^{-1}\left(\mathbf{A}^{-1}\boldsymbol{\mu} + \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}\right). \tag{3.21}$$

Although seemingly different, Eqs. (3.21) and (3.18) are perfectly equivalent; observe that Eq. (3.21) can be seen as the multidimensional version of Eq. (1.85). It is also interesting to compare Eq. (3.21) with Eq. (1.92). Finally, notice that Eq. (3.21) suggests that it would be convenient to directly specify the inverse of the covariance $\mathbf{A}^{-1}$ rather than $\mathbf{A}$ itself. As we shall see, this "more convenient" nature of the inverse of the covariance matrix will emerge in several situations. The inverse covariance is known as the *concentration* or *precision* matrix [2], [107]; it is also called *potential* matrix in [78].

Another interesting appearance for this estimate is obtained with the help of the Mahalanobis distance (see Eq. (1.37)). Eq. (3.20) can be rewritten as

$$\widehat{\mathbf{s}} = \arg\min_{\mathbf{s}}\left\{\|\mathbf{x} - \mathbf{H}\mathbf{s}\|^2_{\mathbf{C}} + \|\mathbf{s} - \boldsymbol{\mu}\|^2_{\mathbf{A}}\right\}, \tag{3.22}$$

clearly showing that a compromise between prior information and data evidence is being sought. Similar criteria appear in non-statistical approaches where Eq. (3.10) is called an inverse problem (usually without explicit reference to the presence of noise); the best known of these approaches is Tikhonov's regularization theory [102]. Furthermore, notice that the ML estimate obtained by maximizing the likelihood in Eq. (3.11) with respect to $\mathbf{s}$ is

$$\widehat{\mathbf{s}}_{\mathrm{ML}} = \arg\min_{\mathbf{s}}\left\{\|\mathbf{x} - \mathbf{H}\mathbf{s}\|^2_{\mathbf{C}}\right\}. \tag{3.23}$$

Comparing Eqs. (3.23) and (3.22) makes it clear why criteria with the form of Eq. (3.22) are often (in non-Bayesian terms) known as *penalized* maximum likelihood estimators.

3.3.3 Particular Cases

Several particular situations deserve special attention and help to give a better understanding of the general case.

Independent and identically distributed noise: When the noise components are independent and identically distributed (often called *white*

[4] The function being minimized in Eq. (3.20) is not convex if $\left[\mathbf{A}^{-1} + \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right]$ is not positive-definite; i.e., if there are any vectors $\mathbf{v}$, such that $\|\mathbf{v}\|^2 > 0$, but $\mathbf{v}^T\left[\mathbf{A}^{-1} + \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right]\mathbf{v} = 0$. Such vectors belong to the *null-space* of the matrix $\left[\mathbf{A}^{-1} + \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right]$ [47]. Rewriting the condition, we get $\mathbf{v}^T\mathbf{A}^{-1}\mathbf{v} + \mathbf{v}^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\mathbf{v} = 0$, which is only possible for some $\mathbf{v}$ such that $\|\mathbf{v}\| > 0$ if the null spaces of $\mathbf{A}$ and $\mathbf{H}$ have a non-empty intersection. In most problems of interest, such as in image restoration, this is not the case.

noise), the covariance matrix of $\mathbf{N}$ becomes diagonal $\mathbf{C} = \sigma^2\mathbf{I}$ and its inverse is simply $\mathbf{C}^{-1} = \mathbf{I}/\sigma^2$. The first Mahalanobis norm in Eq. (3.22) becomes an Euclidean norm while Eq. (3.21) turns into

$$\widehat{\mathbf{s}} = \left(\sigma^2\mathbf{A}^{-1} + \mathbf{H}^T\mathbf{H}\right)^{-1}\left(\sigma^2\mathbf{A}^{-1}\boldsymbol{\mu} + \mathbf{H}^T\mathbf{x}\right). \tag{3.24}$$

No noise: The absence of noise corresponds to $\sigma^2 = 0$ which implies simplifying Eq. (3.24) into

$$\begin{aligned}\widehat{\mathbf{s}} &= \left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T\mathbf{x} & (3.25)\\ &= \arg\min_{\mathbf{s}}\left\{\|\mathbf{Hs} - \mathbf{x}\|^2\right\}, & (3.26)\end{aligned}$$

where $\left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T \equiv \mathbf{H}^\dagger$ is known as the Moore-Penrose *pseudo* (or *generalized*) *inverse* of matrix $\mathbf{H}$ (see, e.g., [7] or [21]). If $\mathbf{H}^{-1}$ exists, then $\mathbf{H}^\dagger = \mathbf{H}^{-1}$; if $\mathbf{H}$ is not invertible, then $\mathbf{H}^\dagger$ provides its *least-squares* sense pseudo-solution (see Eq. (3.26) and references [21], [7]). Notice that, due to the absence of noise, this estimate does not depend on the prior parameters $\mathbf{A}$ and $\boldsymbol{\mu}$ because the observed data is considered absolutely trustworthy. Finally, the similarity between Eqs. (3.26) and (3.23) shows that the least squares solution coincides with the maximum likelihood estimate when the noise components are assumed zero mean Gaussian, independent, and identically distributed (regardless of the variance). In the regression literature, this scenario is called least squares regression.

Prior covariance up to a factor: If the prior covariance matrix is written with a multiplicative factor, i.e., $\mathbf{A} = \phi^2\mathbf{B}$, then ϕ^2 can be seen as (proportional to) the "prior variance". Matrix $\mathbf{B}$ (as $\mathbf{A}$) is positive definite, thus having a unique symmetric square root $\mathbf{D}$ (in the matrix sense, *i.e.*, $\mathbf{DD} = \mathbf{D}^T\mathbf{D} = \mathbf{B}$); this allows rewriting Eq. (3.22), still with $\mathbf{C} = \sigma^2\mathbf{I}$ and using Euclidean norms, as

$$\widehat{\mathbf{s}} = \arg\min_{\mathbf{s}}\left\{\|\mathbf{x} - \mathbf{Hs}\|^2 + \frac{\sigma^2}{\phi^2}\|\mathbf{D}(\mathbf{s} - \boldsymbol{\mu})\|^2\right\}. \tag{3.27}$$

In regularization theory parlance, $\|\mathbf{D}(\mathbf{s} - \boldsymbol{\mu})\|^2$ is called the regularizing term, and σ^2/ϕ^2 the regularization parameter. Concerning Eq. (3.24), it becomes

$$\widehat{\mathbf{s}} = \left(\frac{\sigma^2}{\phi^2}\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{H}\right)^{-1}\left(\frac{\sigma^2}{\phi^2}\mathbf{B}^{-1}\boldsymbol{\mu} + \mathbf{H}^T\mathbf{x}\right), \tag{3.28}$$

which clearly reveals how the ratio σ^2/ϕ^2 controls the relative weight of the prior and the observed data. In a regression context, Eq. (3.28) is called a *ridge estimator*.

Flat prior: Letting the prior variance approach infinity, i.e., $\phi^2 \to \infty$, we obtain the non-informative prior; notice from Eq. (3.11) that $\mathbf{s}$ is clearly a (multidimensional) location parameter. From Eq. (3.28) it is clear that this is equivalent to taking σ^2 to zero and the corresponding estimate is again the one in Eq. (3.25). This is the *maximum likelihood* estimate, as can be confirmed by maximizing Eq. (3.11) with respect to s.

Flat prior, different noise variances: Finally, we consider the situation where we have flat prior ($\phi^2 \to \infty$), and where each observation is contaminated by noise with different variance, *i.e.*, $\mathbf{C} = \text{diag}(\sigma_1^2, ..., \sigma_n^2)$ (the so-called heteroscedastic model). In this case, the estimate is given by

$$\widehat{\mathbf{s}} = \left[\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right]^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}, \tag{3.29}$$

which is known as the *weighted least squares* (WLS) estimate. This designation stems from the fact that it is the solution of

$$\widehat{\mathbf{s}} = \arg\min_{\mathbf{s}} \sum_{i=1}^{n} \frac{\left((\mathbf{H}\,\mathbf{s})_i - x_i\right)^2}{2\sigma_i^2}, \tag{3.30}$$

which is a least squares estimation problem where each observation has its own weight.

3.4 Additive Loss Functions

3.4.1 General Result: Marginal Criteria

Let us now consider the loss functions that can be written as

$$L(\boldsymbol{s}, \mathbf{a}) = \sum_{i=1}^{d} L_i(s_i, a_i), \tag{3.31}$$

those we called *additive*. It will be assumed (without loss of generality, since arbitrary constants can be added to loss functions) that each individual loss function satisfies $L_i(s_i, a_i) \geq 0$. We start by considering the discrete case, i.e., a pure classification problem where $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_d$, with all the $\mathcal{S}_i$'s being discrete sets. Invoking the definition of optimal Bayes' estimator

(Eqs. (1.11) and (1.8)) and using the fact that the loss function is additive,

$$\begin{aligned}
\boldsymbol{\delta}(\mathbf{x}) &= \arg\min_{\mathbf{a}\in\mathcal{S}} \sum_{\boldsymbol{s}\in\mathcal{S}} \underbrace{\sum_{i=1}^{d} L_i(s_i,a_i)}_{L(\mathbf{s},\mathbf{a})}\; p_{\mathbf{S}}(\boldsymbol{s}|\mathbf{x}) && (3.32)\\
&= \arg\min_{\mathbf{a}\in\mathcal{S}} \sum_{i=1}^{d} \sum_{\boldsymbol{s}\in\mathcal{S}} L_i(s_i,a_i)\, p_{\mathbf{S}}(\boldsymbol{s}|\mathbf{x}) && (3.33)\\
&= \arg\min_{\mathbf{a}\in\mathcal{S}} \sum_{i=1}^{d} \sum_{s'\in\mathcal{S}_i} L_i(s',a_i) \sum_{\boldsymbol{s}\in\mathcal{S}:s_i=s'} p_{\mathbf{S}}(\boldsymbol{s}|\mathbf{x}). && (3.34)
\end{aligned}$$

In Eq. (3.34), the notation $\boldsymbol{s}\in\mathcal{S}:s_i=s'$ indicates that the corresponding summation sweeps all the configurations $\mathbf{s}\in\mathcal{S}$ such that the i-th component equals a certain value, $s_i = s'$. Of course, this is simply the posterior marginal probability function, i.e.,

$$\sum_{\boldsymbol{s}\in\mathcal{S}:s_i=s} p_{\mathbf{S}}(\boldsymbol{s}|\mathbf{x}) = p_{S_i}(s|\mathbf{x}); \qquad (3.35)$$

then, we can write

$$\begin{aligned}
\boldsymbol{\delta}(\mathbf{x}) &= \arg\min_{\mathbf{a}\in\mathcal{S}} \sum_{i=1}^{d} \sum_{s_i\in\mathcal{S}_i} L_i(s_i,a_i)\, p_{S_i}(s_i|\mathbf{x}) && (3.36)\\
&= \Bigg[\arg\min_{a_1\in\mathcal{S}_1} \sum_{s_1\in\mathcal{S}_1} L_1(s_1,a_1)\, p_{S_1}(s_1|\mathbf{x}),\ldots,\\
&\qquad \arg\min_{a_d\in\mathcal{S}_d} \sum_{s_d\in\mathcal{S}_d} L_d(s_d,a_d)\, p_{S_d}(s_d|\mathbf{x})\Bigg]^T\\
&= \Bigg[\arg\min_{a_1\in\mathcal{S}_1} E_{S_1}\left[L_1(s_1,a_1)|\mathbf{x}\right],\ldots,\\
&\qquad \arg\min_{a_d\in\mathcal{S}_d} E_{S_d}\left[L_d(s_d,a_d)|\mathbf{x}\right]\Bigg]^T && (3.37)
\end{aligned}$$

because minimizing a sum of independent (each term of the outer summation in Eq. (3.36) is only a function of one a_i) non-negative (recall that $L_i(s_i,a_i)\geq 0$) functions is equivalent to minimizing each one individually. This important result, first brought to the attention of the computer vision/image analysis communities by Marroquin [74], simply states that in the presence of a loss function that is the sum of individual (non-negative) loss functions, the optimal Bayes' rule consists of a (parallel) set of non-interacting Bayes' rules, each relative to the corresponding *a posteriori* marginal probability function. Each individual term $E_{S_i}[L_i(s_i,a_i)|\mathbf{x}]$ can be called the *a posteriori marginal expected loss*. In the continuous case, an equivalent result can be derived by replacing the summations over the $\mathcal{S}_i$'s by the appropriate integrals. The general result in Eq. (3.37) can now be particularized to the common loss functions considered in Chapter 2.

3.4.2 Application to the common loss functions

Quadratic Error Loss Function

For the quadratic loss function, there is already the result in Eq. (3.8), which also covers the additive case (obtained when matrix $\mathbf{Q}$ is chosen to be diagonal). This is a simple consequence of the fact that the mean of a random vector is computed separately with respect to each marginal; it is an intrinsically marginal concept. If the sets $\mathcal{S}_i$ are discrete, then the resulting estimates will have to be *thresholded* (as in Eq. (1.100)), leading to what are called the *thresholded posterior marginal means* (TPMM).

