Bayesian Inference Methods Applied to Cancer Research

by

Rudy Gunawan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The purpose of this Thesis is to present a Bayesian analysis of oncological data sets with particular focus on cervical carcinomas and ovarian cancers.

Bayesian methods of data analysis have a very long history, and have been used with great success in many disciplines, from Physics to Econometrics. Nonetheless, they remain very controversial among statisticians who belong to the orthodox - i.e, frequentist school, and are not well known by the medical community. To help in that direction, we reviewed in the introductory chapter the basic philosophical and practical differences between the two schools, and in the second chapter, we briefly reviewed the history of Bayesian methodology, from the early efforts of Thomas Bayes and of Pierre Simon de Laplace to the modern contributions of Harold Jeffreys, Richard Cox, and Edwin Jaynes.

In many aspects of medical research, we deal with experimental data from which a certain proposition or hypothesis is validated. Unlike in physics, where we have strong and solid foundations such as Newton's law of motion, Snell's optical laws, Kirchoff's laws, Einstein's relativity theory, and many more, we do not have such privileges in medical research. Hence, many hypotheses are constantly tested as new evidence becomes available. One of the actively-researched medical areas is cancer research about which our understanding is still in its infancy. Numerous experiments (both in vivo and in vitro) and clinical trials have been conducted to further our knowledge; thus, Bayesian methodology finds its place to aid us in obtaining scientific inferences about certain propositions or hypotheses from available data and resources.

In this work, we use data given to us by our medical collaborators at the Princess Margaret Hospital (PMH) in Toronto to carry out two main projects: Firstly, to make an inference about the oxygenation status (oxygen partial pressure, pO_2) within human cervical carcinomas and secondly, an inference about the effectiveness of various molecularly-targeted agents (MTAs) in phase II clinical trials of relapsed ovarian cancer patients.

In the first problem, we address the challenges of tumor hypoxia - a state of oxygen deprivation in tumors. Currently, there are two methods to obtain tumor oxygen status, namely the direct Eppendorf needle electrode and the indirect immunohistochemical assay of a protein marker, Carbonic Anhydrase IX (CAIX). In this project, we introduce Bayesian probability theory to obtain inferences about tumor oxygenation from each technique and the concordance between the two techniques. From this study, we conclude that under certain conditions, two biopsies are sufficient to infer the tumor oxygenation level based on the immunohistochemical assays of CAIX. Additionally, there is a fair concordance between the direct and the indirect measurements of tumor oxygenation.

In the latter problem, ovarian cancer is the topic of study. Ovarian cancer has the highest mortality rate among gynecological cancers and one that is known to relapse. CA-125 is still the most inexpensive biomarker for monitoring ovarian cancers. From the phase II clinical trial data, we demonstrate the survival advantage of CA-125 responsive group of patients by means of a non-parametric Kaplan-Meier statistic.

Acknowledgements

First, and foremost, I would like to greatly acknowledge Pino Tenti for teaching me how to think and write critically in a scientific reasearch. Additionally, I would like to express my gratitude to Siv Sivaloganathan for giving me an opportunity to explore Bayesian data analysis in the group of Mathematical Medicine.

Second, I would like to acknowledge several medical professionals at the Princess Margaret Hospital in Toronto, Canada for making this Thesis possible: Dr. Vladimir Iakovlev and Dr David Hedley, who shared their data sets on immunohistochemical assays of CAIX, Dr. Michael Milosevic, whose clinical insights on the pO_2 measurements are appreciated, Dr. Amit Oza, who invited us to join the project for phase II clinical trials of relapsed ovarian cancers, and Lisa Wang, from whom I learn most of the statistics for clinical trials.

Third, last but not least, I kindly appreciate the critiques from Professor Devinder Sivia and Professor Mary Thompson and their time in reading this Thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This Thesis provides an account of Bayesian data analysis applied to oncological research with particular focus on cervical carcinomas and ovarian cancers.

The data were obtained in a prospective study that evaluated hypoxia in patients with cervix cancer treated at the Princess Margaret Hospital (PMH) in Toronto, Canada. Patient selection criteria, measurements of tumor oxygenation, and tissue handling were performed according to the methods described in the literature (Fyles et al., 2002; Hedley et al., 2003; Iakovlev et al., 2007).

All living tissues need a certain amount of oxygen for the cells to function well, and if its concentration has an optimum value the tissue is called *oxic*. However, for various reasons some regions of the tissue may have a lower than normal concentration of oxygen; in that case, the tissue is said to be *hypoxic*. Obviously, the oxygen is carried to the cells by the blood microcirculation, and since tumor cells tend to be tightly packed together, tumor vasculature tends to be highly chaotic, and it is very likely that there will be regions where cells do not receive adequate oxygenation.

It has been known for a long time that oxygen concentration levels within a tumor have a significant impact on the effectiveness of radiation therapy (Hall, 1994), and it may also be the case that oxygen is "the most important determinant of response [to radiotherapy] among tumors of the same type" (Harrison et al., 2002).

In fact, it is considered an established fact among medical researchers that hypoxia is "related to poor response to radiation and chemotherapy, genetic instability, selection for resistance to apoptosis, and increased risk of invasion and metastasis" (Fyles et al., 2002; Vaupel and Harrison, 2004), and more recent studies give further confirmation of these findings (Vaupel and Mayer, 2007; Tatum et al., 2006).

Given this knowledge, it is quite understandable that the assessment of tumor hypoxia has become a central concern of cancer researchers. The *gold standard* of tumor hypoxia assessment is a direct measurement of pO_2 (oxygen tension) *in vivo* by the Eppendorf polarographic electrode (Fatt, 1976), which is an invasive technique restricted to accessible biological sites and samples ($\sim$ 100-150 points) within the tumor. An attractive alternative is immunohistochemical assay (staining) to detect proteins expressed by cells during hypoxia. Carbonic anhydrase IX (CAIX) is an enzyme expressed on the cell membrane during hypoxia in response to balance the immediate extracellular microenvironment, and is widely regarded as a surrogate marker of chronic hypoxia in various cancers (Loncaster et al., 2001; Hoskin et al., 2003; Mseide et al., 2004; Olive et al., 2001; Hoogsteen et al., 2005).

The study conducted by the PMH group produced data on hypoxia assessment by means of both the Eppendorf electrode and by CAIX staining, and have been published along with a statistical analysis carried out with the well-established techniques of conventional (orthodox) statistics (Fyles et al., 2002; Iakovlev et al., 2007). Despite the use of quite sophisticated techniques, this analysis left practically unresolved one of the most important objectives of the study, namely whether CAIX staining can be used with confidence in assessing tumor hypoxia, thus making the Eppendorf polarographic electrode obsolete.

It is, therefore, clearly of interest to re-analyze the data with a different methodology, such as the Bayesian approach taken in this Thesis. With the availability of today's powerful computers, Bayesian data analysis has become more and more attractive to researchers in various disciplines. Thus, for example, this approach has been popular in Econometrics (Zellner, 1971), in the analysis of Nuclear Magnetic

Resonance (NMR) spectra (Bretthorst, 1988), in astrophysical studies (Gregory, 2005), in image processing (Sivia, 2006), and in many other areas. In medical research, however, Bayesian methods are not yet widely used, though there are clear signs of appreciation of its advantages and power (Berry, 2006, 2005), as well as strong proponents of the Bayesian approach against the blind reliance on P-values and hypothesis testing (Goodman, 1999, 2005; Goodman and Sladky, 2005). Nonetheless, Bayesian methods are still considered non-orthodox and controversial, and it is therefore important to have a clear grasp of why this controversy still exists.

The fundamental difference between the two schools of thought rests on the interpretation of *probability*. According to the conventional or orthodox school, probability theory can only be applied to random variables. A classic example is the number on the up face of a rolled fair die: Any of the integers 1, 2, ... , 6 can be the outcome. When the die is rolled n times under the same conditions, the result will be a sequence of n outcomes, possibly not all different, and if the order in which the various outcomes occurred is not of interest, it is often convenient to summarize such a data sequence in terms of *relative frequencies* of outcomes. Experiments of this type have several possible outcomes, which are totally *random* in the sense that it is impossible to say in advance which one will occur. However, experience has shown that in many repetitions of the experiment the relative frequencies with which the various outcomes occur will stabilize and tend to a fixed values. Although it is impossible to predict which face will show up when the fair die is rolled just once, we can say with some confidence that in a large number of rolls each face will show up about $\frac{1}{6}$ of the time.

From this point of view, *probabilities are defined as the limiting values approached by relative frequencies as* $n \to \infty$. The probability of an outcome is then the fraction of the time that the outcome would occur in infinitely many repetitions of the experiment. In other words, probabilities are approximated, or estimated, by the corresponding relative frequency, and since the latter are measured it follows

that probabilities are objective quantities like masses, lengths, temperatures, and so on.