Absolute Error Loss Function

The compound version of the absolute error loss function (with $\mathcal{A} = \mathcal{S} = I\!R^n$) is naturally additive and defined as

$$L(\boldsymbol{s}, \mathbf{a}) = \sum_{i=1}^{d} |s_i - a_i|; \tag{3.38}$$

Then, the optimal Bayes' estimator of each s_i is the median of the respective posterior marginal, called *median of posterior marginal density* (MPMD).

"0/1" Loss Function

For estimation problems, the additive version of the "0/1" loss function is written as Eq. (3.2) with each $L_i(\cdot, \cdot)$ being a "0/1" loss function of scalar argument. Concerning classification problems, the additive "0/1" loss function is called *total error* loss function because it equals the number of incorrect classifications. In both cases, according to the general result for additive loss functions, the optimal Bayes' decision rule is

$$\boldsymbol{\delta}_{\text{MPM}}(\mathbf{x}) = \left[\arg \max_{s_1 \in \mathcal{S}_1} p_{S_1}(s_1|\mathbf{x}), \ldots, \arg \max_{s_N \in \mathcal{S}_d} p_{S_d}(s_d|\mathbf{x})\right]^T. \tag{3.39}$$

This corresponds to the MAP criterion applied to each posterior marginal, and was designated in [74] as the *maximizer of posterior marginals* (MPM).

Example 3.4.1

This example illustrates, with a toy problem, the difference between the MPM criterion just presented and the joint MAP in Eq. (3.4). Let

$\mathcal{S} = \{0,1\}^2$, that is, the unknown is a pair of binary variables $\mathbf{s} = (s_1, s_2)$. Prior information states that it is more probable that $s_1 = s_2$ than not; specifically, $p_{\mathbf{S}}(0,0) = p_{\mathbf{S}}(1,1) = 0.4$ and $p_{\mathbf{S}}(0,1) = p_{\mathbf{S}}(1,0) = 0.1$. Observations $\mathbf{x} = (x_1, x_2)$ are obtained according to a very simple Gaussian likelihood, with the conditional independence property:

$$f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}) = \mathcal{N}(x_1|s_1, \sigma^2)\mathcal{N}(x_2|s_2, \sigma^2)$$

For the joint MAP criterion,

$$(\widehat{s}_{1\ \mathrm{MAP}}, \widehat{s}_{2\ \mathrm{MAP}}) = \arg \max_{(s_1,s_2)\in\{0,1\}^2} p_{\mathbf{S}}(s_1, s_2|x_1, x_2),$$

with $p_{\mathbf{S}}(\mathbf{s}|\mathbf{x}) \propto p_{\mathbf{S}}(\mathbf{s}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\mathbf{s})$. Actually, this is a quaternary classification problem, of the type studied in Section 1.4.3; in the notation there used, $\boldsymbol{\mu}_1 = [0,0]^T$, $\boldsymbol{\mu}_2 = [0,1]^T$, $\boldsymbol{\mu}_3 = [1,0]^T$, $\boldsymbol{\mu}_4 = [1,1]^T$, and $\mathbf{C}_i = \sigma^2\mathbf{I}$, for $i = 1,2,3,4$.

For the MPM criterion, the *a posteriori* marginals have to be separately maximized:

$$\begin{aligned} \widehat{s}_{1\ \mathrm{MPM}} &= \arg \max_{s_1=0,1} \left\{ \mathcal{N}(x_1|s_1, \sigma^2) \sum_{s_2=0,1} \mathcal{N}(x_2|s_2, \sigma^2) p_{\mathbf{S}}(s_1, s_2) \right\} \\ \widehat{s}_{2\ \mathrm{MPM}} &= \arg \max_{s_2=0,1} \left\{ \mathcal{N}(x_2|s_2, \sigma^2) \sum_{s_1=0,1} \mathcal{N}(x_1|s_1, \sigma^2) p_{\mathbf{S}}(s_1, s_2) \right\} \end{aligned}$$

To study the difference between the two criteria, we generated 2000 sample from $p_{\mathbf{S}}(\mathbf{s})$ and the corresponding observations according to the model described. Those observations were then used by the two criteria and the following results were collected: number of correct patterns (i.e., $\widehat{\mathbf{s}} = \mathbf{s}$), denoted n_0, number of patterns with one error (i.e., $\widehat{\mathbf{s}}$ and $\mathbf{s}$ differing in exactly one position), denoted n_1, and number of patterns with two errors (i.e., $\widehat{s}_1 \neq s_1$ and $\widehat{s}_2 \neq s_2$), denoted n_2. Another interesting number is $n_1 + 2\,n_2$ which counts the total number of wrong "bits".

For $\sigma^2 = 0.2$, we have a low-noise situation and both criteria give similar results:

MAP			
$n_0 = 1975$	$n_1 = 25$	$n_2 = 0$	$n_1 + 2\,n_2 = 25$

MPM			
$n_0 = 1976$	$n_1 = 24$	$n_2 = 0$	$n_1 + 2\,n_2 = 24$

When the noise increases to $\sigma^2 = 0.8$, differences between MPM and MAP arise:

MAP			
$n_0 = 1334$	$n_1 = 385$	$n_2 = 281$	$n_1 + 2\,n_2 = 947$

MPM			
$n_0 = 1281$	$n_1 = 503$	$n_2 = 216$	$n_1 + 2\,n_2 = 935$

The MAP criterion yields a larger number of fully correct patterns (1334 versus 1281), which is a natural result since the underlying loss function "insists" in that the estimate be exactly equal to the true $\mathbf{s}$. However, the MAP criterion produces more estimates that are completely wrong (error in both "bits") than MPM does (281 versus 216); this is again a natural consequence of the underlying loss functions. The additive "0/1" loss counts the number of wrong positions, while the non-additive one, once a position is wrong, does not care about the other. A final consequence of this effect is that MPM achieves a smaller number of "bit" errors (935 versus 947). This is a natural result, because of the additive nature of the loss function underlying the MPM criterion.

Reference [74] contains other examples where MPM is compared against MAP in the context of image restoration/reconstruction problems.

End of Example 3.4.1

Example 3.4.2

Another interesting example of the difference between the MPM and MAP concerns the joint estimation of the mean and variance of a set of Gaussian observations. Consider n real-valued observations $\mathbf{x} = \{x_1, ...x_n\}$ obtained according to the likelihood function

$$f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}.$$

Let us concentrate on maximum likelihood estimates by adopting flat priors for both parameters: $p(\mu, \sigma^2) \propto$ "constant". Accordingly, the *a posteriori* probability density function is

$$p(\mu, \sigma^2|\mathbf{x}) \propto f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2)$$

which, despite the use of improper priors, turns out to be integrable. The MAP/ML estimate of the unknown parameters yields

$$\begin{aligned} \left(\widehat{\mu}_{\mathrm{MAP}}, \widehat{\sigma^2}_{\mathrm{MAP}}\right) &= \arg\max_{\mu,\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \widehat{\mu}_{\mathrm{MAP}}\right)^2\right). \qquad (3.40) \end{aligned}$$

It is well known that this variance estimate is biased (see section 2.7.1, for the definition of an unbiased estimate). In fact, it easy to check that

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\frac{1}{n}\sum_{j=1}^{n}x_i\right)^2\middle|\mu,\sigma^2\right]=\frac{n-1}{n}\sigma^2,$$

with the unbiased estimate thus being

$$\widehat{\sigma^2}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\widehat{\mu}_{\text{MAP}})^2=\frac{n}{n-1}\,\widehat{\sigma^2}_{\text{MAP}}$$

To obtain the MPM estimates of μ and σ^2 we need to compute the corresponding posterior marginals.

$$\begin{aligned} p(\mu|\mathbf{x}) &= \int_0^\infty p(\mu,\sigma^2|\mathbf{x})\,d\sigma^2 \propto \left(n\,\widehat{\sigma^2}_{\text{MAP}}+n\,(\mu-\widehat{\mu}_{\text{MAP}})^2\right)^{\frac{2-n}{2}}\\ p(\sigma^2|\mathbf{x}) &= \int_{-\infty}^\infty p(\mu,\sigma^2|\mathbf{x})\,d\mu \propto (2\pi\sigma^2)^{-\frac{n-1}{2}}\exp\left\{-\frac{n}{2\sigma^2}\widehat{\sigma^2}_{\text{MAP}}\right\} \end{aligned}$$

Maximization of these marginals leads to

$$\widehat{\mu}_{\text{MPM}} = \widehat{\mu}_{\text{MAP}} \tag{3.41}$$

$$\widehat{\sigma^2}_{\text{MPM}} = \frac{n}{n-1}\,\widehat{\sigma^2}_{\text{MAP}}; \tag{3.42}$$

that is, the mean estimate is the same as the one obtained under the MAP criterion, while the variance estimate appears with correction needed to make it unbiased. Of course, the difference between the two criteria vanishes when the number of data points increases, but for small samples this may be a significant effect.

End of Example 3.4.2

3.5 Priors for Compound Problems

Most of Chapter 2 was devoted to studying techniques for building priors in a formal way. From a conceptual point of view, most of what we saw there carries over unchanged to the compound inference scenario; usually, only computational and/or analytical difficulties will emerge. In this section, we will revisit each topic covered in Chapter 2 concerning prior building (improper priors, exponential families, conjugate priors, non-informative priors, and maximum entropy priors), examining the impact of the multivariate nature of the unknown. Minimum description length (MDL) priors will be considered in the next chapter where the problem of estimating parameter vectors of unknown dimension is addressed.

3.5.1 Improper Priors and Maximum Likelihood Inference

Just as for an univariate s, a (joint) prior $p_{\mathbf{S}}(\mathbf{s})$ for a multivariate unknown $\mathbf{s}$ is called improper if the corresponding integral fails to converge, i.e., if

$$\int \cdots \int p_{\mathbf{S}}(s_1, ..., s_d)\, ds_1 \cdots ds_d = \infty,$$

or a similar property involving summations in the discrete case. As for a scalar unknown, as long as the *a posteriori* probability function is well defined, valid Bayesian rules can still be obtained.

However, a new issue arises which was not present in univariate problems: a prior can be proper for some aspects of the unknown vector and improper for others. To clarify this statement, let us consider a very simple example:

Example 3.5.1

Consider an estimation problem where $\mathcal{S} = I\!\!R^2$, $\mathbf{s} = [s_1,\ s_2]^T$; the observed vector is also bidimensional $\mathbf{x} = [x_1,\ x_2]^T$ and the likelihood function is simply $f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}) = \mathcal{N}(\mathbf{x}|\mathbf{s}, \sigma^2\mathbf{I})$. This problem is in the class considered in Section 3.3 with $\mathbf{H} = \mathbf{I}$ and $\mathbf{C} = \sigma^2\mathbf{I}$. Now, consider the prior

$$p_{\mathbf{S}}(s_1, s_2) \propto \exp\left\{-\frac{(s_1 - s_2)^2}{2\phi^2}\right\} = \exp\left\{\frac{-1}{2\phi^2}\,\mathbf{s}^T \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \mathbf{s}\right\};$$

this prior can be used to express the knowledge that s_1 and s_2 are expected to be "close" to each other, although it clearly says nothing about "where" they are expected to be (because $p_{\mathbf{S}}(s_1, s_2) = p_{\mathbf{S}}(s_1 + k, s_2 + k)$, for any k). Notice that this is a Gaussian prior, although of a particular kind since the corresponding covariance matrix does not exist. Nevertheless, we can still obtain the *a posteriori* joint probability density function which is

$$p_{\mathbf{S}}(\mathbf{s}|\mathbf{x}) = \mathcal{N}\left(\mathbf{s}|\widehat{\mathbf{s}}, \mathbf{P}\right)$$

where, letting $\alpha = \sigma^2/\phi^2$,

$$\widehat{\mathbf{s}} = \begin{bmatrix} \widehat{s_1} \\ \widehat{s_2} \end{bmatrix} = \frac{1}{1 + 2\alpha} \begin{bmatrix} 1 + \alpha & \alpha \\ \alpha & 1 + \alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The key aspect of this example is the following. The average of the two estimates,

$$\frac{\widehat{s_1} + \widehat{s_2}}{2} = \frac{(1 + \alpha)x_1 + \alpha x_2 + \alpha x_1 + (1 + \alpha)x_2}{2(1 + 2\alpha)} = \frac{x_1 + x_2}{2}$$

is the same as the average of the observations. However, the difference

$$\widehat{s_1} - \widehat{s_2} = \frac{(1 + \alpha)x_1 + \alpha x_2 - \alpha x_1 - (1 + \alpha)x_2}{(1 + 2\alpha)} = \frac{x_1 - x_2}{1 + 2\alpha}$$

is smaller than the difference between the observations $(x_1 - x_2)$, because $\alpha = \sigma^2/\phi^2 \geq 0$ and thus $(1 + 2\alpha) > 1$. These facts are in accordance with the previous interpretation of the prior: it is flat with respect to the location of the unknowns, but informative with respect to their relative difference.

Another way to look at this prior is to notice that the matrix that appears in the exponent as a zero eigenvalue whose corresponding eigenvector is (any vector proportional to) $[1\ 1]^T$; its other eigenvalue is 2 associated with eigenvectors of the form $[-1\ 1]^T$. In other words, this prior, behaves as having infinite variance along the direction of $[1\ 1]^T$, yielding a maximum likelihood estimate, and finite variance along the orthogonal direction. Of course, this example can be generalized to dimensions higher than two, as we shall see concerning Gauss Markov random fields.

From a technical point of view, as we mentioned concerning the scalar case, such a prior can be viewed as the limit of a family of proper priors. For example, we could have started with

$$p_{\mathbf{S}}(s_1, s_2) \propto \exp\left\{ \frac{-1}{2\phi^2}\, \mathbf{s}^T \left[\begin{array}{cc} 1+\varepsilon & -1 \\ -1 & 1+\varepsilon \end{array} \right] \mathbf{s} \right\},$$

for $\varepsilon > 0$, which is a perfectly proper Gaussian density with a valid covariance matrix (as can be seen by inverting the matrix in the exponent). We could have done all the derivations above keeping $\varepsilon > 0$, and in the end found limits for $\varepsilon \to 0$; our conclusions would, of course, have been the same. For example, the eigenvalues of this matrix are ε and $2 + \varepsilon$, with the same eigenvectors as above.

End of Example 3.5.1

3.5.2 Exponential Families and Conjugate Priors

The concept of conjugate prior can be carried absolutely unchanged from the univariate to the compound scenario. Only a slight generalization of notation appears in the following definition:
Let $\mathcal{F} = \{f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}),\ \mathbf{s} \in \mathcal{S}\}$ be a class of likelihood functions; let $\mathcal{P}$ be a class (set) of probability (density or mass) functions; if, for any $\mathbf{x}$, any $p_{\mathbf{S}}(\mathbf{s}) \in \mathcal{P}$, and any $f_{\mathbf{X}}(\mathbf{x}|\mathbf{s}) \in \mathcal{F}$, the resulting a posteriori (joint) probability function $p_{\mathbf{S}}(\mathbf{s}|\mathbf{x}) \propto f_{\mathbf{X}}(\mathbf{x}|\mathbf{s})\, p_{\mathbf{S}}(\mathbf{s})$ is still in $\mathcal{P}$, then $\mathcal{P}$ is a conjugate family (a family of conjugate priors), for $\mathcal{F}$.

For example, the multivariate Gaussian priors studied in Section 3.3 are conjugate priors for Gaussian observation models with respect to a multidimensional location parameter (mean).

As we have seen in Section 2.12, deriving conjugate priors for likelihoods in an exponential family is a straightforward task. The concept of exponential families in the case of multidimensional parameters is also a simple generalization (mostly of notational nature) of the unidimensional case. Specifically, a parameterized family of probability functions which can be written as

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = \phi(\mathbf{x})\,\psi(\boldsymbol{\theta})\,\exp\{\boldsymbol{\xi}(\boldsymbol{\theta})^T\mathbf{t}(\mathbf{x})\} \tag{3.43}$$

where $\mathbf{t}(\mathbf{x})$ is a k-dimensional sufficient statistic and $\boldsymbol{\xi}(\boldsymbol{\theta})$ is a k-dimensional function of the parameter vector $\boldsymbol{\theta}$; $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ is called an exponential family [19], [93]. The role of $\psi(\boldsymbol{\theta})$ is to ensure that each member of the family is adequately normalized. As in the scalar parameter case, $\boldsymbol{\xi}(\boldsymbol{\theta})$ and $\mathbf{t}(\mathbf{x})$ are called the *natural* (or *canonical*) parameter vector and sufficient statistic, respectively. The change of variables $\mathbf{t} = \mathbf{t}(\mathbf{x})$ and the reparameterization $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\theta})$ allow casting an exponential family into its *natural form*, shown in Eq. (2.134).

As in the unidimensional case, in the presence of a set of i.i.d. observations $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, the resulting joint likelihood belongs to an exponential family with the same natural parameter. The (joint) natural sufficient statistic is the summation of the individual ones,

$$\mathbf{t}(\mathbf{x}) = \mathbf{t}(\mathbf{x}_1) + \mathbf{t}(\mathbf{x}_2) + ... + \mathbf{t}(\mathbf{x}_n), \tag{3.44}$$

which is still k-dimensional. The importance of this property is that it decouples the dimension of the sufficient statistic from the sample size. The Pitman-Koopman theorem, referred in Section 2.12, is of course still valid for multidimensional parameters.

Recall that a natural exponential family (regardless of the dimensionality of its parameter) can be written as in Eq. (2.138)), explicitly revealing the normalizing constant (partition function) $Z(\boldsymbol{\xi}) = 1/\psi(\boldsymbol{\xi})$. As referred in Section 2.12, this has the important consequence that means, variances, covariances, or any other moments of the natural statistics can be computed directly from the derivatives of $\log Z(\boldsymbol{\xi})$.

We will now look at some representative exponential families with multidimensional parameters, and corresponding conjugate priors.

Example 3.5.2

The Bernoulli and binomial distributions (which we have considered in several examples in the previous chapter) are used to model situations where each trial has one of two possible outcomes; the classical example is coin tossing. The *multinomial* distribution generalizes the *binomial* to cases where each observation is one of a set of k possible outcomes; for example, dice tossing, where $k = 6$.

Let us consider a sequence of m tosses of a die; each outcome is denoted as x_i and belongs to $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, so $k = 6$. The die is characterized by the probabilities of each of its 6 faces $\theta_1,\ \theta_2, ..., \theta_6$; of course, $\theta_1 + \theta_2 + ... + \theta_6 = 1$ must hold, which is equivalent to saying that there are only 5 free parameters, e.g., $\boldsymbol{\theta} = (\theta_1,\ \theta_2, ..., \theta_5)$, with $\theta_6 = 1-\theta_1-\theta_2-\theta_3-\theta_4-\theta_5$. Now, let n_i be the number of face i outcomes observed in some sequence of m tosses; the probability of a particular configuration $\mathbf{n} = (n_1,\ n_2, ..., n_6)$

(of course with $n_1 + n_2 + ... + n_6 = m$, so $n_6 = m - n_1 - \cdots - n_5$) is given by the multinomial probability function

$$f_{\mathbf{N}}(\mathbf{n}|\boldsymbol{\theta}) = \left(\begin{array}{c} m \\ n_1 \; n_2 \; ... \; n_6 \end{array} \right) \prod_{i=1}^{6} \theta_i^{n_i}, \tag{3.45}$$

where the *multinomial coefficients* are given by

$$\left(\begin{array}{c} m \\ n_1 \; n_2 \; ... \; n_6 \end{array} \right) = \frac{m!}{n_1! n_2! \cdots n_6!}.$$

Notice that the particular case where $k = 2$ is precisely the binomial distribution.

The multinomial can be written in exponential family form as

$$f_{\mathbf{N}}(\mathbf{n}|\boldsymbol{\theta}) = \underbrace{\left(\begin{array}{c} m \\ n_1 ... n_6 \end{array} \right)}_{\phi(\mathbf{n})} \underbrace{(1 - \theta_1 - \ldots - \theta_5)^m}_{\psi(\boldsymbol{\theta})^m} \exp \left\{ \sum_{i=1}^{5} n_i \log \frac{\theta_i}{1 - \theta_1 - \cdots - \theta_5} \right\}$$

revealing that the natural parameter is

$$\begin{aligned} \boldsymbol{\xi}(\boldsymbol{\theta}) &= \left[\log \frac{\theta_1}{1 - \theta_1 - \cdots - \theta_5}, \ldots, \log \frac{\theta_5}{1 - \theta_1 - \cdots - \theta_5} \right] \\ &= [\xi_1, \, \xi_2, \, \xi_3, \, \xi_4, \, \xi_5]^T \end{aligned} \tag{3.46}$$

and the natural sufficient statistic is $\mathbf{t}(\mathbf{n}) = [n_1, \, n_2, \, n_3, \, n_4, \, n_5]^T$. We can conclude that multinomial distribution with k possible outcomes constitute a $k - 1$ dimensional exponential family (a natural consequence of the fact that there are only $k - 1$ degrees of freedom).

As we saw in Section 2.12, it is now possible to write conjugate priors as

$$p_\Theta(\boldsymbol{\theta}) \propto \psi(\boldsymbol{\theta})^\nu \exp \left\{ \boldsymbol{\xi}(\boldsymbol{\theta})^T \boldsymbol{\gamma} \right\}. \tag{3.47}$$

In the current case, we have

$$\begin{aligned} p_\Theta(\boldsymbol{\theta}) &\propto (1 - \theta_1 - \ldots - \theta_5)^\nu \exp \left\{ \sum_{i=1}^{5} \gamma_i \log \frac{\theta_i}{1 - \theta_1 - \cdots - \theta_5} \right\} \\ &\propto (1 - \theta_1 - \cdots - \theta_5)^{\nu - \gamma_1 - \ldots - \gamma_5} \prod_{i=1}^{5} \theta_i^{\gamma_i}. \end{aligned} \tag{3.48}$$

This is called a 6−dimensional Dirichlet distribution; usually it is written in a slightly different way (with $\gamma_6 = \nu - \gamma_1 - ... - \gamma_5$ and $\alpha_i = \gamma_i + 1$), now including the normalization factor, as

$$\mathcal{D}_6(\theta_1, ..., \theta_6 | \alpha_1, ..., \alpha_6) = \frac{\Gamma(\alpha_1 + ... + \alpha_6)}{\prod_{i=1}^{6} \Gamma(\alpha_i)} \prod_{i=1}^{6} \theta_i^{\alpha_i - 1} \mathbf{1}_{\left(\sum_{i=1}^{6} \theta_i = 1\right)}$$

where $\mathbf{1}_A$ is an *indicator function*, i.e., it is equal to one if A is true, and equal to zero if A is false. Such a prior has the clear interpretation of additional data: $\nu = \gamma_1 + ... + \gamma_6$ additional trials, with γ_1 face 1 outcomes, γ_2 face 2 outcomes, and so on. Notice that, naturally, a $\mathcal{D}_2(\theta_1, \theta_2|\alpha_1, \alpha_2)$ density coincides with a beta $\text{Be}(\theta|\alpha_1, \alpha_2)$ prior.

Finally, the resulting *a posteriori* probability density function is again Dirichlet,

$$p_\Theta(\boldsymbol{\theta}|\mathbf{n}) = \mathcal{D}_6(\theta_1, ..., \theta_6|n_1 + \alpha_1, ..., n_6 + \alpha_6). \tag{3.49}$$

The MPM and PM estimates are simple functions of the parameters of the *a posteriori* density (see, e.g., [46]):

$$\widehat{\theta}_{i\text{PM}} = \frac{\alpha_i + n_i}{\sum_{j=1}^{6} \alpha_j + n_j} = \frac{\alpha_i + n_i}{m + \sum_{j=1}^{6} \alpha_j}, \tag{3.50}$$

$$\widehat{\theta}_{i\text{MPM}} = \frac{\alpha_i + n_i - 1}{m - 6 + \sum_{j=1}^{6} \alpha_j}. \tag{3.51}$$

Notice that in the limit (infinitely long sequence of tosses), these estimates converge to those dictated by the ML criterion, which are simply $\widehat{\theta}_{i\text{ML}} = n_i/m$, for $i = 1, .., 6$.

End of Example 3.5.2

In the following set of examples we will focus on the important and ubiquitous Gaussian likelihood. We will start by considering the case of univariate Gaussian likelihoods where both the mean and variance are unknown parameters; after that, the multivariate Gaussian will be addressed.

Example 3.5.3

In Section 2.12 we only considered two separate cases: unknown mean but known variance, and unknown variance but known mean. In this example we will look at the more general situation where both the mean μ and variance σ^2 are unknown, i.e., $\boldsymbol{\theta} = [\mu,\ \sigma^2]^T$. If $\mathbf{x} = (x_1, ..., x_n)$ is a sequence of i.i.d. observations such that $f_{X_i}(x_i|\mu, \sigma^2) = \mathcal{N}(x_i|\mu, \sigma^2)$,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right\} \\ &= \underbrace{\left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\}}_{\psi(\boldsymbol{\theta})^n} \exp\left\{\frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2\right\} \end{aligned}$$

showing that $\boldsymbol{\xi} = [\mu/\sigma^2,\ -1/(2\sigma^2)]^T$ is a natural parameterization corresponding to the sufficient statistic

$$\mathbf{t}(\mathbf{x}) = \left[\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} x_i^2 \right]^T .$$

The standard approach to obtaining a conjugate family for this class of likelihoods can now be adopted:

$$p_{\Theta}(\boldsymbol{\theta}|\nu, \gamma_1, \gamma_2) = \psi(\boldsymbol{\theta})^{\nu} \exp\left\{\gamma_1 \mu/\sigma^2 - \gamma_2/(2\sigma^2)\right\}.$$

Interestingly, this joint prior can be decomposed as (after changing into a more meaningful parameterization; see, e.g., [46]),

$$p(\mu, \sigma^2|\nu_0, \mu_0, \sigma_0^2, \kappa_0) = p(\mu|\mu_0, \sigma^2, \kappa_0)p(\sigma^2|\nu_0, \sigma_0^2)$$

where

$$p(\mu|\mu_0, \sigma^2, \kappa_0) \quad \propto \quad \mathcal{N}\left(\mu|\mu_0, \frac{\sigma^2}{\kappa_0}\right) \tag{3.52}$$

$$p(\sigma^2|\nu_0, \sigma_0^2) \quad \propto \quad (\sigma^2)^{-(\nu_0/2+1)} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}. \tag{3.53}$$

The prior on μ depends on σ^2; this is a natural feature since a high σ^2 value suggests a high variance on the prior for μ [46]. Densities with the form of Eq. (3.53) are called inverse-gamma, and are conjugate priors with respect to Gaussian likelihoods of unknown variance. The *a posteriori* joint density of μ and σ^2 is still (due to conjugacy) the product of a Gaussian with respect to μ and an inverse-gamma for σ^2, from which the estimates can be explicitly obtained (see [46] for more details).