The restriction of probability to random variables poses serious problems in data analysis. In most cases, scientists want to use data to make inferences about the values of unknown parameters entering the model, or theory, which is supposed to accurately describe the phenomenon under study. But these parameters are constants, not random variables; hence, probability theory cannot be directly applied to them. In order to cope with this problem, mathematicians created a new subject – Statistics. To estimate a parameter, one must first relate it to the data through a function of the data called *a statistic*; and since the data are subject to noise, the statistic becomes the random variable to which the rules of probability theory can be applied. This, of course, poses immediately another serious problem: How should one choose the statistic ? Most of the last century saw statisticians occupied with the solution of this problem. The masters, such as Fisher, Neyman, and Pearson, created a variety of different principles, which in the hands of others resulted in a plethora of tests and procedures without clear underlying rationale.

The Bayesian school of thought, on the other hand, sees probability in a completely different light. To pioneers like Bayes and above all Laplace, a probability $P(A|E)$ represented a *degree-of-belief*, or *plausibility*, of the truth of a proposition A given the evidence E at hand. Thus all probabilities are *conditional*, in contrast to the frequentist interpretation where probabilities are absolute. To the 19^{th} century scholars this seemed too vague and subjective an idea to be the basis of a rigorous mathematical theory, and Laplace's uses of probability theory were scorned for over a century. (A brief review of the history of this conflict, up to the present, is given in chapter two.)

A second point of contention between the two schools was Laplace's use of Bayes' theorem as the main tool for making inductive inferences from the data and the available prior information. Due to the fact that Bayesian probability is *not necessarily* a frequency, but rather a measure of the plausibility of the truth of

a proposition, this point of view allowed Laplace to solve scientific problems that were beyond the capability of the frequentists, such as his famous estimate of the mass of Saturn (which is recounted in chapter two). Thus Laplace was taking the first steps toward the use of probability theory as *generalized logic of science*, as spelled out by Edwin Jaynes in his recent book (Jaynes, 2003).

The most important contribution to this view of probability was given by Richard Cox in the middle of the last century (Cox, 1946, 1961). Instead of getting involved in the controversy about whether or not Laplace gave us the right "calculus of inductive reasoning", Cox took the constructive path of asking if such a calculus is possible in the first place, and if so, whether or not there is a consistent set of mathematical rules for carrying out plausible, rather than deductive, reasoning. He started out with a few simple and common-sensical desiderata (axioms) on which all reasonable people can agree. Firstly, given propositions A, B, and C, the plausibilities of their truth within a context K must possess the *transitive property*: If A is more plausible than B, and B is more plausible than C, then we must assert that A is more plausible than C. To do otherwise would lead us to argue in circles. This can be simply accomplished by assigning a real number – in the interval $[0, 1]$ for convenience – to each proposition; the larger the number $P(A|K)$ is, the higher the plausibility that A is true. Of course, the interval endpoints represent certainty, namely, $P(A|K) = 0$ means that A is false whereas $P(A|K) = 1$ means that A is true. A second straightforward assumption is that once the plausibility $P(A|K)$ that A is true has been specified, the plausibility $P(\bar{A}|K)$ that A is false is also specified automatically. Thirdly, if the plausibility $P(B|K)$ that B is true is specified, and then the plausibility that A is true *given* that B is true $P(A|B, K)$ is also specified, then the plausibility that *both* A and B are true $P(AB|K)$ is also automatically specified.

Cox did not prescribe the functional relations between the various plausibilities, but only assumed their existence. Then, using Boolean algebra and the constraint of *consistency* (in the sense that if two different methods of calculation are permitted

by the rules, then they should yield the same result), he showed rigorously that the consistency conditions take the form of two functional equations, whose general solutions he found. Those solutions uniquely determine the two following rules: $P(A|K) + P(\bar{A}|K) = 1$, and $P(AB|K) = P(A|B,K) \times P(B|K)$. By mathematical transformations one can, of course, change *the form* of these rules; but what Cox proved is that any change of their *content* will produce inconsistencies, in the sense that two methods of calculations, each permitted by the rules, will yield different results. But the two rules above are just the sum and product rules of the mathematical theory of probability, and all other equations needed for applications can be derived from them. Therefore, *the plausibility $P(A|K)$ is nothing but the probability that A is true given the context K*. The proofs of all these results are given in the references cited above, and a simplified modern version of them is presented in Appendix B of (Sivia, 2006).

One thing missing from the outstanding work of Cox was a theory of how to derive from fundamental logical principles the rules for assigning the prior probabilities that are necessary to use Bayes' theorem as the basic tool for making scientific inferences. Fortunately, soon afterwards the seminal work of Shannon in information theory and the realization by Jaynes of its relevance to the logical assignment of priors, resulted in the Maximum Entropy (MaxEnt) algorithm that put to rest the main objections to Bayesian methodology. Their contributions – along with those of Harold Jeffreys – are reviewed in chapter two, to which the reader is referred.

Chapter three gives a rather brief review of the principles of Bayesian inference, with particular regard to parameter estimation and model selection. Then, in chapter four, we begin with the applications of Bayesian methodology for inferring tumor oxygenation status within human cervical carcinomas. This work has been done in collaboration with Dr. Vladimir Iakovlev (at the Keenan Research Centre, St. Michael's Hospital) and Dr. Michael Milosevic (at the Radiation Medicine Program, Princess Margaret Hospital). In chapter five, we introduce Bayesian methods to infer the effectiveness of various molecularly-targeted agents (MTAs)

from four phase II clinical trials of relapsed ovarian cancer patients. This project has been done in collaboration with Dr. Amit Oza (of the Drug Development Program, Princess Margaret Hospital). Finally, we conclude the Thesis with a critical discussion of the results of the Bayesian analysis versus those of the orthodox statistical methods, and point out future useful directions for application of the Bayesian methodology.

Chapter 2

A brief history of Bayesian methodology

It is convenient, for completeness' sake, to start out with a brief historical review of the developments of Bayesian methods of data analysis.

2.1 Bayes and Laplace

Traditionally, the starting point is associated with the paper of a British clergyman (and amateur mathematician), Thomas Bayes (1763), but it was really Pierre Simon de Laplace who gave impetus to the use of Bayes' theorem as the ideal tool for making inferences from a data set. In his famous memoir on the "probabilities of causes" (Laplace, 1774), he had already pointed out the fundamental difference between pure mathematics and scientific reasoning. The former relies on logical deduction as the tool for making inferences, with the result that there is no room for propositions of the type "This is *probably* the cause of that event". The latter, on the other hand, is concerned with the problem of inferring the probable cause of an observed event or phenomenon, as schematically indicated in Figure 2.1, and no certainty can ever be produced in this process. As already pointed out long ago by Hume, no amount of observations of the sun rising in the East in the past can

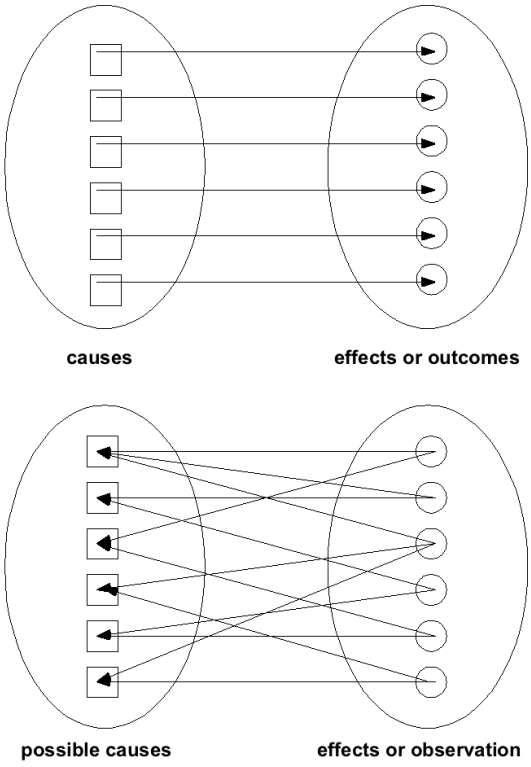

Figure 2.1: Schematic process of deductive logic (top) versus inductive logic (bottom)

be used to deduce that it will rise in the East tomorrow as well. The best we can do in Science is to learn from our past experience and infer from that and any new data the most probable cause of the observed phenomenon.

The manner in which Laplace used probability theory for making inductive inferences about causes is succinctly summarized by Edwin Jaynes (Jaynes, 1978) as follows. Suppose E is an observable event which could result from a set (C_1, C_2, ... , C_N) of conceivable causes. Suppose also that we have found, according to some model, the "sampling distribution" or "direct" probabilities of E for each cause: $P(E|C_i)$, $i = 1, 2, ..., N$. Finally, assume that initially the possible causes C_i are equally likely, i.e, assume a uniform prior distribution of C_i. Then, says Laplace, the "inverse" probability $P(C_i|E)$, i.e, the posterior probability, is given by

$$P(C_i|E) = \frac{P(E|C_i)}{\sum_{k=1}^{N} P(E|C_k)}, \tag{2.1}$$

which Laplace used with great success in solving many problems in astronomy, meteorology, geodesy, population statistics, and more.

Later on, Laplace generalized (2.1) by noting that if initially the causes cannot be considered equally likely, but have prior probabilities $P(C_i|I)$, where I stands for the prior information, then the terms in (2.1) should be weighted according to these priors to give

$$P(C_i|E, I) = \frac{P(E|C_i, I)P(C_i|I)}{\sum_{k=1}^{N} P(E|C_k, I)P(C_k|I)}, \tag{2.2}$$

which, following a long-established custom, is always referred to in the literature as "Bayes' theorem".

Despite Laplace's success in using (2.1), there were clear difficulties with this approach. Firstly, in many practical problems, it is not possible to list a finite number of causes in order to apply the so-called "principle of insufficient reason". This had already been noticed by James (Jakob) Bernoulli, who in his posthumously published *Ars Conjecturandi* (1713) remarked that

> " ... this may be done only in a very few cases and almost nowhere other than in games of chance ... But what mortal will ever determine, for example, the number of diseases – these and other such things depend upon causes completely hidden from us –".

Secondly, Laplace never derived (2.1) and (2.2) from fundamental logical principles, but simply stated them as intuitive recipes, thus leaving room for doubt as to their logical justification and uniqueness. Finally, Laplace was unable to tell how the prior probabilities $P(C_i|I)$ in (2.2) were to be assigned in a non-arbitrary manner, leaving room for the critics to attack him on the grounds of non-objectivity, and hence non-scientificity, of his results.

2.2 The frequentist reaction

In view of the great practical achievements by Laplace with (2.1), it would seem reasonable to expect that the difficulties mentioned above should have stimulated the next generation of statisticians to build constructively on the foundations laid down by him; in particular, they should have focused their attention on seeking new and more general principles for determining prior probabilities. Instead, what followed was a concerted series of increasingly violent attacks on Laplace's works, some of which can be found in the works of Ellis (1842), Boole (1854), Venn (1866), and von Mises (1928); they were so successful in their opposition that the Bayes/Laplace uses of probability theory laid discredited for a century.

The objectives voiced by these writers were not directed against Bayes' theorem *per se*, for this theorem follows trivially from the product and sum rules of probability theory, namely

$$P(AB|C) = P(A|BC) \times P(B|C), \tag{2.3}$$

and

$$P(A|B) + P(\bar{A}|B) = 1, \tag{2.4}$$

where

- A, B, and C can be any event or proposition;
- $AB|C$ denotes proposition A and B are both true given that proposition C is true; and finally,
- $\bar{A}$ stands for proposition A is false.

Now by the product rule, we have

$$\begin{aligned} P(AB|C) &= P(A|BC) \times P(B|C), \\ P(BA|C) &= P(B|AC) \times P(A|C), \end{aligned}$$

and, since $P(AB|C)$ must equal $P(BA|C)$, subtracting term by term gives

$$0 = P(A|BC) \times P(B|C) - P(B|AC) \times P(A|C),$$

or else

$$P(A|BC) = \frac{P(B|AC) \times P(A|C)}{P(B|C)}, \quad \text{provided } P(B|C) > 0, \tag{2.5}$$

which is just Bayes' Theorem. As long as A, B, and C refer to random variables and the probability is interpreted as frequency, as explained in the Introduction, the orthodox statisticians have no objections to using (2.5). But that is not the way Bayes/Laplace intended it. To see this, I can do no better than to follow the explanation given by Sivia in his delightful introductory textbook (Sivia, 2006), namely to read (2.5) as

$$P(\text{hypothesis}|\text{data}, I) = \frac{P(\text{data}|\text{hypothesis}, I) \times P(\text{hypothesis}|I)}{P(\text{data}|I)}, \tag{2.6}$$

where

- $P(\text{hypothesis}|I)$ is referred to the 'prior probability'; it is our state-of-knowledge before we analyze the data or before we do any experiment regarding the hypothesis at hand.

- $P(\text{data}|\text{hypothesis}, I)$ is the 'direct probability', also known as the 'likelihood function'. It is the implication of the hypothesis on the data.
- $P(\text{data}|I)$ is the 'evidence'. It only acts as a normalization constant in the *parameter estimation* case; however, it plays a crucial role in the *model selection* case. This is the evidence representing the 'merit' of a particular hypothesis.
- $P(\text{hypothesis}|\text{data}, I)$ is referred to the 'posterior probability'; it represents our updated state-of-knowledge in light of the new data.

Thus, in (2.6), we have a mathematical representation of the process of learning; that is, of the scientific method. In the application of (2.5), one of the outstanding successes of Laplace is recounted by Jaynes (Jaynes, 1985) as follows:

> "For example – a famous example that Laplace actually did solve – proposition A might be the statement that the unknown mass M_S of Saturn lies in a specified interval, B the data from observatories about the mutual perturbations of Jupiter and Saturn, C the common sense observation that M_S cannot be so small that Saturn would lose its rings; or so large that Saturn would disrupt the solar system. Laplace reported that, from the data available up to the end of 18th Century, Bayes' theorem estimates M_S to be (1/3512) of the solar mass, and gives a probability of 0.99991, or odds of 11,000:1, that M_S lies within 1% of that value. Another 150 years' accumulation of data has raised the estimate 0.63 percent."

It goes without saying that this failed to impress the statisticians of the frequentist school. In the first place, they pointed out, we cannot use probability theory in estimating the mass of Saturn since it is a constant, not a random phenomenon; hence, it cannot have a frequency distribution. In the frequency theory, one has to imagine hypothetical *ensembles* of universes in which everything remains identical except the mass of Saturn. Then, one has to relate the mass of Saturn to the observations through a function called a *statistic*. The data are obtained from some experiments; hence, they are subject to measurement error, which is regarded as 'random noise'. As a consequence, the statistic becomes a random variable to which the rules of probability can be applied. In the second place, they strongly objected to the use of the prior, for if probability is just a measure of our state of

knowledge then it cannot be objective, since different people will assign different priors; therefore, objective (scientific) knowledge would not be possible to achieve.

2.3 Jeffreys and Cox

It was not until the middle of the 20$^{\text{th}}$ century that Laplace's work in probability theory was taken up and greatly expanded. In 1939, Harold Jeffreys published his book entitled *Theory of Probability*, which saw several editions and reprints (Jeffreys, 1958), and contained a strong defense of plausible, i.e, inductive reasoning, as the true scientific method. Jeffreys pointed out that just as deduction must start from axioms that must be accepted without proof, so induction must start from a set of **rules of reasoning** that we cannot prove a priori but can only be justified a posteriori. He then listed the following five rules that in his view are *essential* (Jeffreys, 1958):

1. All hypotheses used must be explicitly stated and the conclusions must follow from the hypotheses.
2. The theory must be self-consistent; that is, it must not be possible to derive contradictory conclusions from the postulates and any given set of observational data.
3. Any rule given must be applicable in practice. A definition is useless unless the thing defined can be recognized in terms of the definition when it occurs. The existence of a thing or the estimate of a quantity must not involve an impossible experiment.
4. The theory must provide explicitly for the possibility that inferences made by it turn out to be wrong. A law may contain adjustable parameters, which may be wrongly estimated, or the law itself may be afterwards found to need modification – the relativity and quantum theories providing conspicuous instances – and there is no conclusive reason to suppose that any of our present laws are final. But we do accept inductive inference in some sense; we have a certain amount of confidence that it will be right in any particular sense, though this confidence does not amount to logical certainty.
5. The theory of induction must not deny any empirical proposition a priori; any precisely stated empirical proposition must be formally capable of being accepted in the sense of the last rule, given a moderate amount of relevant evidence.

In addition to the five essential rules, Jeffreys stated three more "useful guides":

6. The number of postulates should be reduced to a minimum.

7. Although we do not regard the human mind as a perfect reasoner, we must accept it as a useful one and the only one available. The theory need not represent actual thought processes in detail but should agree with them in outline.

8. In view of greater complexity of induction, we cannot hope to develop it more thoroughly than deduction. We therefore take it as a rule that an objection carries no weight if an analogous objection invalidates part of generally accepted pure mathematics.

Note that number six is essentially a statement of *Ockham's razor*: the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory. Furthermore, the seventh "useful guide" shows that a theory of induction must agree with common sense in evaluating propositions about empirical phenomena.

In addition to formulating rules for induction or plausible reasoning, Jeffreys attempted to rebut the accusation of subjective prior assignment in Bayesian methodology. It is not at all arbitrary but required by logical consistency to represent a state-of-knowledge *before* analyzing the data. Prior probability represents our 'ignorance' apriori. At any state of knowledge it is legitimate to ask about a given proposition or hypothesis, 'How do you know ?' The answer will usually depend on observational data. If we ask further, 'What did you think of the proposition or the hypothesis before you had these data ?' we may be told of some less convincing data. If we ask further questions, we shall reach a state where the answer must be: 'I thought the matter worth considering, but had no opinion about whether it was true' (Jeffreys, 1958). In the event, when we do not have any information regarding the value of a certain proposition or hypothesis, we can assign equal probabilities. They are not in any way an assertion that they must occur equally often in any 'random experiment'. As Jeffreys emphasized, taking a uniform or equal probabilities prior is "not a statement of any belief about the actual composition of the

world, nor is it an inference from previous experience; it is merely the formal way of expressing ignorance" (Jeffreys, 1958).