End of Example 3.5.3

Example 3.5.4

The multivariate Gaussian probability density function plays a central role in many areas of applied statistics, this being particularly true in statistical signal and image analysis. In Section 3.3, we studied the problem of estimating an unknown mean of a multivariate Gaussian likelihood with known covariance matrix (in fact we studied a more general problem, but this simpler case can be obtained with $\mathbf{H} = \mathbf{I}$). We saw there that multivariate Gaussian priors are conjugate for that class of likelihoods. In this example we will look at a scenario where the unknown is the covariance matrix. Consider n i.i.d. vector observations $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ drawn from a d-dimensional Gaussian density of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}$

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{A}) = (2\pi)^{-\frac{n\,d}{2}} |\mathbf{A}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{A}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$$

where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. The key step to writing this likelihood in exponential family form is the well known equality for quadratic forms, $\mathbf{v}^T\mathbf{M}\mathbf{v} = \text{tra}(\mathbf{M}\mathbf{v}\mathbf{v}^T)$, where $\text{tra}(\cdot)$ denotes the trace operator (the sum of the elements in the main diagonal). With this in hand,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{A}) \;\; \propto \;\; & \sqrt{|\mathbf{A}^{-1}|}\exp\{-\frac{n}{2}\boldsymbol{\mu}^T\mathbf{A}^{-1}\boldsymbol{\mu}\} \\ & \exp\left\{\boldsymbol{\mu}^T\mathbf{A}^{-1}\sum_{i=1}^{n}\mathbf{x}_i - \frac{1}{2}\text{tra}\left(\mathbf{A}^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)\right\} \end{aligned} \tag{3.54}$$

Notice that the argument of the second exponential is linear on the elements of the inverse of the covariance matrix $\mathbf{A}^{-1}$ (the concentration or precision matrix, see Section 3.3.2), but not with respect to the elements of $\mathbf{A}$ itself. This shows that the natural parameter is $\mathbf{A}^{-1}$ rather than $\mathbf{A}$ (see the discussion in the paragraph following Eq. (3.21)). The other natural parameter is $\boldsymbol{\mu}^T\mathbf{A}^{-1}$. Finally, the corresponding vector of natural sufficient statistics is

$$\mathbf{t}(\mathbf{x}) = \left[\sum_{i=1}^{n}\mathbf{x}_i, \; \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right]^T .$$

We already saw that a Gaussian prior on $\boldsymbol{\mu}$ is a conjugate prior for that parameter. Let us now focus on the prior for $\mathbf{B} \equiv \mathbf{A}^{-1}$, by taking $\boldsymbol{\mu} = 0$. Using the standard approach, we obtain (where $\boldsymbol{\Gamma}$ has to be a positive definite $d \times d$ matrix)

$$p(\mathbf{B}|\nu, \boldsymbol{\Gamma}) \propto \left(\sqrt{|\mathbf{B}|}\right)^{\nu}\exp\left\{-\frac{1}{2}\text{tra}\left(\mathbf{B}\boldsymbol{\Gamma}\right)\right\}, \tag{3.55}$$

for $\mathbf{B}$ positive definite, and zero elsewhere, which is known as a Wishart distribution; actually, in a more standard notation, Eq. (3.55) would be called a Wishart $\mathcal{W}_d\left(\nu + d + 1, \boldsymbol{\Gamma}\right)$ density; a detailed study of this multivariate density can be found in [2].

The resulting a posteriori density is of course again Wishart; specifically, it is $\mathcal{W}_d\left(\nu + n + d + 1, \left(\boldsymbol{\Gamma} + \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)\right)$. The PM estimate obtained from this *a posteriori* density is the mean of the Wishart density (see [2]), which is

$$\widehat{\mathbf{B}} = (\nu + n)\left(\boldsymbol{\Gamma} + \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}$$

explicitly revealing the usual "additional data" nature of the conjugate prior.

End of Example 3.5.4

3.5.3 Non-informative Jeffreys' Priors

When estimating a single parameter, we saw in Section 2.6 that non-informative priors allow formalizing the concept of "ignorance" by relying on invariance arguments. In the compound case, there may be different invariance criteria involved since the parameters may play different kinds of roles; e.g., when estimating the mean and the standard deviation from a set of Gaussian observations, one of the unknowns is a location parameter while the other is a scale parameter. Another important new question arising when dealing with sets of unknowns concerns the choice of a dependence structure (among the variables) expressing prior ignorance. This point raises even more controversy than non-informative priors for single parameters. For example, in the multidimensional case, Jeffreys' priors may lead to incoherent criteria (however, it should be kept in mind that Jeffreys proposed his approach mainly for the one-dimensional case) [14], [93]. As stressed in [14]: *"In continuous multiparameter situations there is no hope for a single, unique, non-informative prior, appropriate for all the inference problems within a given model"*. We conclude this introductory paragraph by pointing out that a very comprehensive catalog of non-informative priors, for both one-dimensional and multidimensional parameters, is available in [110].

For multi-parameter likelihood functions, the Fisher information (see Section 2.7) is a matrix function of the vector parameter $\boldsymbol{\theta}$, which we will still denote as $\mathcal{I}(\boldsymbol{\theta})$. Its definition is the natural generalization of Eq. (2.59); accordingly, each element $\mathcal{I}_{ij}(\boldsymbol{\theta})$ is given by

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j}\right]. \tag{3.56}$$

The Fisher information matrix also verifies the linear dependence on the size of the data set (for i.i.d. observations) studied in Section 2.7 (see Eq. (2.69)).

There is, of course, a multiparameter version of the Cramer-Rao bound referred in Section 2.7. If $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ is an unbiased estimator of an unknown parameter vector $\boldsymbol{\theta}$, that is, if

$$E_{\mathbf{X}}\left[\widehat{\boldsymbol{\theta}}(\mathbf{x})\middle|\,\boldsymbol{\theta}\right] = \int\cdots\int \widehat{\boldsymbol{\theta}}(\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} = \boldsymbol{\theta}, \tag{3.57}$$

then, the covariance matrix of the estimator verifies

$$E_{\mathbf{X}}\left[\widehat{\boldsymbol{\theta}}(\mathbf{x})\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right)^T\middle|\,\boldsymbol{\theta}\right] - \mathcal{I}^{-1}(\boldsymbol{\theta}) \geq 0 \tag{3.58}$$

where "≥ 0" here means "is a positive semidefinite matrix". In particular, since the diagonal elements of a positive semidefinite matrix can not be negative,

$$E_{\mathbf{X}}\left[\left(\widehat{\theta}_i(\mathbf{x}) - \theta_i\right)^2\middle|\,\boldsymbol{\theta}\right] \geq \left[\mathcal{I}^{-1}(\boldsymbol{\theta})\right]_{ii}. \tag{3.59}$$

In Eq. (3.59), $\widehat{\theta}_i(\mathbf{x})$ and θ_i denote the i-th components of vectors $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ and $\boldsymbol{\theta}$, respectively, while $[\mathcal{I}^{-1}(\boldsymbol{\theta})]_{ii}$ stands for element ii of $\mathcal{I}^{-1}(\boldsymbol{\theta})$, the inverse of Fisher matrix.

The same line of thought described in Section 2.7 can be followed to derive a Jeffreys' non-informative prior for the multiparameter case, although we will now face a more complex notation. Consider the likelihood function $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in I\!R^n$, and a reparameterization $\boldsymbol{\phi} = \mathbf{g}(\boldsymbol{\theta})$ (*i.e.*, $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]^T = \mathbf{g}(\theta_1, \ldots, \theta_n) = [g_1(\theta_1, \ldots, \theta_n), \ldots, g_n(\theta_1, \ldots, \theta_n)]^T$ is a one-to-one continuous mapping) whose associated likelihood function is $f'_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\phi})$. We start with a relation similar to Eq. (2.73), again a simple consequence of the fact that $\mathbf{g}(\cdot)$ is invertible:

$$f'_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\phi}) = f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}), \quad \text{for } \boldsymbol{\theta} = \mathbf{g}^{-1}(\boldsymbol{\phi}). \tag{3.60}$$

Applying the chain rule of the derivative (now more complicated due to the multivariate nature of the variables) twice in a row, we can write

$$\frac{\partial^2 \log f'_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\phi})}{\partial\phi_j \partial\phi_i} = \sum_{l=1}^{n} \frac{\partial}{\partial\theta_l} \left(\sum_{k=1}^{n} \frac{\partial \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_j} \, \frac{\partial\theta_k}{\partial\phi_i} \right) \frac{\partial\theta_l}{\partial\phi_j}, \tag{3.61}$$

where all derivatives w.r.t. elements of $\boldsymbol{\theta}$ are computed at $\boldsymbol{\theta} = \mathbf{g}^{-1}(\boldsymbol{\phi})$. Now, rearranging terms, and computing expected values on both sides,

$$\begin{aligned} \mathcal{I}_{ij}(\boldsymbol{\phi}) \;\; = \;\; & \sum_{l=1}^{n}\sum_{k=1}^{n} \mathcal{I}_{kl}(\boldsymbol{\theta}) \frac{\partial\theta_k}{\partial\phi_i}\frac{\partial\theta_l}{\partial\phi_j} \\ & + \sum_{l=1}^{n}\sum_{k=1}^{n} \underbrace{E_{\mathbf{X}}\left[\frac{\partial \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_j}\right]}_{0, \text{ (see Eq. (2.66))}} \, \frac{\partial^2\theta_k}{\partial\theta_l \partial\phi_i} \, \frac{\partial\theta_l}{\partial\phi_j}. \end{aligned} \tag{3.62}$$

The set of all these equalities, for $i, j = 1, .., n$, can be written compactly in matrix notation as

$$\mathcal{I}(\boldsymbol{\phi}) = \mathbf{G}\,\mathcal{I}(\boldsymbol{\theta} = \mathbf{g}^{-1}(\boldsymbol{\phi}))\,\mathbf{G}^T, \tag{3.63}$$

where $\mathbf{G}$ (which should be written as $\mathbf{G}(\boldsymbol{\theta})$, but we simplify the notation) is a matrix given by

$$\mathbf{G} = \begin{bmatrix} \frac{\partial\theta_1}{\partial\phi_1} & \cdots & \frac{\partial\theta_1}{\partial\phi_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial\theta_n}{\partial\phi_1} & \cdots & \frac{\partial\theta_n}{\partial\phi_n} \end{bmatrix} \quad \text{for } \boldsymbol{\phi} = \mathbf{g}(\boldsymbol{\theta}). \tag{3.64}$$

Finally, obtaining determinants on both sides and then extracting square roots leads to

$$\sqrt{|\det\left[\mathbf{I}(\boldsymbol{\phi})\right]|} = |\det\left[\mathbf{G}\right]| \sqrt{|\det\left[\mathcal{I}(\boldsymbol{\theta} = \mathbf{g}^{-1}(\boldsymbol{\phi}))\right]|} \tag{3.65}$$

which is the multidimensional version of Eq. (2.75). As a conclusion, it follows that the prior suggested by Jeffreys,

$$p(\boldsymbol{\theta}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\theta})|} \tag{3.66}$$

remains unchanged under any one-to-one continuous reparametrization (*i.e.*, change of variable[5]) $\boldsymbol{\phi} = \mathbf{g}(\boldsymbol{\theta})$ in the sense that if $p(\boldsymbol{\theta}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\theta})|}$ (called the Jeffreys' prior), then, according to Eq. (3.65),

$$\begin{aligned} p(\boldsymbol{\phi}) &= p(\boldsymbol{\theta} = \mathbf{g}^{-1}(\boldsymbol{\phi}))\,|\det[\mathbf{G}]| \\ &\propto \sqrt{|\det[\mathcal{I}(\boldsymbol{\theta} = \mathbf{g}^{-1}(\boldsymbol{\phi}))]|}\,|\det[\mathbf{G}]| \\ &\propto \sqrt{|\det[\mathcal{I}(\boldsymbol{\phi})]|}. \end{aligned} \tag{3.69}$$

Naturally, also in the multivariate case, there is a particular form of Fisher information for exponential families (see Section 2.12.4). For an exponential family in canonical form,

$$f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\xi}) = \phi(\mathbf{t})\,\psi(\boldsymbol{\theta})\,\exp\{\boldsymbol{\xi}^T\mathbf{t}\}, \tag{3.70}$$

the Fisher information with respect to the canonical parameter is simply the multivariate generalization of Eqs. (2.144) and (2.160),

$$\mathcal{I}(\boldsymbol{\xi}) = E[\mathbf{t}\mathbf{t}^T|\boldsymbol{\xi}]$$

or

$$\mathcal{I}(\boldsymbol{\xi}) = \nabla^2 \log Z(\boldsymbol{\xi}),$$

meaning

$$\mathcal{I}_{ij}(\boldsymbol{\xi}) = \frac{\partial^2 \log Z(\boldsymbol{\xi})}{\partial \xi_i\,\partial \xi_j}.$$

Before showing some examples, we would like to stress again the importance of the concept of Fisher information. Like we pointed out in Chapter 3, it plays the central role in the differential geometric theories of statistical inference [1], [62], [79], [94] and is the key for the interplay between statistics and information theory [23], [67], [91].