No matter how reasonable Jeffreys' rules appear to open-minded people, the unfortunate fact is that he did not succeed in deriving them logically from any compelling set of desiderata. The result was that the adherents of the frequency school, i.e, the majority of mathematicians and scientists of the time, attacked Jeffreys with the same acrimony as they had reserved for Laplace, and Bayesian methodology kept being shunned for many more years.

This state of affairs did not change even after Richard Cox published what is perhaps the most important work in the history of Bayesian Probability. In an unpretentious little paper (Cox, 1946), he avoided getting involved in the debate about whether or not it is appropriate to use probability as a measure of degree of plausibility for the truth of a proposition, and focused instead on asking the following question: Is it possible to construct a consistent set of mathematical rules for carrying out plausible, rather than deductive, reasoning ? To find out, he assumed the following desiderata:

1. Degrees of plausibility are represented by real numbers;
2. degrees of plausibility must be in qualitative correspondence with common sense – i.e., with rationality; and finally,
3. if a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

Then, using Boolean algebra, he proved that the consistency conditions took the form of two functional equations whose general solution he was able to find. Those solutions determined the rules (2.3) and (2.4) to within a change of variables that can alter their form but not their content. This result was summarized by Jaynes (Jaynes, 1985) as follows:

> "So, thanks to Cox, it was now a theorem that any set of rules for conducting inference, in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to the Laplace-Jeffreys

rules, or inconsistent. The reason for their pragmatic success is then pretty clear. Those who continued to oppose Bayesian methods after 1946 have been obliged to ignore not only the pragmatic success, but also the theorem."

2.4 Shannon and Jaynes

The full scope and generality of Bayesian inference had been already recognized by Jeffreys, and Cox's theorem legitimized what he had been doing all along. But the main bone of contention with the frequentists, namely the alleged subjectivity of Bayesian methods, was still at issue. In particular, there was still no logically derived method to *assign* prior probabilities with scientific objectivity.

A breakthrough occurred soon after the appearance of the fundamental paper *The Mathematical Theory of Communication* by Claude Shannon in 1948 (reprinted in (Shannon, 1971)), which laid the foundations of the new field of *Information Theory*. The communication system considered by Shannon consists of several elements:

- An *information source*, which selects a message from a set of possible messages;
- A *transmitter*, which changes the message into a *signal*;
- A *communication channel*, such as a wire, which carries the signal; and
- A *receiver*, which changes the transmitted signal back into a message and hands the message on to the *destination*.

The kind of questions which Shannon sought to answer were fundamental to communication theory. First of all, he asked how we can measure the *amount of information*. Secondly, how can information be *coded* efficiently and how can we measure the *channel capacity*. Thirdly, and importantly, how do we handle the *noise* which inevitably will try to corrupt the message. Finally, Shannon pointed out that *information* should not be confused with meaning; in his words, "the semantic aspects of communication are irrelevant to the engineering aspect." Now

in a communication process, we have N messages m_1, m_2, ... , m_N from which we are free to choose the one to be sent. Suppose we assign probabilities P_1, P_2, ... , P_N to the messages; these probabilities are known, but that is all we know about which message will be chosen. Then, asked Shannon, can we find a measure of how much "choice" is involved in the selection of the message or of how uncertain we are of the outcome ? (Shannon, 1971). Calling such a measure $H(P_1, P_2, ..., P_N)$ and requiring it to obey some reasonable mathematical constraints, such as continuity in the P_i's and consistency, Shannon proved that the only such measure has the form (Theorem 2 of (Shannon, 1971))

$$H = -K \sum_{i=1}^{N} P_i \log P_i, \tag{2.7}$$

where K is a positive constant. He then went on to stress that quantities of the form $H = -\sum_{i=1}^{N} P_i \log P_i$ play a central role in information theory as measures of *information*, *choice*, and *uncertainty*. Since this form is just the expression for entropy in statistical mechanics, it will be called *the entropy of the probability distribution* P_i.

Next Shannon considers the problem of encoding a message into binary digits in the most efficient way. The most important step is to assign probabilities to each conceivable message, which in general will not be possible. For example, he says,

> "If a source can produce only one particular message its entropy is zero, and no channel is required. For example, a computing machine set up to calculate the successive digits of π produces a definite sequence with no chance element. No channel is required to 'transmit' this to another point. One could construct a second machine to compute the same sequence at the point. However, *this may be impractical.* In such a case we can choose to ignore some or all of the statistical knowledge we have of the source. We might consider the digits of π to be a random sequence in that we construct a system capable of sending any sequence of digits. In a similar way we may choose to use some of our statistical knowledge of English in constructing a code, but not all of it. *In such a case we consider the source with the maximum entropy subject to the statistical conditions we wish to retain.* The entropy of this source determines the

> channel capacity which is necessary and sufficient. In the π example the only information retained is that all of the digits are chosen from the set 0, 1, ... , 9. In the case of English one might wish to use the statistical saving possible due to letter frequencies, but nothing else. The maximum entropy source is then the first approximation to English and its entropy determines the required channel capacity." (Shannon, 1971)

Shannon does not show mathematically what this suggested calculation method produces, nor does he discuss the interpretation of the probability distribution P_i, but simply moves on to other matters.

It was Edwin Jaynes that realized the importance of Shannon's suggestion towards the solution of the old problem of assigning prior probabilities in the most objective way possible. In his words,

> "The great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the 'amount of uncertainty' represented by a discrete probability distribution, which agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one, and satisfies all other conditions which make it reasonable. ... It is now evident how to solve our problem; in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have." (Jaynes, 1985)

Furthermore, as already noted by Shannon, the entropy H is maximum when all P_i's are equal and no information is given except for the enumeration of all possible outcomes. Thus, Jaynes pointed out that the Maximum Entropy Principle (MaxEnt) contains as a particular case the principle of insufficient reason, but with the following essential difference:

> "The maximum entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise. Thus, the concept of

entropy supplies the missing criterion of choice which Laplace needed to remove the apparent arbitrariness of the principle of insufficient reason, and in addition it shows precisely how this principle is to be modified in case there are reasons for 'thinking otherwise'. " (Jaynes, 1985).

Chapter 3

The basics of Bayesian inference

Fundamental to Bayesian methodology is learning to revise a probability representing a degree-of-belief in the truth of a certain proposition in light of new information or empirical data through Bayes' theorem. This process is also referred to as the *principle of inverse probability* and is schematically illustrated in Fig. 3.1. As explained in the introduction chapter of this Thesis, probability is defined as a degree-of-belief or plausibility of truth of a certain proposition, and probability is always conditional on our background information. As should be clear from Fig. 3.1, Bayesian reasoning is a gradual learning process.

The two aspects of Bayesian analysis that are reviewed here are those of pa-

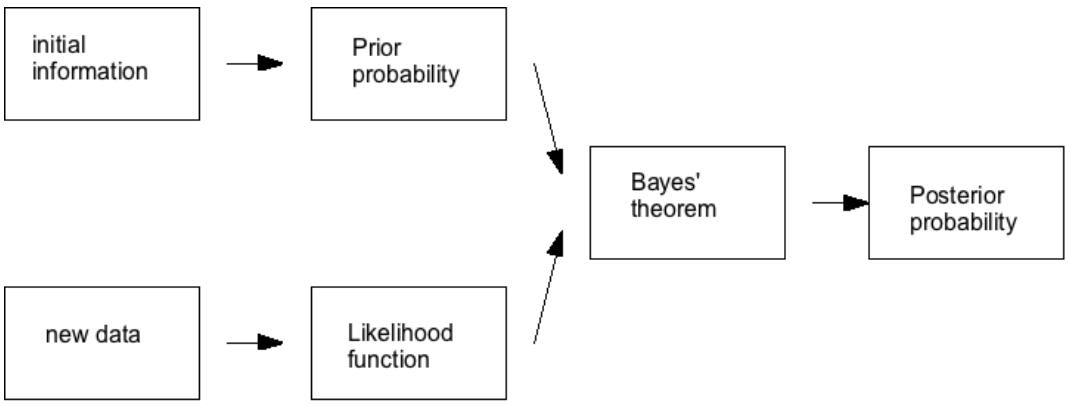

Figure 3.1: The process of revising probabilities. Redrawn from (Zellner, 1971)

rameter estimation and model selection. The former is the simplest to explain; therefore, we will begin with the application of Bayesian methods in estimating a parameter. The latter is important when one wants to compare various hypotheses or propositions about a certain phenomenon.

3.1 Parameter estimation

Consider for a illustration purposes that we are concerned about a parameter θ, which is assumed to be a continuous variable in a finite interval. Following the scheme in Fig. 3.1, we begin the Bayesian analysis by quantifying our knowledge about θ through the prior probability. We may be aware of a plausible value of θ from previous experience, either theoretically or experimentally. If not, we shall not despair. We can use a prior probability that represents total ignorance, namely

$$P(\theta|I) = \begin{cases} \frac{1}{[\theta_{\max}-\theta_{\min}]}, & \text{for } \theta_{\min} \leq \theta \leq \theta_{\max} \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

where $\theta_{\max}$ and $\theta_{\min}$ are the minimum and the maximum value for θ respectively. In any physical application, the parameter θ is bound either by theoretical or practical limits. This type of prior probabilities is also known as the *flat* prior. It basically states that we do not know which particular values of θ are more plausible; equivalently, it assigns equal probability to all values that θ can take. Note, however, that *this does not mean that those values must occur equally often in any "random experiment"*, as the frequentists would have it; *rather*, as emphasized by Jeffreys, *it is only a formal way of saying "I don't know"*

Next we must calculate the likelihood function, $P(D|\theta, I)$, which is basically the implication of our model of $f(\theta_i)$ on the data (D). We can model the data d_i in general as

$$d_i = f(\theta_i) + \epsilon_i, \qquad i = 1, 2, ..., N \tag{3.2}$$

where ϵ_i is the error parameters, commonly referred to as noise, and $f(\theta_i)$ represents

the functional form of the model. For clarity purposes, let us assume that we have a constant model $f(\theta_i) = \theta$ $(\forall i)$ for the data,

$$d_i = \theta + \epsilon_i; \tag{3.3}$$

in scientific experiments, the data may be thought of consisting a constant signal θ and a noise ϵ_i. In Bayesian probability theory, one need not make assumption about the sampling distribution of the noise; rather, one assigns what is actually known about the noise. In other words, we assign to the noise a prior distribution which is consistent with what is known about it. This prior should be as uninformative as possible, as a precaution against "seeing" things in the data which are not there. It can be shown by means of the Maximum Entropy (MaxEnt) principle that *the least informative* prior probability of the errors ϵ_i is given by (Bretthorst, 1988),

$$P(\epsilon_i|s_i, I) = \frac{1}{\sqrt{2\pi}s_i} \exp\left\{-\frac{\epsilon_i^2}{2s_i^2}\right\}, \tag{3.4}$$

where s_i^2 is the second moment of the noise. Now from Eq. 3.3 we may write the error, which is the difference between the assumed model ($f(\theta_i) = \theta$) and the actual data (d_i), in the form $\epsilon_i^2 = (\theta - d_i)^2$, and so the direct probability that we should obtain the data $D = \{d_1, d_2, ..., d_N\}$, given the parameters, will give us the likelihood function (proportional to the direct probability) as

$$P(d_i|\theta, I) = \frac{1}{\sqrt{2\pi}s_i} \exp\left(-\frac{\chi_i^2}{2}\right), \tag{3.5}$$

where $\chi_i^2 = \left(\frac{d_i - \theta}{s_i}\right)^2$ is the mismatch between the data d_i and the model θ. Assuming that each datum is independent, we have

$$\begin{aligned} P(D|\theta, I) &= \prod_{i=1}^{N} P(d_i|\theta, I), \\ &= (2\pi)^{-\frac{N}{2}} \frac{1}{\prod_{i=1}^{N} s_i} \exp\left(-\sum_{i=1}^{N} \frac{\chi_i^2}{2}\right). \end{aligned} \tag{3.6}$$

Then the application of Bayes' theorem gives us our new knowledge about the parameter θ in light of the data encapsulated in a posterior probability density function (pdf),

$$P(\theta|D,I) = \frac{P(D|\theta,I)\,P(\theta|I)}{P(D|I)}. \tag{3.7}$$

Since this is a parameter estimation case, we may absorb the evidence, $P(D|I)$, into the normalization constant, resulting in

$$\begin{aligned} P(\theta|D,I) &\propto P(D|\theta,I)\,P(\theta|I), \\ &= \left[(2\pi)^{-\frac{N}{2}}\,\frac{1}{\prod_{i=1}^{N}\,s_i}\,\exp\left(-\sum_{i=1}^{N}\,\frac{\chi_i^2}{2}\right)\right]\,\left[\frac{1}{[\theta_{\max}-\theta_{\min}]}\right], \\ &\propto \exp\left(-\sum_{i=1}^{N}\,\frac{\chi_i^2}{2}\right); \end{aligned} \tag{3.8}$$

from which the best estimate of θ is found by maximizing this probability density function (pdf). Obviously, for this to happen we must have

$$\left.\frac{d[P(\theta|D,I)]}{d\theta}\right|_{\theta=\hat{\theta}} = 0\,, \tag{3.9}$$

where $\hat{\theta}$ represents the desired maximum, along with a negative second derivative. In most cases, it is much simpler to deal with the (natural) logarithm of the pdf (Sivia, 2006),

$$L = \ln[P(\theta|D,I)] \propto -\frac{1}{2}\sum_{i=1}^{N}\,\chi_i^2. \tag{3.10}$$

Expanding about its maximum $\theta = \hat{\theta}$ and ignoring terms higher than quadratic, we obtain an approximation for the pdf,

$$P(\theta|D,I) \approx A\,\exp\left[\frac{1}{2}\,\left.\frac{d^2L}{d\theta^2}\right|_{\theta=\hat{\theta}}\,(\theta-\hat{\theta})^2\right], \tag{3.11}$$

where A is a normalization constant. We therefore approximate the posterior pdf with the well-known Gaussian distribution.

The simplest case is when all the error-bars can be taken to be equal - i.e, when

$s_1 = s_2 = ... = s$. Then Eq. 3.11 can be written as

$$P(\theta|D, I) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta - \hat{\theta})^2}{2\sigma^2}\right\}, \tag{3.12}$$

where we have adopted the notation $\sigma = \left(\frac{d^2L}{d\theta^2}\Big|_{\theta=\hat{\theta}}\right)^{-\frac{1}{2}}$. The posterior pdf thus has the classical bell-shaped form and the uncertainty can be represented by the full width at half maximum (FWHM). For a Gaussian pdf, it can be shown easily that the FWHM is $2 \times \sqrt{2 \ln 2}\sigma \approx 2.35\sigma$ (Sivia, 2006). In such a case, we can summarize the result of our inference by the expression

$$\theta = \hat{\theta} \pm \sigma. \tag{3.13}$$

In general, however, it may not be possible to encapsulate the best estimate from the posterior pdf in two simple numbers such as $\hat{\theta}$ and s. For an asymmetric pdf, it is not appropriate to use the FWHM of the posterior pdf to quantify the uncertainty of the estimate; however, we may utilize a *credibility interval* to express our uncertainty, namely

$$P(\theta_1 \leq \theta \leq \theta_2|D, I) = \int_{\theta_1}^{\theta_2} P(\theta|D, I)\, d\theta \approx 1 - \alpha, \tag{3.14}$$

where $\theta_2 - \theta_1$ is as small as possible. The region $\theta_1 \leq \theta \leq \theta_2$ is called the shortest $1-\alpha$ credibility interval, where α is usually taken as 0.05. We can then proceed with the Lagrange's optimization technique to find θ_1 and θ_2, by solving the straightforward problem:

$$\begin{aligned} \min \quad & \theta_2 - \theta_1, \\ \text{subject to} \quad & \int_{\theta_1}^{\theta_2} P(\theta|D, I)\, d\theta = 1 - \alpha. \end{aligned} \tag{3.15}$$

For a multimodal pdf which has one peak that is relatively much higher the others, we may neglect the smaller peaks. However, if we have more than one peaks

that have relatively the same height, we cannot use only one number for the best estimate of the parameter. Hence, the most honest thing we can do in that case is just displaying the posterior pdf itself (Sivia, 2006).

So far we have assumed that the error-bars are known. But very frequently they are not, so that the simplest case is when we have to deal with two unknown parameters, θ and s. Our posterior pdf is then the joint pdf, $P(\theta, s|D, I)$, and we still would like to get the best estimate of θ. This situation poses great difficulties to the frequentist school, and that is the reason why orthodox statisticians refer to parameter like s as *nuisance parameters*. In the Bayesian approach, nuisance parameters pose no problem, for all we have to do is using marginalization,

$$P(\theta|D, I) = \int_s P(\theta, s|D, I)ds, \tag{3.16}$$

where the integral is over the whole range of s. With the application of Bayes' theorem and the product rule to the integrand, we obtain

$$\begin{aligned} P(\theta|D, I) &\propto \int_s P(D|\theta, s, I)\, P(\theta, s|I)ds, \\ &= \int_s \underbrace{P(D|\theta, s, I)}_{\text{Eq. 3.6}}\, P(\theta|s, I)\, P(s|I)ds. \end{aligned} \tag{3.17}$$

We may use the assumption of independence for each datum and between the data and the error-bar; thus, $P(\theta|I) = P(\theta|s, I)$. For the prior probabilities, we use the flat prior for $P(\theta|I)$; however, since s is a scale parameter, we shall use the Jeffrey's prior for the error parameter (Jeffreys, 1958),

$$P(s|I) = \begin{cases} \frac{1}{s}, & \text{for } 0 < s < \infty \\ 0 & \text{otherwise.} \end{cases} . \tag{3.18}$$

As a result, the posterior pdf becomes

$$\begin{aligned} P(\theta|D,I) &\propto \int_s \frac{1}{s^{N+1}} \exp\left(-\sum_{i=1}^{N} \frac{(d_i-\theta)^2}{2\,s^2}\right) ds, \\ &= \frac{A}{\left(\sum_{i=1}^{N}(d_i-\theta)^2\right)^{\frac{N}{2}}}, \end{aligned} \tag{3.19}$$

where the integration is easily evaluated by a change of variable $t = \frac{1}{s}$ and A is the normalization constant. The posterior pdf in Eq. 3.19 has a Student-t distribution as opposed to Gaussian distribution. The best estimate of θ and its uncertainty can be obtained by following the same procedures as in the previous case.