[5] Recall the rule for performing a change of variable on a multivariate probability density function: let $\mathbf{X}$ be a continuous ($I\!R^n$ valued) random vector with p.d.f. $f_{\mathbf{X}}(\mathbf{x})$; now, let $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, where $\mathbf{g}(x_1,\ldots,x_n) = [y_1,\ldots,y_n]^T = [g_1(x_1,\ldots,x_n),\ldots,g_n(x_1,\ldots,x_n)]^T$ is a one-to-one continuous mapping. Then

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))\left|\det\left(J_{\mathbf{g}^{-1}}(\mathbf{y})\right)\right|, \tag{3.67}$$

where $\mathbf{g}^{-1}(\cdot)$ denotes the inverse function of $\mathbf{g}(\cdot)$ (which exists because $g(\cdot)$ is one-to-one) and $J_{\mathbf{g}^{-1}}(\cdot)$ is its Jacobian determinant

$$J_{\mathbf{g}^{-1}}(\mathbf{y}) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \tag{3.68}$$

Example 3.5.5

Let us go back to Example 3.5.3, where the data $\mathbf{x}$ is a single sample (extension to several observations is straightforward) from a Gaussian random variable of mean $\mu = \theta_1$ and variance $\sigma^2 = \theta_2$.

For the log-likelihood in hand, $\log f_X(x|\theta_1, \theta_2) = (-1/2)\log(2\pi\theta_2) - (x - \theta_1)^2/(2\theta_2)$, the Fisher information matrix is

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= -E\left[\begin{array}{cc} -\frac{1}{\theta_2} & -\frac{x-\mu}{\theta^2} \\ -\frac{x-\mu}{\theta^2} & \frac{1}{2\,\theta^2} - \frac{(x-\mu)^2}{\theta^3} \end{array}\right] \\ &= \left[\begin{array}{cc} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{array}\right] = \left[\begin{array}{cc} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2(\sigma^2)^2} \end{array}\right], \end{aligned} \tag{3.71}$$

with the corresponding Jeffreys' prior being

$$p(\mu, \sigma^2) \propto \sqrt{\frac{1}{(\sigma^2)^3}} = \frac{1}{(\sigma^2)^{3/2}}. \tag{3.72}$$

This prior is usually criticized by arguing that in an "ignorance" situation, μ and σ^2 should be considered *a priori* independent; then, one should take $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ with each of the factors derived from invariance arguments [93]. In this case, the resulting joint prior, according to the results seen in Section 2.7, would be $p(\mu, \sigma^2) = 1/\sigma^2$ rather than the one in Eq. (3.72). An alternative way to obtain this prior $p(\mu, \sigma^2) = 1/\sigma^2$ is through Bernardo's reference analysis [14]; although this approach is beyond the scope of this text, a point worth mentioning is that it is the type of information theoretical approach briefly reviewed in Section 2.9.3 that leads to reference priors.

End of Example 3.5.5

Example 3.5.6

Let us now consider the problem of estimating an unknown covariance matrix of a p-variate Gaussian likelihood with known mean (for simplicity taken as zero). In this example we follow the approach in [18]. The log-likelihood function (for one observation) is

$$\log f_{\mathbf{X}}(\mathbf{x}|\mathbf{A}) \propto \frac{1}{2}\log\det\left[\mathbf{A}^{-1}\right] - \frac{1}{2}\text{tra}\left[\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^T\right]. \tag{3.73}$$

Since $\mathbf{A}$ is necessarily symmetric, it only has $p(p+1)/2$ distinct (*i.e.*, free) elements. Computing derivatives of the two terms w.r.t. the elements B_{ij} of the concentration matrix $\mathbf{B} = \mathbf{A}^{-1}$, we have:

$$\begin{aligned} \frac{1}{2}\frac{\partial \log\det\left[\mathbf{A}^{-1}\right]}{\partial B_{ij}} &= \frac{1}{2}\frac{1}{\det[\mathbf{B}]}\frac{\det[\mathbf{B}]}{\partial B_{ij}} \\ -\frac{1}{2}\frac{\partial \text{tra}\left[\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^T\right]}{\partial B_{ij}} &= -\frac{1}{2}\frac{\partial \text{tra}\left[\mathbf{B}\mathbf{x}\mathbf{x}^T\right]}{\partial B_{ij}} = -\frac{x_i\, x_j}{2} \end{aligned}$$

We start by noticing that the second derivative of the second term is zero, because the first is a constant with respect to any B_{kl}. If we now recall that a determinant can always be written as

$$\det \mathbf{B} = \sum_{j=1}^{n} B_{ij} c_{ij}, \tag{3.74}$$

where c_{ij} (called the cofactor) equals $(-1)^{i+j}$ times the determinant of the submatrix of $\mathbf{B}$ obtained by deleting row i and column j, it results that $\partial \det [\mathbf{B}] / \partial B_{ij} = c_{ij}$. But $c_{ij} / \det \mathbf{B}$ is simply the element A_{ij} of $\mathbf{A} = \mathbf{B}^{-1}$. Now the second derivative becomes

$$\frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{x}|\mathbf{A})}{\partial B_{kl} \partial B_{ij}} = \frac{1}{2} \frac{\partial^2 \log \det \left[\mathbf{A}^{-1}\right]}{\partial B_{kl} \partial B_{ij}} = \frac{1}{2} \frac{\partial A_{ij}}{\partial B_{kl}}, \tag{3.75}$$

for $i, j, k, l = 1, \ldots, p$, with $j \geq i$ and $l \geq k$; consequently, in compact notation,

$$\det [\mathcal{I}(\mathbf{B})] = \frac{1}{2} \det \left[\frac{\partial \mathbf{A}}{\partial \mathbf{B}}\right], \tag{3.76}$$

where $\partial \mathbf{A} / \partial \mathbf{B}$ denotes a matrix of partial derivatives of the type $\partial A_{ij} / \partial B_{kl}$. It can be shown (see [2], [18]) that since we only have $p(p+1)/2$ free parameters (recall that matrices $\mathbf{B}$ and $\mathbf{A}$ have to be symmetric), then

$$\det \left[\frac{\partial \mathbf{A}}{\partial \mathbf{B}}\right] = \det [\mathbf{A}]^{(p+1)}. \tag{3.77}$$

Finally, invoking the transformation rule in Eq. (3.63) leads to

$$\det [\mathcal{I}(\mathbf{A})] = \det [\mathcal{I}(\mathbf{B})] \; \det \left[\frac{\partial \mathbf{A}}{\partial \mathbf{B}}\right]^{-2} = \frac{1}{2} \det [\mathbf{A}]^{-(p+1)} \tag{3.78}$$

and to the corresponding Jeffreys' prior

$$p(\mathbf{A}) \propto \det [\mathbf{A}]^{-\frac{1}{2}(p+1)}.$$

Notice that when $p = 1$, $\mathbf{A} \equiv \sigma^2$, this naturally reduces to $p(\sigma^2) \propto 1/\sigma^2$, coinciding with Eq. (2.79). Of course we also derived, in passing, a Jeffreys' prior for $\mathbf{B}$, which has the same form as the one for $\mathbf{A}$, specifically

$$p(\mathbf{B}) \propto \det [\mathbf{B}]^{-\frac{1}{2}(p+1)}.$$

Notice how this prior for the precision matrix can be seen as the limit of a Wishart density shown in Eq. (3.55),

$$\det [\mathbf{B}]^{-\frac{1}{2}(p+1)} \propto \lim_{\nu \to -(p+1)} \lim_{\boldsymbol{\Gamma} \to 0} \left(\sqrt{|\mathbf{B}|}\right)^{\nu} \exp\left\{-\frac{1}{2} \text{tra}\,(\mathbf{B}\boldsymbol{\Gamma})\right\} \tag{3.79}$$

i.e. a Wishart $\mathcal{W}_p\,(0,0))$ density. From this observation, the PM estimate of $\mathbf{B}$, given n observations, and under the Jeffreys' prior, can then easily be found to be

$$\widehat{\mathbf{B}}_{\mathrm{PM}} = (-p-1+n)\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1}.$$

End of Example 3.5.6

3.6 Maximum Entropy Priors

There is nothing specifically univariate in the information theoretic concepts and techniques described in Section 2.8. The quantities there described, namely the entropy, the Kullback-Leibler divergence, and the mutual information can be easily extended to multivariate versions. However, some aspects do deserve special attention and will be considered in this section.

3.6.1 Independence Maximizes Entropy

We start by referring an important inequality that is relevant for the design of maximum entropy priors. This inequality is supported on an intermediate result known as the chain rule for entropies which we now review [23]. Let $\mathbf{S} = [S_1, ..., S_d]^T$ be a vector random variable (*i.e.*, a set of random variables) whose probability function is $p_{\mathbf{S}}(\mathbf{s})$; then, the joint entropy of $S_1, ..., S_d$, that is, the entropy of $\mathbf{S}$, can be decomposed according to

$$H(\mathbf{S}) = H(S_1, ..., S_d) = \sum_{i=1}^{d} H(S_i|S_{i-1}, ..., S_1)$$

which is a simple consequence of Bayes law. The conditional entropies involved in this decomposition are of course bounded above by the fact that conditioning can not increase entropy (see Section 2.9.1); that is $H(S_i|S_{i-1}, ..., S_1) \leq H(S_i)$. This inequality, applied to the chain rule, leads to

$$H(S_1, ..., S_d) \leq \sum_{i=1}^{d} H(S_i) \tag{3.80}$$

with equality if and only if the S_i's are independent of each other, in which case $H(S_i|S_{i-1}, ..., S_1) = H(S_i)$.

A consequence of this fact, that is relevant for the derivation of maximum entropy priors, is: in the absence of any specific prior information regarding the dependency structure of the unknowns $S_1, ..., S_d$, a maximum entropy prior is one that specifies independence, that is,

$$p_{\mathbf{S}}(\mathbf{s}) = p_{S_1}(s_1) \cdot p_{S_2}(s_2) \cdots p_{S_d}(s_d).$$

3.6.2 Discrete Problems

When the elements of $\mathbf{S} = [S_1, ..., S_d]^T$ are discrete variables, the prior has to be a probability mass function $p_{\mathbf{S}}(\mathbf{s})$. Like in the scalar situation, the available information has to be formally expressed; the multivariate version of Eq. (2.88) is, naturally, a set of $m+1$ equalities now with the form

$$\sum_{\mathbf{s}} p_{\mathbf{S}}(\mathbf{s})\, g_k(\mathbf{s}) = \mu_k, \quad \text{for } \; k = 0, 1, ..., m, \tag{3.81}$$

where $g_0(\mathbf{s}) = 1$ and $\mu_0 = 1$ (the zero-th constraint simply imposes that $\sum p(\mathbf{s}) = 1$). With this type of constraints, the ME probability mass function is

$$p_S^{\mathrm{ME}}(\mathbf{s}) = \exp\{\lambda_0 + \sum_{k=1}^{m} \lambda_k\, g_k(\mathbf{s})\} \quad \text{for } \; \mathbf{s} \in \mathcal{S}, \tag{3.82}$$

where the parameters λ_k are solved for by requiring that $p_S^{\mathrm{ME}}(\mathbf{s})$ satisfies the constraints in Eq. (3.81) [23].

Example 3.6.1

Let us look at a multivariate version of Example 2.8.1, by letting $\mathcal{S} = \mathbb{N}_0^d$, that is, $\mathbf{s} = [s_1, s_2, ..., s_d]^T$ with $s_i \in \mathbb{N}_0 = \{0, 1, 2, ...\}$. Consider that, in addition to the zero-th constraint, there are d other constraints defined by a set of d functions such that

$$g_i(\mathbf{s}) = s_i$$

and the corresponding parameters $\mu_1,\ \mu_2, ..., \mu_d$. Then, according to Eq. (3.82),

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \exp\{\lambda_0 + \sum_{i=1}^{d} \lambda_i\, g_i(\mathbf{s})\} = \exp\{\lambda_0\} \prod_{i=1}^{d} (\exp\{\lambda_i\})^{s_i} \tag{3.83}$$

The normalization condition allows us to write the following equality

$$\exp\{\lambda_0\} = \prod_{i=1}^{d} (1 - \exp\{\lambda_i\})\,.$$

In conclusion, the maximum entropy probability mass function is the product of d univariate geometric distributions,

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \prod_{i=1}^{d} (1 - \theta_i)\, \theta_i^{s_i}, \tag{3.84}$$

with $\theta_i = \exp\{\lambda_i\}$. Invoking the constraints on the expected values leads to $\exp\{\lambda_i\} = \theta_i = \mu_i/(1+\mu_i)$.