In general, if we have more than one parameters to be estimated $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_k)$, the above derivation extends naturally; we will have a joint posterior pdf instead of single variable posterior pdf,

$$P(\vec{\theta}|D,I) \propto \exp\left(-\sum_{i=1}^{N} \frac{[d_i - f(\vec{\theta})]^2}{2\,s_i^2}\right), \tag{3.20}$$

if the s_i are known. If they are unknown, we have

$$P(\vec{\theta}|D,I) \propto \frac{1}{\left(\sum_{i=1}^{N}[d_i - f(\vec{\theta})]^2\right)^{\frac{N}{2}}}. \tag{3.21}$$

In both cases, we will have a posterior hyper-surface rather than a posterior curve. For illustration purposes, interested readers are encouraged to visit chapter three of (Sivia, 2006) in which a straight-line fitting example is presented. Furthermore, for curious readers, a paper by Stephen Gull entitled "Bayesian data analysis - Straight line fitting" (Gull, 1988) is a good starting point for the case where both dependent and independent variables contain uncertainties.

x	σ
1.867	0.014
1.837	0.012
1.847	0.003
1.853	0.003
1.858	0.003
1.906	0.020

Table 3.1: Adapted from (Hamilton, 1964)

3.2 An example from theoretical chemistry

As a simple example of application of the theory reviewed in the previous section, consider the following scientific problem. Crystallographic techniques have been used to measure the length of the bond between an atom of Phosphor and an atom of Carbon in a certain molecule ($(PC_3)_4$). Six measurements have been performed, and the results are displayed in Tab. 3.1.

In Bayesian reasoning, we ask ourselves what the plausible value of the P-C bond length is in light of these six observations. Let us denote by μ the estimate for the P-C bond length, which is a constant. We want to obtain $P(\mu|\text{data}, I)$. Thanks to Bayes' theorem, we can write

$$P(\mu|\text{data}, I) \propto P(\text{data}|\mu, I)\, P(\mu|I), \tag{3.22}$$

and since this is a parameter estimation case, we absorb $P(\text{data}|I)$ into the normalization constant. For the prior probability, we shall assign a uniform prior because we have no reason to believe any particular value is more preferable than others. The most honest way to assign the prior probability is

$$P(\mu|I) = \begin{cases} \frac{1}{[\mu_{\max} - \mu_{\min}]}, & \text{for} \quad \mu_{\min} \leq \mu \leq \mu_{\max} \\ 0 & \text{otherwise.} \end{cases} . \tag{3.23}$$

For the likelihood function formulation, we need to look at the implication of

the constant model. We can model the data as

$$x_i = \mu + \epsilon_i, \qquad (i = 1, 2, ..., 6), \tag{3.24}$$

where ϵ_i is the noise. Since the errors follow a Gaussian distribution, we have,

$$P(x_i|\mu, I) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left\{-\frac{1}{2\sigma_i^2}(x_i - \mu)^2\right\}. \tag{3.25}$$

Assuming that each observation is independent, the likelihood function is simplified into

$$P(\text{data}|\mu, I) = \frac{1}{(\sqrt{2\pi})^6}\,\frac{1}{\sigma_1\sigma_2\,...\,\sigma_6} \exp\left\{-\frac{1}{2}\sum_{i=1}^{6}\frac{(x_i - \mu)^2}{\sigma_i^2}\right\}. \tag{3.26}$$

The posterior density function (pdf) is essentially the product of the likelihood function (Eq. 3.26) and the prior probability (Eq. 3.23); thus,

$$P(\mu|\text{data}, I) = C\ \exp\left\{-\frac{1}{2}\sum_{i=1}^{6}\frac{(x_i - \mu)^2}{\sigma_i^2}\right\}, \tag{3.27}$$

where all the constants are absorbed into C, which is determined so that the area under the curve is unity. Following the procedures explained in the previous section on data with known error-bars, we obtain the following estimate

$$\hat{\mu} = 1.853 \pm 0.0017, \tag{3.28}$$

for the most plausible P-C bond length based on the six observations. Had we not known the uncertainties or error-bars on each observation, we could have used marginalization as shown in previous section on data with unknown error-bars. The estimate for the most probable P-C bond length is then

$$\hat{\mu} = 1.861 \pm 0.013. \tag{3.29}$$

Fig. 3.2 illustrates the posterior probability density for each case.

It is important to keep in mind that the estimate $\hat{\mu}$ is nothing more than the most plausible value in light of these six measurements and our prior information. As it is evident from Fig. 3.2 the full width at half maximum (FWHM) represents the reliability of our estimate. As a result, it is appropriate to summarize our estimate in the form (*most plausible value* $\pm$ *something*) as long as the resulting posterior pdf is *symmetric* and *unimodal*.

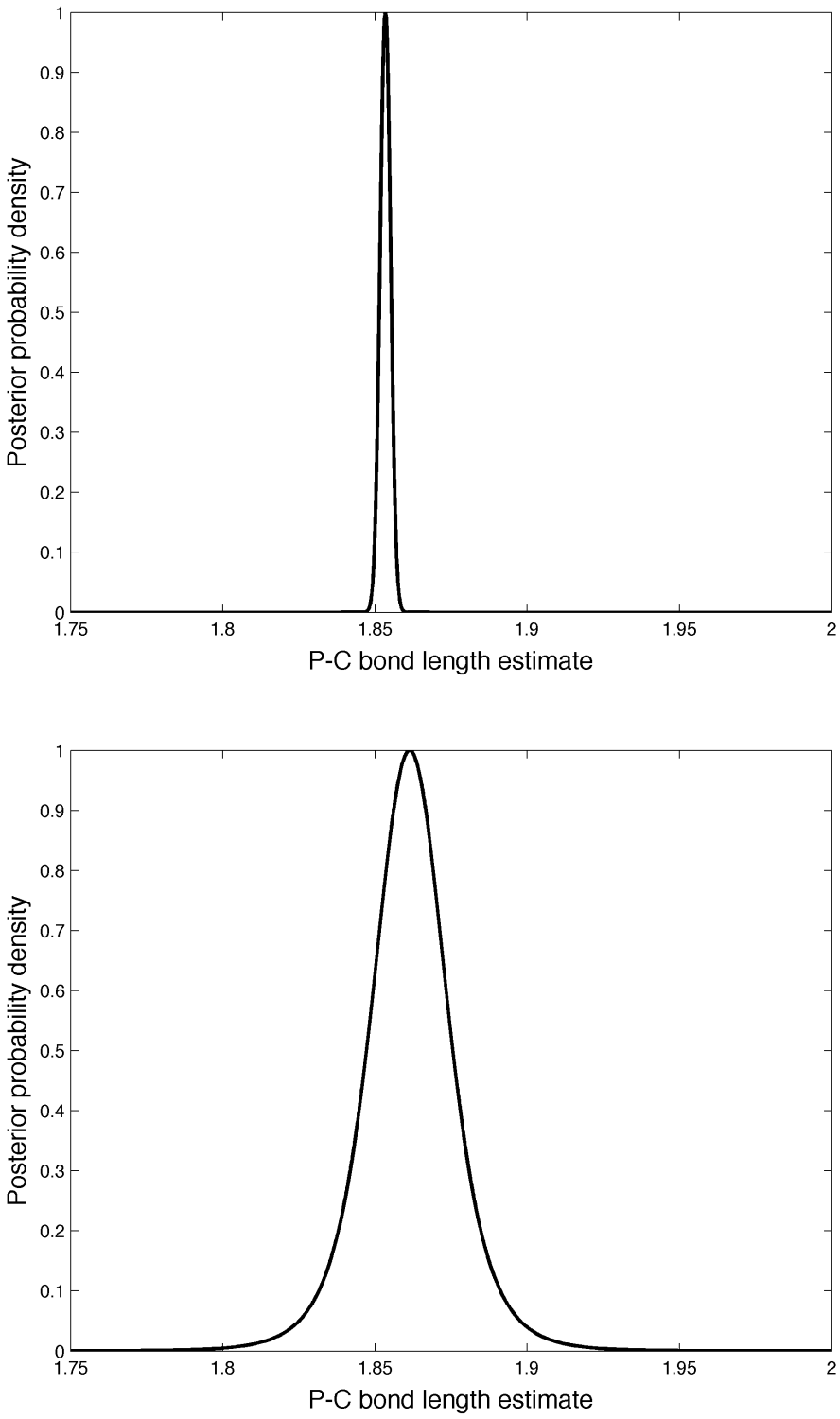

Figure 3.2: The posterior probability density function as a result of known (left) and unknown (right) error-bars respectively.

3.3 Model Selection

The assumed form of the model $f(\theta_i)$ in Section 3.1 was simply a constant. In many other situations, however, we might be interested in inferring whether the model should be a linear, or a quadratic, or an even more complicated function. Choosing a model in cases when there is uncertainty as to which one of a set of alternative models is most suitable is called *model selection*.

Confronted with this problem, we might think of choosing the model on the basis of how well the alternatives fit the data. However, this thought is naive because models with more adjustable parameters, of which they may be infinitely many, will always fit the data better. Still, most people would prefer the model with the fewest possible parameters, as the main goal of science is to explain the observed phenomena with *the simplest* possible theory.

To explain how the decision is made within the Bayesian framework, it is best to repeat the elementary formulation (originally due to Jeffreys (Jeffreys, 1958)) presented in chapter four of Sivia (Sivia, 2006). Suppose that in a graph-fitting context we formulate two models:

- M_1 – The noisy measurements of y versus x are described by $y = 0$;
- M_2 – The same measurements are described by $y = c$, with c an adjustable parameter.

Which is more plausible on the basis of data D ? In the Bayesian framework, we quantify the merit of the two theories based on each posterior probability, $P(M_1|D, I)$ and $P(M_2|D, I)$, respectively. Equivalently, we may use the posterior ratio,

$$\text{posterior ratio} = \frac{P(M_1|D, I)}{P(M_2|D, I)}; \tag{3.30}$$

if the posterior ratio is greater than one (in order of magnitude) then the theory M_1 is more plausible than that of M_2. If the posterior ratio is less than one then M_2 is more plausible. Finally, if the posterior ratio is approximately of order one, then

the current data is insufficient to make an informed judgement. Applying Bayes' theorem on the right-hand side gives

$$\text{posterior ratio} = \frac{P(D|M_1, I)}{P(D|M_2, I)} \times \underbrace{\frac{P(M_1|I)}{P(M_2|I)}}_{\text{prior ratio}}, \tag{3.31}$$

where the evidence $P(D|I)$ cancels out. Probability theory tells us that the merit of the two theories depends partly on what each individual knows *prior* to the analysis of the data. If we assign the prior ratio to be unity, representing fairness, the posterior ratio becomes the ratio of the likelihood functions, and to assign the latter, we need to be able to compare the data with the predictions of M_1 and M_2: The larger the mismatch, the lower the corresponding probability. The calculation for M_1 is straight-forward, but the one for M_2 is complicated by the fact that the parameter c is unknown. This is a clear example of the usefulness of marginalization, for we can write

$$\begin{aligned} P(D|M_2, I) &= \int_{c\text{-range}} P(D, c|M_2, I)dc, \\ &= \int_{c\text{-range}} P(D|c, M_2, I)P(c|M_2, I)dc, \end{aligned} \tag{3.32}$$

where in the second step we just applied the product rule. The first factor under the integral is the ordinary likelihood function (on a par with $P(D|M_1, I)$) with the value of c given. The second factor, on the other hand, is M_2's prior probability for c. Hence, it is the duty of the theorist to articulate his state-of-knowledge, or ignorance, before seeing the data. We will illustrate how this idea can be developed in the application presented in the next section.

3.4 Another example from theoretical chemistry

Consider a set of data representing the natural rubber content in mixture of natural and synthetic rubber obtained by infrared spectroscopy (Mandel, 1984). The

i	x_i (%)	y_i
1	0	0.734
2	20	0.885
3	40	1.050
4	60	1.191
5	80	1.314
6	100	1.432

Table 3.2: The experiments on the mixture of natural and synthetic rubber (x_i) and the natural rubber content (y_i)

experimental results are tabulated in Tab. 3.2 and plotted in Fig. 3.3. From the experimental data, we want to infer which model is more plausible: a linear or a quadratic model.

The two propositions of interests are H_1 and H_2, where respectively,

1. H_1 is a proposition of a linear model: $y_i^{(1)} = a_1 + b_1 x_i + \epsilon_i$, and
2. H_2 is a proposition of a quadratic model: $y_i^{(2)} = a_2 + b_2 x_i + c_2 x_i^2 + \epsilon_i$,

where $i = 1, ..., 6$ and ϵ_i denotes the measurement error (noise) for the i^{th} datum. In the first model H_1, we have two adjustable parameters (a_1 and b_1); whereas, in H_2, we have three adjustable parameters (a_2, b_2, and c_2). We begin with the posterior pdf for each proposition through Bayes' theorem,

$$P(H_k|D, I) = \frac{P(D|H_k, I)\, P(H_k|I)}{P(D|I)}, \tag{3.33}$$

where $k = 1, 2$ denoting each proposition. The denominator in Eq. 3.33 cancels out if we use the posterior ratio (odds ratio),

$$O_{12} = \frac{P(D|H_1, I)}{P(D|H_2, I)} \times \underbrace{\frac{P(H_1|I)}{P(H_2|I)}}_{\text{prior ratio}}. \tag{3.34}$$

If we assign equal plausibility to each proposition, the odds ratio simplifies into the likelihood ratio,

$$O_{12} = \frac{P(D|H_1, I)}{P(D|H_2, I)}. \tag{3.35}$$

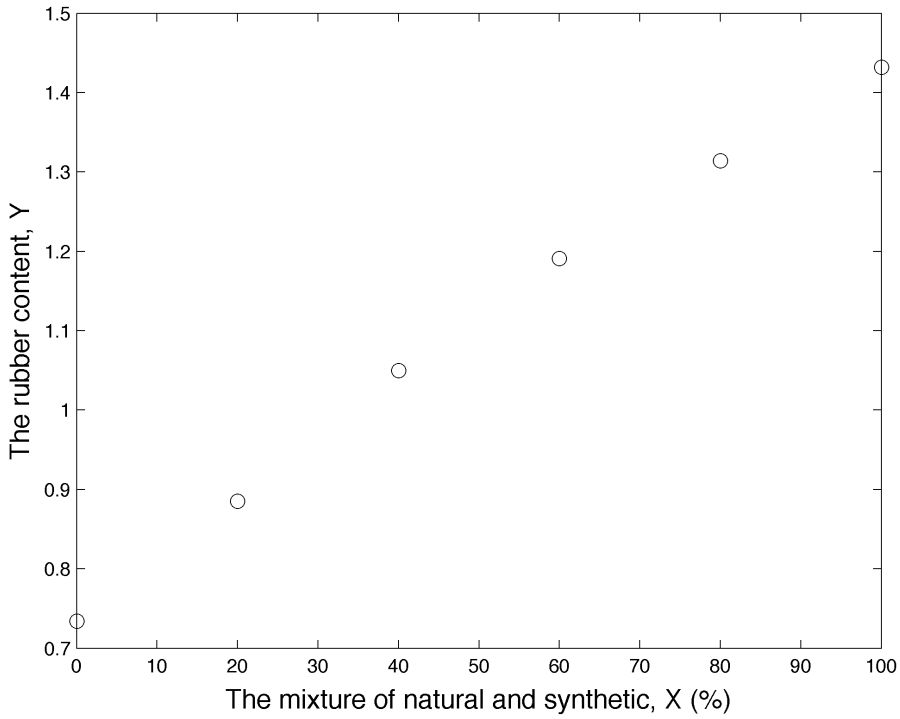

Figure 3.3: The scatter-plot of the experiments

Thus, we need to compare the data in light of each proposition H_1 and H_2; the larger the mismatch the smaller the probability.

We cannot directly compare the data with each of the propositions because each proposition contains adjustable parameters which are unknown. As it was explained earlier, the measurement errors follow a Gaussian distribution with an unknown variance, s. Thus, we have three unknowns for H_1 and four unknowns for H_2. Instead of estimating each unknown which is not our concern in evaluating the merit of a proposition, we may use marginalization available to us; therefore,

for H_1,

$$\begin{aligned} P(D|H_1, I) &= \int_{b_1}\int_{a_1}\int_{s} P(D, a_1, b_1, s|H_1, I) ds da_1 db_1, \\ &= \int_{b_1}\int_{a_1}\int_{s} P(D|a_1, b_1, s, H_1, I) P(a_1, b_1, s|H_1, I) ds da_1 db_1, \\ &= \int_{b_1}\int_{a_1}\int_{s} \underbrace{P(D|a_1, b_1, s, H_1, I)}_{\text{likelihood}} \underbrace{P(a_1, b_1|s, H_1, I)}_{*} P(s|H_1, I) ds da_1 db_1, \end{aligned} \tag{3.36}$$

where we applied consecutive product rules. In general, we need not find the estimates for each parameter; however, as we see later, in order to find analytically tractable solution, we may need to do so. It is reasonable to assume that the adjustable parameters are independent of the variance of the noise, so that $*$ becomes simply $P(a_1, b_1|H_1, I)$, which is the joint prior for parameter a_1 and b_1. $P(s|H_1, I)$ is just the prior for the standard deviation of the noise. The limits of integration for each adjustable parameter are determined by its range of possible values. Since we have no prior information about the plausibility of the values for each parameter, we then assign a flat prior for each adjustable parameter,

$$P(a_i|I) = \begin{cases} \frac{1}{\triangle a_i}, & \text{for} \quad a_i^{\min} \leq a_i \leq a_i^{\max} \\ 0 & \text{otherwise.} \end{cases}, \tag{3.37}$$

where $\triangle a_i = a_i^{\max} - a_i^{\min}$, $i = 1, 2$, and similarly for the adjustable parameter b and c. For the scale parameter s, we assign the Jeffrey's prior,

$$P(s|I) = \begin{cases} \frac{1}{s}, & \text{for} \quad s \in (0, \infty) \\ 0 & \text{otherwise.} \end{cases}. \tag{3.38}$$