The interesting point in this example is the following: due to the particular form of the functions $g_i(\cdot)$, the constraints only concern marginal properties of the unknowns. As a consequence, and in accordance to what we just saw in Section 3.6.2, the obtained ME distribution expresses independence among the components of $\mathbf{S}$.

End of Example 3.6.1

Example 3.6.2

Consider that $\mathbf{s} \in \mathcal{S} = \{0,1\}^d$, the set of all binary sequences of length d. Let us take a single ($m = 1$) restriction (apart from the zero-th one), on the expected value of the number of ones, $g_1(\mathbf{s}) = n_1(\mathbf{s})$, required to be μ_1. Then, according to Eq. (2.89),

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \exp\{\lambda_0\}\left(\exp\{\lambda_1\}\right)^{n_1(\mathbf{s})}. \tag{3.85}$$

Invoking the normalization condition

$$\begin{aligned}\exp\{\lambda_0\}\sum_{\mathbf{s}\in\mathcal{S}}\left(\exp\{\lambda_1\}\right)^{n_1(\mathbf{s})} &= \exp\{\lambda_0\}\sum_{n_1(\mathbf{s})=0}^{n}\left(\begin{array}{c} n \\ n_1(\mathbf{s})\end{array}\right)\left(\exp\{\lambda_1\}\right)^{n_1(\mathbf{s})}\\ &= \exp\{\lambda_0\}\left(1+\exp\{\lambda_1\}\right)^{n} \;=\; 1,\end{aligned}$$

we get $\exp\{\lambda_0\} = \left(1+\exp\{\lambda_1\}\right)^{-n}$. Now, by similarity with Eq. (2.146), we immediately conclude that this is a Bernoulli probability distribution with cannonical parameter equal to λ_1 and natural parameter n_1. In the usual Bernoulli form,

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \theta^{n_1(\mathbf{s})}(1-\theta)^{n-n_1(\mathbf{s})},$$

with $\theta = \exp\{\lambda_1\}/(1+\exp\{\lambda_1\}) = \mu_1/n$.

End of Example 3.6.2

3.6.3 Estimation Problems

Let us generalize the results of Section 2.8.3 to multivariate situations. Consider that we have agreed on some non-informative prior $q_{\mathbf{S}}(\mathbf{s})$ and that the available information is expressed by a set of integral equalities

$$\int_{\mathcal{S}} p_{\mathbf{S}}(\mathbf{s}) g_k(\mathbf{s})\, ds = \mu_k, \quad \text{for } \; k = 0, 1, ..., m. \tag{3.86}$$

Again, $g_0(\mathbf{s}) = 1$ and $\mu_0 = 1$, which are necessary to normalize $p_{\mathbf{S}}(\mathbf{s})$. The *maximum entropy* (ME) prior (or better, the *least informative prior* with respect to $q_{\mathbf{S}}(\mathbf{s})$) is

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = q_{\mathbf{S}}(\mathbf{s})\exp\{\lambda_0 + \sum_{k=1}^{m}\lambda_k\, g_k(\mathbf{s})\} \quad \text{for } \; \mathbf{s}\in\mathcal{S}, \tag{3.87}$$

where the λ_k's must be obtained from the constraints; this result can (as in the discrete case) be derived by Lagrange multipliers or through the information inequality [23]. As in the univariate case, in the absence of additional constraints apart from normalization, i.e., if $m = 0$, then, the ME prior coincides with the adopted non-informative density $q_{\mathbf{S}}(\mathbf{s})$. As for scalar unknowns, in the continuous case the ME prior is specified up to a reference density $q_{\mathbf{S}}(\mathbf{s})$ which has to be chosen *a priori*. In all the examples below, we take $q_{\mathbf{S}}(\mathbf{s})$ to be the uniform density.

Example 3.6.3

Let $\mathbf{s} = [s_1, s_2, ..., s_d]^T$ with each s_i being real and non-negative. Suppose that the constraints are specified by a set of functions $g_i(\mathbf{s}) = s_i$, and the corresponding expected values $\mu_i > 0$, for $i = 1, ..., d$. The ME probability density function is, according to Eq. (3.87),

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \exp\{\lambda_0 + \sum_{i=1}^{d} \lambda_i g_k(\mathbf{s})\} = \exp\{\lambda_0\} \prod_{i=1}^{d} \exp\{\lambda_i s_i\}.$$

The normalization constraint (possible if all λ_i's are smaller than zero) requires that $\exp\{\lambda_0\} = \prod_{i=1}^{d}(-\lambda_i)$, while the other constraints lead to $\lambda_i = -1/\mu_i$ (thus compatible with $\lambda_i < 0$). Putting these two results together yields

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \prod_{i=1}^{d} \frac{1}{\mu_i} \exp\left\{-\frac{s_i}{\mu_i}\right\}, \qquad (3.88)$$

that is, a product of exponential densities. Just like in Example 3.6.1, since the constraints only concern marginal properties, the obtained ME distribution expresses independence among the components of $\mathbf{S}$.

End of Example 3.6.3

Example 3.6.4

This example shows that the maximum entropy nature of the Gaussian distribution is still valid in the multivariate case. Let the configuration space be $\mathcal{S} = \mathbb{R}^d$ and consider two sets of constraints: one involving only single variables, expressing constraints on their expected values, $E[g_i(\mathbf{s})] = E[s_i] = \mu_i$, for $i = 1, 2, ..., d$; another one involving pairs of variables, that is, $h_{ij}(\mathbf{s}) = s_i s_j$ with expected values γ_{ij} (with, necessarily, $\gamma_{ij} = \gamma_{ji}$), for $i, j = 1, 2, ..., d$. Once more, the resulting ME prior, according to Eq. (3.87), is

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \exp\{\lambda_0\} \exp\left\{ \sum_{i=1}^{d} \mu_i s_i + \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_{ij} s_i s_j \right\} \qquad (3.89)$$

Some manipulation allows writing this density with the form

$$p_{\mathbf{S}}^{\mathrm{ME}}(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}, \mathbf{C}), \qquad (3.90)$$

a d-variate Gaussian density, with $\boldsymbol{\mu} = [\mu_1, ..., \mu_d]^T$ and $\mathbf{C} = \boldsymbol{\Gamma} - \boldsymbol{\mu}\,\boldsymbol{\mu}^T$, where $\boldsymbol{\Gamma}$ is the matrix with γ_{ij} in the (i,j) position.

End of Example 3.6.4

3.6.4 Maximum Entropy and Exponential Families

As we pointed out at the end of Section 2.12.1, maximum entropy distributions constitute exponential families, as is clear from the similarity between Eqs. (3.82) and (3.87), and Eq. (3.43). Although we could have looked at the consequences of this fact for the univariate scenario, in the compound/multivariate case it is far more interesting and useful.

Recall that the constraints for the maximum entropy formalism specify the expected values of certain functions of the random variable at hand (see Eqs. (3.81) and (3.86)),

$$E_{\mathbf{X}}\left[g_k(\mathbf{x})|p_{\mathbf{X}}(\mathbf{x})\right] = \mu_k, \quad \text{for } k = 0, 1, ..., m, \tag{3.91}$$

where $E_{\mathbf{x}}\left[\cdot|p_{\mathbf{X}}(\mathbf{x})\right]$ denotes expected value with respect to the probability function $p_{\mathbf{X}}(\mathbf{x})$. With C denoting the set of constraints imposed, i.e., $C = \{g_k(\cdot), \mu_k, \;\; k = 1, ..., m\}$, we follow [111] by writing

$$\Omega_C = \{p_{\mathbf{X}}(\mathbf{x}) : \;\; E_{\mathbf{X}}\left[g_k(\mathbf{x})|p_{\mathbf{X}}(\mathbf{x})\right] = \mu_k, \quad \text{for } k = 0, 1, ..., m\}$$

as the set of probability functions that verify the set of constraints C. Of course, with this notation, the maximum entropy prior can be written as

$$p_{\mathbf{X}}^{\mathrm{ME}}(\mathbf{x}) = \arg \max_{p_{\mathbf{X}}(\mathbf{x}) \in \Omega_C} H(\mathbf{X})$$

(here it would make more sense to use the notation $H[p_{\mathbf{X}}(\mathbf{x})]$, because the entropy strictly depends on the probability function). The resulting probability functions can be written as (see Eqs. (3.82) and (3.87), with respect to a flat non-informative distribution $q_{\mathbf{X}}(\mathbf{x})$)

$$p_{\mathbf{X}}^{\mathrm{ME}}(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left\{\sum_{k=1}^{m} \lambda_k \, g_k(\mathbf{x})\right\} = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left\{\boldsymbol{\lambda}^T \mathbf{g}\,(\mathbf{x})\right\}, \tag{3.92}$$

that is, in exponential family form, where $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_m]^T$ is the natural parameter and $\mathbf{g}\,(\mathbf{s}) = [g_1(\mathbf{x}), ..., g_1(\mathbf{x})]^T$ the associated statistic. The specific values of the parameters are obtained from the constraints and depend on the μ_k.

Now, consider a slightly different scenario where, rather than having expected value constraints (Eq. (3.91)), one has a set of n independent and identically distributed observations $\mathbf{x}_{(n)} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$; consider that the goal is to estimate the probability function $f_{\mathbf{X}}(\mathbf{x})$. Given a set of m

statistics $\{g_k(\mathbf{x}), \;\; k=1,...,m\}$ (in some contexts, called *features* [82]), it is possible to compute their sample averages:

$$\overline{g_k(\mathbf{x})} = \frac{1}{n}\sum_{i=1}^{n} g_k(\mathbf{x}_i), \quad \text{for } \; i=1,...,m.$$

The set of probability functions that agree with the data is Ω_C, with $C = \{g_k(\cdot), \mu_k, \;\; k=1,...,m\}$, where $\mu_k = \overline{g_k(\mathbf{x})}$. In the absence of any other information, the maximum entropy criterion dictates that the maximum entropy distribution from Ω_C be chosen, as the one which is the least informative [56], [57], [58]; in other words, it is the one that assumes less. As we have a seen above, the resulting probability function is given by Eq. (3.92), with the specific values of the parameters obtained from the constraints in Eq. (3.91) together with the $\mu_k = \overline{g_k(\mathbf{x})}$.

Another possible approach to the problem of estimating the probability function $f_{\mathbf{X}}(\mathbf{x})$ is to assume that it belongs to an exponential family with a certain (given) set of natural statistics $\{g_k(\mathbf{x}), \;\; k=1,...,m\}$, that is

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\lambda})}\exp\left\{\boldsymbol{\lambda}^T \mathbf{g}\left(\mathbf{x}\right)\right\}. \tag{3.93}$$

or, for the full data set,

$$f_{\mathbf{X}_{(n)}}(\mathbf{x}_{(n)}|\boldsymbol{\lambda}) = \prod_{i=1}^{n}\frac{1}{Z(\boldsymbol{\lambda})}\exp\left\{\boldsymbol{\lambda}^T \mathbf{g}\left(\mathbf{x}\right)\right\} = \frac{1}{Z(\boldsymbol{\lambda})^n}\exp\left\{\boldsymbol{\lambda}^T \sum_{i=1}^{n}\mathbf{g}\left(\mathbf{x}_i\right)\right\}.$$

By the maximum likelihood criterion,

$$\widehat{\boldsymbol{\lambda}}_{\text{ML}} = \arg\max_{\boldsymbol{\lambda}}\left[-n\log Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T\sum_{i=1}^{n}\mathbf{g}\left(\mathbf{x}_i\right)\right]. \tag{3.94}$$

Computing derivatives, invoking the fact that

$$\frac{\partial \log Z(\boldsymbol{\lambda})}{\partial \lambda_i} = E\left[g_i(\mathbf{x})|\boldsymbol{\lambda}\right]$$

(see Eq. (2.143)), and equating to zero leads to

$$E_{\mathbf{X}}\left[g_k(\mathbf{x})|f_{\mathbf{X}}(\mathbf{x}|\widehat{\boldsymbol{\lambda}}_{\text{ML}})\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}\left(\mathbf{x}_i\right), \;\; \text{for } \; k=1,...,m. \tag{3.95}$$

But this last set of equations is precisely the one that is solved to find the parameters in the maximum entropy approach. This shows that finding maximum likelihood estimates of parameters of exponential families and looking for maximum entropy distributions, both based on the same set of statistics, are dual problems that lead to the same solution [67]. Of course, a more difficult problem is how to select the best set of statistics; in the context of Markov random fields, this has been explored in [82], [111].

3.7 Summary

This chapter extended most of the ideas presented in previous ones to situations where rather than a single unknown there is a set of unknowns, or an unknown vector.

For sets of unknowns, we divided the possible loss functions in those that can be expressed as a sum of individual loss functions, one for each unknown, which we called *additive*, from those that do not allow such a decomposition. The non-additive loss functions lead to joint criteria, that is, inference rules that have to be solved simultaneously with respect to all the unknown variables. On the other hand, additive losses lead to what we called *marginal criteria*; these are characterized by depending on the marginal *a posteriori* probability function of each unknown. These general results were particularized to the common loss functions studied in Chapter 1. The scenario with Gaussian prior and linear-Gaussian observation, which is a classical setup for many problems and applications, was given special attention.

The remaining sections of this chapter were devoted to extending prior design techniques reviewed in Chapter 2 to the multivariate case. Specifically, we started by looking at improper priors and maximum likelihood inference, exponential families and how they allow a simple derivation of conjugate priors, and Jeffreys' non-informative priors. These topics were illustrated with examples of specific models: multivariate Gaussian of unknown mean with the corresponding Gaussian prior, multinomial likelihood and Dirichlet prior, univariate Gaussian of unknown mean and variance, multivariate Gaussian with unknown mean and covariance matrix.

The maximum entropy principle for multivariate variables was the topic of the last section. Several classical probability functions were shown to result from this criterion under simple constraints. Finally, we studied the intimate relation that exists between maximum entropy distributions and exponential families, which is relevant and will be used in latter chapters in the context of Markov random fields.

Appendix A
Notation

Although the reader is assumed to be familiar with the elementary concepts of probability theory, we will briefly present the notational conventions to be used throughout this book. A random variable (r.v.) is a function (denoted by a capital letter, e.g. X, Y, or S) from an event space into a *sample space* (denoted by caligraphic style, e.g., $\mathcal{X}$, $\mathcal{Y}$, or $\mathcal{S}$). Particular outcomes of a r.v. will be denoted by the corresponding lower case (e.g., x, y, or s, respectively). In the case of sets of random variables, e.g. processes and vectors (1-D) or fields (2D), boldface will be used (e.g., $\mathbf{X}$ or $\mathbf{x}$). If the sample space is continuous, e.g. $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathbb{R}^n$, the respective (continuous) r.v. is characterized by its probability density function (p.d.f.) denoted as $f_X(x)$ (or $f_{\mathbf{X}}(\mathbf{x})$, for a vector or field); a p.d.f. $f_X(x)$ is a function $f_X(x) : \mathcal{X} \rightarrow \mathbb{R}$ such that $\int_{\mathcal{X}} f_X(x)dx = 1$. We will usually drop the subscript and write simply $f(x)$ or $p(x)$. For a discrete (i.e., countable, finite or infinite) sample space, e.g. $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, the (discrete) r.v. is characterized by its probability mass function (p.m.f.), still denoted as $f(x)$; this is still a function $f(x) : \mathcal{X} \rightarrow \mathbb{R}$ which can also be seen as a set of probability values, one for each outcome, e.g., $f(x) = \{f(x) : x \in \mathcal{X}\}$; the constraint $\sum_{x \in \mathcal{X}} f(x) = 1$ must obviously also be met. Exceptions to these conventions will be individually pointed out and justified.

Subsets of sample spaces are called *events*, e.g. if X is a r.v. taking values in $\mathcal{X}$ then any $A \subseteq \mathcal{X}$ is an event. The notation $P(A)$ will denote *the probability of event* A which is given by[1]

$$P(A) = \begin{cases} \displaystyle\int_A f_X(x)\,dx, & \text{if} \quad X \;\text{ is a continuous r.v.} \\ \displaystyle\sum_{x \in A} f(x), & \text{if} \quad X \;\text{ is a discrete r.v.} \end{cases} \tag{A.1}$$

The expected value (or expectation) of a function $g(x)$ with respect to a random variable $X \in \mathcal{X}$ is denoted $E[g(x)]$ and defined to be

$$E[g(x)] = \begin{cases} \displaystyle\int_{\mathcal{X}} g(x) f_X(x)\,dx, & \text{if} \quad X \;\text{ is a continuous r.v.} \\ \displaystyle\sum_{x_i \in \mathcal{X}} g(x_i) f(x_i), & \text{if} \quad X \;\text{ is a discrete r.v.} \end{cases} \tag{A.2}$$

Sometimes (to avoid possible confusions) the r.v. with respect to which the expectation is being computed will be indicated by a subscript, e.g., $E_X[\cdot]$.

[1] For the sake of simplicity, and to make it readable by a wider audience, this tutorial will exclude any measure theoretic aspects [3]; all subsets/events will be considered as having a well defined probability.

Appendix B

Markov Processes and Chains: a Brief Review

The theory of Markov processes is a branch of the study of stochastic processes that underlies much of statistical signal processing and many areas of applied statistics. It is a fairly standard topic covered in many textbooks devoted to the study of random processes, and in many specialized texts. A few examples are [6], [39], [49], [55], [65], [80]; a usefull review paper is [32].

B.1 Discrete-Index Stochastic Processes

A discrete-time (or discrete-index) random (or stochastic) process is simply a sequence of random variables indexed by a (discrete) parameter (which may or may not possess some temporal meaning),

$$\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_n, ...)\,.$$

Full knowledge of a random process implies the ability to write the joint probability function of any set of random variables in the process, in particular,

$$p_{S_1,...,S_n}(s_1, ..., s_n).$$

A stochastic process is said stationary if the joint probability functions are invariant under index shifs, i.e.,

$$p_{S_{k+1},...,S_{k+n}}(s_1, ..., s_n) = p_{S_1,...,S_n}(s_1, ..., s_n).$$

A sequence of independent and identically distributed (according to some probability function $p_S(s)$) random variables is the simplest example of a stationary process; in that case,

$$p_{S_1,...,S_n}(s_1, ..., s_n) = \prod_{i=1}^{n} p_S(s_i),$$

which is clearly stationary.

B.2 Markov Processes

Arguably the simplest type of dependency that can be exhibited by the variables of a random process is the one found in first-order Markov processes: each variable S_i dependes only the preceding one, S_{i-1}; moreover, conditionally on S_{i-1}, it is independent of all other preceding variables. Formally, the process is called a *first-order Markov process* when

$$p_{S_n}(s_n|s_{n-1}, s_{n-2}, ..., s_1) = p_{S_n}(s_n|s_{n-1}). \tag{B.1}$$

The joint probability function of any process (of any set of random variables) can be factored as

$$\begin{aligned} &p_{S_1,...,S_n}(s_1, ..., s_n) = \\ &\qquad p_{S_n}(s_n|s_{n-1}, ..., s_1)\, p_{S_{n-1}}(s_{n-1}|s_{n-2}, ..., s_1) \cdots p_{S_2}(s_2|s_1)\, p_{S_1}(s_1), \end{aligned}$$

which is a trivial chain application of $p(A|B)p(B) = p(A, B)$.

One of the most important consequence of the Markovianity of a process is that its factorization becomes simply

$$\begin{aligned} &p_{S_1,...,S_n}(s_1, ..., s_n) = \\ &\qquad p_{S_n}(s_n|s_{n-1})\, p_{S_{n-1}}(s_{n-1}|s_{n-2}) \cdots p_{S_2}(s_2|s_1)\, p_{S_1}(s_1). \end{aligned} \tag{B.2}$$

Accordingly, a Markov process is completely characterized (i.e., it is possible to compute any joint probability function) once the initial probability function $p_{S_1}(s_1)$, and the sequence of *transition* probability functions $p_{S_i}(s_i|s_{i-1})$ are given.

Consider a Markov process such that each S_i can take values on a finite set (the i-th *state space*) $\mathcal{S}_i = \{1, 2, ..., M_i\}$ (without loss of generality here identified with sets of integers; notice that these are merely labels). In this case, the process is called a *finite* Markov process, $p_{S_1}(s_1)$ is a set of M_1 probability values, and the transition probability functions $p_{S_i}(s_i|s_{i-1})$ define $M_{i-1} \times M_i$ *transition matrices* $\mathbf{P}(i) = [P_{kl}(i)]$ according to

$$P_{kl}(i) = p_{S_i}(s_i = l|s_{i-1} = k) \geq 0. \tag{B.3}$$

Given their meaning, these matrices must verify

$$\sum_{l=1}^{M_i} P_{kl}(i) = \sum_{l=1}^{M_i} p_{S_i}(s_i = l | s_{i-1} = k) = 1,$$

and are called *stochastic matrices.* If everything in the previous definitions is index-invariant, i.e., $\mathcal{S}_i = \mathcal{S}$ (the state space, of course with $M_i = M$) and $\mathbf{P}(i) = \mathbf{P}$, we have a so-called *time-invariant* or *homogeneous* Markov *chain.*

If the probability function of variable S_n is $p_{S_n}(s_n)$, then that of the "next" variable, S_{n+1}, can easily be obtained by noting that

$$\begin{aligned} p_{S_{n+1}}(s_{n+1}) &= \sum_{s_n \in \mathcal{S}_n} p_{S_n, S_{n+1}}(s_n, s_{n+1}) \\ &= \sum_{s_n \in \mathcal{S}_n} p_{S_{n+1}}(s_{n+1} | s_n) p_{S_n}(s_n) \end{aligned} \tag{B.4}$$

(with integrals taking place of the summations in the case of continuous state spaces).

If we are the presence of a time invariant chain (or process), then a probability function that remains unchanged from index n to the next index $n+1$, i.e., such that

$$p_{S_{n+1}}(b) = \sum_{s_n \in \mathcal{S}_n} p_{S_{n+1}}(b | s_n) p_{S_n}(s_n) = p_{S_n}(b) \tag{B.5}$$

(again, with integrals instead of summations in the case of continuous state spaces), is called a *stationary distribution.*

B.3 Irreducibility and Stationary Distributions

A Markov chain is called *irreducible* if

$$p_{S_n}(s_n = l | s_i = k) > 0,$$

for all $l \neq k \in \mathcal{S}$ and all $\infty > n > i \geq 1$. This means that if the chain is in (any) state k, at time i, then it can reach (any other) state l in finite time n.

In the finite discrete case (homogeneous finite Markov chain), the stationarity condition can be compactly written in matrix notation by letting

$$\mathbf{p}_n \equiv [p_{S_n}(s_n = 1)\; p_{S_n}(s_n = 2) \ldots p_{S_n}(s_n = M)]^T;$$

then, Eq. (B.5) becomes

$$\mathbf{p}_n = \mathbf{p}_n \, \mathbf{P}. \tag{B.6}$$

This is of course only possible if matrix $\mathbf{P}$ possesses a unit left eigenvalue, called the *Perron-Frobenius eigenvalue.* The *Perron-Frobenius theorem* (simply put) states that if matrix $\mathbf{P}$ corresponds to an irreducible Markov chain, then it will have a unit left eigenvalue; moreover, to this eigenvalue corresponds a one dimensional eigenspace $\alpha\mathbf{v}$, i.e., $\alpha\mathbf{v}\mathbf{P} = \alpha\mathbf{v}$. By choosing α such that $\alpha \sum_{i=1}^{M} v_i = 1$, then $\alpha\mathbf{v}$ can be identified with the stationary distribution of the Markov chain.

B.4 Chapman-Kolmogorov Equations

The dependence relationship between non-adjacent variables of any discrete time process may be obtained according to

$$p_{S_n}(s_n|s_{n-i}) = \int p(s_n, s_{n-1}, \ldots, s_{n-i+1}|s_{n-i})\, ds_{n-1} \cdots ds_{n-i+1} \qquad \text{(B.7)}$$

(with the integral replaced by summations in the case of discrete configurations spaces). In the case of a Markov process, the integrand is simply

$$\begin{aligned} & p(s_n, s_{n-1}, \ldots, s_{n-i+1}|s_{n-i}) = \\ & \qquad p_{S_n}(s_n|s_{n-1})\, p_{S_{n-1}}(s_{n-1}|s_{n-2}) \cdots p_{S_{n-i+1}}(s_{n-i+1}|s_{n-i}). \end{aligned}$$

Of course, these relations can be chained,

$$p_{S_n}(s_n|s_i) = \int p_{S_n}(s_n|s_k)\, p_{S_k}(s_k|s_i)\, ds_k, \qquad \text{(B.8)}$$

for any $n > k > i > 1$; in these form these are usually called Chapman-Kolmogorov equations.

B.5 Other Properties of Markov Processes

A Markov process $\mathbf{S}$ is also Markovian if the "time" direction is reversed; i.e., if the Markov property in Eq. (B.1) holds, then so does

$$p_{S_n}(s_n|s_{n+1}, s_{n+2}, ..., s_{n+k}) = p_{S_n}(s_n|s_{n+1}). \qquad \text{(B.9)}$$

In fact, omitting the subscripts from the probability functions,

$$\begin{aligned} p(s_n|s_{n+1}, s_{n+2}, ..., s_{n+k}) &= \frac{p(s_n, s_{n+1}, s_{n+2}, ..., s_{n+k})}{p(s_{n+1}, s_{n+2}, ..., s_{n+k})} \\ &= \frac{p(s_{n+k}|s_{n+k-1}) \cdots p(s_{n+1}|s_n)\, p(s_n)}{p(s_{n+k}|s_{n+k-1}) \cdots p(s_{n+2}|s_{n+1})\, p(s_{n+1})} \\ &= \frac{p(s_{n+1}|s_n)\, p(s_n)}{p(s_{n+1})} = p(s_n|s_{n+1}) \end{aligned}$$

where the last equality is simply Bayes law.

In a Markov process, the "past" and the "future" are independent when conditioned on the "present", that is, for any integers $k, l > 0$,

$$\begin{aligned} p(s_{n+k}, s_{n-l}|s_n) &= \frac{p(s_{n-k}, s_{n+l}, s_n)}{p(s_n)} \\ &= \frac{p(s_{n+k}|s_n, s_{n-l})p(s_n|s_{n-l})p(s_{n-l})}{p(s_n)} \\ &= p(s_{n+k}|s_n)\, p(s_{n-l}|s_n), \end{aligned} \tag{B.10}$$

because, again by Bayes law, $p(s_n|s_{n-l})p(s_{n-l})/p(s_n) = p(s_{n-l}|s_n)$.

Finally, and in close relation with Markov random fields, Markov processes also exhibit a *bilateral* Markov property stated as: if $\mathbf{S}$ is a Markov process, then, on each interval $\{1, 2, ..., n\}$, and for any $k \neq 1$,

$$p_{S_k}(s_k|\{s_i\,; 1 \leq i \leq n, i \neq k\}) = p_{S_k}(s_n|s_{k-1}, s_{k+1}) \tag{B.11}$$

The proof of this property follows the same lines as the previous two and is left as an exercise to the reader or may be found, e.g., in [108].

References

[1] S. Amari. *Differential Geometry in Statistical Inference.* Institute of Mathematical Statistics, Hayward, CA, 1987.

[2] T. Anderson. *Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, 1984. 2nd Edition.

[3] R. Ash. *Real Analysis and Probability.* Academic Press, New York, 1972.

[4] H. Avi-Itzhak and T. Diep. Arbitrarily tight upper and lower bounds on the Bayesian probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):89–91, 1996.

[5] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, IT-44(6):2743–2760, 1998.

[6] A. Barucha-Reid. *Elements of the Theory of Markov Processes and their Applications.* McGraw-Hill, New York, 1960.

[7] A. Ben-Israel and T. Greville. *Generalized Inverses: Theory and Applications.* Robert E. Krieger Publishing Company, New York, 1980.

[8] J. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1980.

[9] J. Berger. Bayesian salesmanship. In P. Goel and A. Zellner, editors, *Bayesian Decision and Inference Techniques: Essays in Honor of Bruno de Finetti*, pages 473–488. North-Holland, 1986.

[10] J. Berger and R. Wolpert. *The Likelihood Principle.* Institute of Mathematical Statistics, Hayward, CA, 1984.

[11] J. Bernardo, J. Berger, A. Dawid, and A. Smith (Editors). *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting.* Oxford University Press, Oxford, 1992.

[12] J. Bernardo, J. Berger, A. Dawid, and A. Smith (Editors). *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting.* Oxford University Press, Oxford, 1996.

[13] J. Bernardo, M. DeGroot, D. Lindley, and A. Smith (Editors). *Bayesian Statistics 3: Proceedings of the Third Valencia International Meeting.* Oxford University Press, Oxford, 1988.

[14] J. Bernardo and A. Smith. *Bayesian Theory.* J. Wiley & Sons, Chichester, UK, 1994.

[15] A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269–326, 1962.

[16] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, 1995.

[17] R. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, Reading, MA, 1987.

[18] G. Box and G. Tiao. *Bayesian Inference in Statistical Aanalyis.* John Wiley and Sons, New York, 1973.

[19] L. Brown. *Foundations of Exponential Families.* Institute of Mathematical Statistics, Hayward, CA, 1986.

[20] S. Burch, S. Gull, and J. Skilling. Image restoration by a powerful maximum entropy method. *Computer Vision, Graphics, and Image Processing*, 23:113–128, 1983.

[21] S. Campbell and Jr. C. Meyer. *Generalized Inverses of Linear Transformations.* Pitman, London, 1979.

[22] D. Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, Oxford, 1987.

[23] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[24] R. Cox. Probability, frequency, and reasonable expectation. *American Journal of Physics*, 14, 1946.

[25] M. Crouse, R. Nowak, and R. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46:886–902, 1998.

[26] S. Dalal and W. Hall. Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society (B)*, 45, 1983.

[27] A. Dawid. Invariant prior distributions. In *Encyclopedia of Statistical Sciences, Vol. 2*, pages 228–236. Wiley, New York, 1987.

[28] A. Dawid. Prequential analysis, stochastic complexity, and Bayesian inference. In J. Bernardo, J. Berger, A Dawid, and F. Smith, editors, *Bayesian Statistsics 4: Proceedings of the Fourth Valencia International Meeting*, pages 1019–125. Oxford University Press, 1992.

[29] C. deBoor. *A Practical Guide to Splines.* Springer Verlag, New York, 1978.

[30] B. DeFinetti. *Theory of Probablity.* John Wiley & Sons, New York, 1970.

[31] M. DeGroot. *Optimal Statistical Decisions.* McGraw-Hill, New York, 1970.

[32] H. Derin and P. Kelly. Discrete-index Markov-type random processes. *Proceedings of the IEEE*, 77(10):1485–1510, October 1989.

[33] L. Devroye. *A Course in Density Estimation.* Birkhäuser, Boston, 1987.

[34] L. Devroye, L. Györfy, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, New York, 1996.

[35] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *Annals of Statistics*, 7:269–281, 1979.

[36] P. Dierckx. *Curve and Surface Fitting with Splines.* Oxford University Press, Oxford, 1993.

[37] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[38] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* John Wiley & Sons, New York, 2001.

[39] E. Dynkin. *Theory of Markov Processes.* Pergamon Press, New York, 1961.

[40] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21:194–203, 1975.

[41] M. Figueiredo, J. Leitão, and A. K. Jain. Adaptive B-splines and boundary estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – CVPR'97*, pages 724–729, San Juan (PR), 1997.

[42] R. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Trans. of the Royal Society of London (A)*, 222:309–368, 1922.

[43] R. Fisher. Mathematical probability in the natural sciences. *Technometrics*, 1:21–29, 1959.

[44] B. Frieden. Restoring with maximum likelihood and maximum entropy. *Journal of the Optical Society of America*, G2:511–518, 1972.

[45] K. Fukunaga. *Statistical Pattern Recognition.* Academic Press, Boston, M.A., 1990.

[46] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis.* Chapman & Hall, London, 1995.

[47] G. Golub. *Matrix Computations.* Johns Hopkins University Press, Baltimore, 1989.

[48] I. Good. Weight of evidence, corroboration, explanatory power, information, and the utility of experiments. *Journal of the Royal Statistical Society B*, 22:319–331, 1960.

[49] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, Oxford, 1992. Second edition.

[50] S. Gull and J. Skilling. Maximum entropy method in image processing. *IEE Proceedings (F): Communications, Radar, and Signal Processing*, F-131(6):646–650, 1984.

[51] R. Haralick. Decision making in context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:417–428, 1983.

[52] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, Upper Saddle River, N.J., 1999. 2nd Edition.

[53] C. Helstrom. *Elements of Signal Detection and Estimation.* Prentice Hall, Englewood Cliffs, N.J., 1995.

[54] T. Irony and N. Singpurwalla. Noninformative priors do not exist: A dialogue with José M. Bernardo. *Journal of Statistical Planning and Inference*, 65:159–189, 1997.

[55] D. Isaacson. *Markov Chains: Theory and Applications.* John Wiley & Sons, New York, 1976.

[56] E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4:227–241, 1968.

[57] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70:939–956, 1982.

[58] E. T. Jaynes. *Probability Theory: The Logic of Science.* Available on the World Wide Web, at `http://omega.albany.edu:8008/JaynesBook.html`, 1993.

[59] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society (A)*, 186:453–461, 1946.

[60] H. Jeffreys. *Theory of Probability.* Oxford University Press, Oxford, 1961.

[61] R. Kass and L.Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, September 1996.

[62] R. Kass and P. Vos. *Geometric Foundations of Asymptotic Inference.* John Wiley & Sons, New York, 1997.

[63] S. Kay. *Fundamentals of Statistical Signal Processing. Volume I: Estimation Theory.* Prentice Hall, Englewood Cliffs, N.J., 1993.

[64] S. Kay. *Fundamentals of Statistical Signal Processing. Volume II: Detection Theory.* Prentice Hall, Englewood Cliffs, N.J., 1998.

[65] J. Kemeney and J. Snell. *Finite Markov Chains.* van Nostrand, Princeton, N.J., 1960.

[66] D. Knill and W. Richards (editors). *Perception as Bayesian Inference.* Cambridge University Press, 1996.

[67] S. Kullback. *Information Theory and Statistics.* John Wiley & Sons, New York, 1959.

[68] E. Lee and D. Messerschmitt. *Digital Communications.* Kluwer Academic Publishers, Boston, 1988.

[69] J. Lee, R. Haralick, and L. Shapiro. Morphologic edge detection. *IEEE Journal of Robotics and Automation*, RA-3(2):142–156, April 1987.

[70] E. Lehmann. *Theory of Point Estimation.* John Wiley & Sons, New York, 1983.

[71] E. Lehmann. *Testing Stastistical Hypotheses.* Springer Verlag, New York, 1986. 2nd Edition.

[72] D. Lindley. *Bayesian Statistics: A Review.* SIAM, Philadelphia, PA, 1972.

[73] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, 1998.

[74] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89, March 1987.

[75] G. McLachlan. *Discriminant Analyis and Statistical Pattern Recognition.* John Wiley & Sons, New York, 1992.

[76] J. Melsa and D. Cohn. *Decision and Estimation Theory.* McGraw-Hill, New York, 1976.

[77] D. Middleton. *Am Introduction to Statistical Communication Theory.* IEEE Press, New York, 1990. Originally published by McGraw-Hill, New York, 1960.

[78] J. Moura and N. Balram. Recursive structure of noncausal Gauss-Markov random fields. *IEEE Transactions on Information Theory*, IT-38(2):334–354, March 1992.

[79] M. Murray and J. Rice. *Differential Geometry and Statistics.* Chapman & Hall, London, 1993.

[80] A. Papoulis. A new algorithm in spectral analysis and band-limited extrapolation. *IEEE Transactions on Circuits and Systems*, CAS-22:735–742, 1975.

[81] G. Parisi. *Statistical Field Theory.* Addison Wesley Publishing Company, Reading, Massachusetts, 1988.

[82] S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, April 1997.

[83] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.

[84] H. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, New York, 1988.

[85] M. Powell. Radial basis functions for multivariate interpolation. In J. Mason and M. Cox, editors, *Algorithms for Approximation*, pages 143–167, Oxford, 1987. Clarendon Press.

[86] J. Rawlings, S. Pantula, and D. Dickey. *Applied Regression Analysis: A Research Tool.* Springer Verlag, New York, 1998.

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

[88] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[89] J. Rissanen. Minimum-description-length principle. In *Encyclopedia of Statistical Sciences, Vol. 5*, pages 523–527, New York, 1987. Wiley.

[90] J. Rissanen. *Stochastic Complexity in Stastistical Inquiry.* World Scientific, Singapore, 1989.

[91] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, IT-42(1):40–47, January 1996.

[92] C. Robert. Intrinsic losses. Technical report, Centre de Recherche en Economie et Statistique, Université de Rouen, France, 1993. Available at `ftp.ensae.fr`.

[93] C. Robert. *The Bayesian Choice: A Decision Theoretic Motivation.* Springer-Verlag, New York, 1994.

[94] C. Rodriguez. Bayesian robustness: a new look from geometry. Available at `http://omega.albany.edu:8008/robust.ps`, 1995.

[95] R. Rozenkrantz. *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science.* Reidel, Boston, 1977.

[96] D. Sakrison. *Communication Theory: Transmission of Waveforms and Digital Information.* John Wiley & Sons, New York, 1968.

[97] L. Savage. *The Foundations of Statistics.* Dover, New York, 1954.

[98] L. Scharf. *Statistical Signal Processing.* Addison Wesley Publishing Company, Reading, Massachusetts, 1991.

[99] A. Sen and M. Srivastava. *Regression Ananlysis: Theory, Methods, and Applications.* Springer Verlag, New York, 1997.

[100] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26:26–37, 1980.

[101] B. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London, 1986.

[102] A. Tikhonov, A. Goncharsky, and V. Stepanov. Inverse problems in image processing. In A. Tikhonov and A. Goncharsky, editors, *Ill-Posed Problems in the Natural Sciences*, pages 220–232. Mir Publishers, Moscow, 1987.

[103] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, Chichester (U.K.), 1985.

[104] H. Van Trees. *Detection, Estimation and Modulation Theory*, volume I. John Wiley, New York, 1968.

[105] R. von Mises. *Mathematical Theory of Probability and Statistics.* Academic Press, New York, 1964.

[106] G. Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1990.

[107] J. Whittaker. *Graphical Models in Applied Multivariate Statistics.* John Wiley & Sons, Chichester, UK, 1990.

[108] G. Winkler. *Image analysis, random fields, and dynamic Monte Carlo systems.* Springer-Verlag, Berlin, 1995.

[109] J. Wozencraft and I. Jacobs. *Principles of Communications Engineering.* John Wiley & Sons, New York, 1965.

[110] R. Yang and J. Berger. A catalog on noninformative priors. Technical Report ISDS Discussion Paper 97-42, Institute of Statistical and Decision Sciences, Duke University, Durham, NC, 1997. Available at `http://www.isds.duke.edu/ berger/papers/catalog.html`.

[111] S. Zhu, Z. Wu, and D. Mumford. Minimax entropy principle and its application to texture modelling. *Neural Computation*, 9(8):1627–1660, 1997.

[112] X. Zhuang, E. Østevold, and R. Haralick. The principle of maximum entropy in image recovery. In H. Stark, editor, *Image Recovery. Theory and Applications*, pages 157–193. Academic Press, 1987.