For each likelihood, we assume independent, additive, Gaussian noise; as a result, the inner most integration due to the marginalization of the variance of the noise converges as $s \to \infty$, giving a Student-t distribution or also known as Cauchy distribution (Section 3.2). Similar procedures apply for H_2 but we have one more

integral due to the additional adjustable parameter c_2. The odds ratio now becomes, after some simplifications,

$$O_{12} = \underbrace{\frac{\triangle a_2 \triangle b_2 \triangle c_2}{\triangle a_1 \triangle b_1}}_{\text{Ockham's factor}} \times \frac{\int_{b_1}\int_{a_1} \frac{1}{\left[\sum_{i=1}^{6}(a_1+b_1x_i-y_i^{(1)})^2\right]^3} da_1 db_1}{\int_{c_2}\int_{b_2}\int_{a_2} \frac{1}{\left[\sum_{i=1}^{6}(a_2+b_2x_i+c_2x_i^2-y_i^{(2)})^2\right]^3} da_2 db_2 dc_2}. \tag{3.39}$$

We then may proceed either analytically by approximating the likelihood function or numerically by computing the multi-dimensional integrals. The latter can be done by using a Monte Carlo integration; since we only have up to three dimensional integration, a crude Monte Carlo method still suffices. For higher dimensions however, it may be more efficient to use importance sampling. The former approach is better in a sense that it is analytically tractable. We outline this approach briefly. If the joint prior for the adjustable parameters does not cut off too much of the integrand $P(D|a_1, b_1, s, H_1, I)$ - i.e, if most of the distribution falls into the rectangle $[a_1^{\min}, a_2^{\max}] \times [b_1^{\min}, b_2^{\max}]$ - then we can make the *approximation*:

$$P(D|a_1, b_1, s, H_1, I) \approx P(D|\hat{a}_1, \hat{b}_1, s, H_1, I) \exp\left\{-\frac{1}{2}\left[\frac{(a_1-\hat{a}_1)^2}{\delta a_1^2} + \frac{(b_1-\hat{b}_1)^2}{\delta b_1^2}\right]\right\}, \tag{3.40}$$

where $\hat{a}_1$ and $\hat{b}_1$ are the best estimate for each adjustable parameter in H_1 and δa_1 and δb_1 are the uncertainty of the estimates. We can generalize Eq. 3.40 by including the correlation term; however, for illustration purposes, we ignore the correlation between a_1 and b_1. The best estimates, along with their uncertainties, can be obtained by following the parameter estimation method explained in Section 3.1. The parameter estimates are tabulated in Tab. 3.3. Substituting Eq. 3.40, along with its respective prior probabilities, into Eq. 3.36, and evaluating the integral respectively, we obtain

$$P(D|H_1, I) \propto \frac{1}{\left[\sum_{i=1}^{6}(\hat{a}_1+\hat{b}_1x_i-y_i^{(1)})^2\right]^3} \frac{\delta a_1 \delta b_1}{\triangle a_1 \triangle b_1}; \tag{3.41}$$

similar procedures for H_2 except that we have an additional integration. Thus, the

	Estimate
$\hat{a}_1$	7.5×10^{-1}
$\hat{b}_1$	7.1×10^{-3}
δa_1	1.1×10^{-2}
δb_1	1.9×10^{-4}
$\hat{a}_2$	7.3×10^{-1}
$\hat{b}_2$	8.5×10^{-3}
$\hat{c}_2$	-1.5×10^{-4}
δa_2	4.2×10^{-3}
δb_2	2.0×10^{-4}
δc_2	1.9×10^{-6}

Table 3.3: The best estimates, along with their uncertainties, of each proposition.

posterior ratio O_{12} becomes

$$\begin{aligned} O_{12} &= \overbrace{\frac{\left[\sum_{i=1}^{6}(\hat{a}_1 + \hat{b}_1 x_i - y_i^{(1)})^2\right]^{-3}}{\left[\sum_{i=1}^{6}(\hat{a}_2 + \hat{b}_2 x_i + \hat{c}_2 x_i^2 - y_i^{(2)})^2\right]^{-3}}}^{\text{likelihood ratio}} \times \underbrace{\frac{\delta a_1 \delta b_1}{\delta a_2 \delta b_2 \delta c_2} \times \frac{\triangle a_2 \triangle b_2 \triangle c_2}{\triangle a_1 \triangle b_1}}_{\text{Ockham's factor}}, \\ &\approx \mathcal{O}(10^5), \end{aligned} \tag{3.42}$$

where the prior ranges for each parameter are taken as follows

- $a_i \in [0, 100]$; when we do not have any rubber in the mixture, the probability of the rubber content is zero. The maximum value however can be as large as possible (for simplicity, we set the maximum to be 100),
- $b_i \in [0, 100]$; the argument is that the more rubber we have in the mixture, the rubber content shall increase, therefore, this eliminates the possibility of negative slopes,
- $c_2 \in [-100, 100]$; this however can be negative since we can also have a negative concavity when we have a positive slope.

where $i = 1, 2$. We keep in mind that in determining of possible ranges for each parameter, we should be as uninformative as possible unless we have had experience with similar experiments before. Additionally, we should not be influenced with the scatter-plot of the current data since, by definition, prior probability is assigned *before* we look at the current data.

The latter approach can be done easily with a Monte Carlo integration (see Appendix A, for a brief overview of Monte Carlo integration), leading to the same conclusion which is in favor of the linear model. Evidently, the likelihood ratio prefers the more complicated model; whereas, the Ockham's factor punishes the more complicated model because of its additional parameter. Thus, based on the experimental data and our prior knowledge, we conclude that the most plausible proposition is the linear one as opposed to the quadratic one.

In the next chapter, we begin applying Bayesian methods to cancer research problems, namely tumor hypoxia inference in cervical carcinomas and clinical trials' data analysis in relapsed ovarian cancers.

Chapter 4

Bayesian analysis of tumor hypoxia data

In this chapter, we present the results of the assessment of tumor hypoxia by using Bayesian methods to analyze the data generated by the PMH group by means of both direct oxygen partial pressure (pO_2) measurements in vivo and immunohistochemical assays of an intrinsic protein marker, namely Carbonic Anhydrase IX (CAIX). Additionally, we analyze whether or not the two tumor hypoxia quantifications are in concordance with each other. First, however, a description of tumor hypoxia, more detailed than the one given in the Introduction is in order.

4.1 Tumor hypoxia

Hypoxia is defined as a state of oxygen deficiency. It occurs in an early stage of tumor growth when tumor cells depend on the diffusion of oxygen from nearby blood vessels. At this stage, the tumor is said to be in an avascular stage. Some examples of avascular tumors include carcinomas, lymphomas, and sarcomas. A vascular tumor, on the other hand, is a matured one. It does not depend on the diffusion of oxygen any longer because it has already completed angiogenesis, which is the formation of new blood vessels. Thus, a vascular tumor tends to be

aggressive because it has its own supply of nutrients; moreover, tumor cells are capable of entering blood vessels and traveling to distant sites, forming metastasis.

A more precise definition of tumor hypoxia is a state of a decreased dissolved oxygen concentration (pO_2) below a critical level (Hckel and Vaupel, 2001). The definite critical oxygen level is still debatable. Nevertheless, many agree that if the tissue oxygen level falls below 10 mmHg, the particular tissue is considered hypoxic (Hckel and Vaupel, 2001; Milosevic et al., 2004). Based on the causes at tissue level, tumor hypoxia can be classified into three different groups, known as radiobiological hypoxia (Raleigh et al., 1996; Horsman, 1998).

The first type of radiobiological hypoxia is a *chronic (diffusion-limited)* hypoxia. Thomlinson and Gray demonstrated that tumor cells that are beyond 150 μm of nearby blood vessels become starved of oxygen and die, forming necrosis (Hall, 1994; Dewhirst, 1998; Vaupel and Harrison, 2004). Additionally, diffusion-limited hypoxia may also be caused by the deterioration of diffusion "geometry", for example, concurrent versus countercurrent blood flow within the tumor microvessel network (Vaupel and Harrison, 2004). The second type is *acute* or *perfusion-limited* hypoxia which is transient in nature (Hall, 1994; Dewhirst, 1998; Vaupel and Harrison, 2004). This type of hypoxia is usually present in the stroma of solid tumors. Perfusion-limited hypoxia is caused by inadequate blood flow in tissues due to severe structural and functional abnormalities in tumor blood flows (Vaupel and Harrison, 2004). Tumor microvessels are widely known to be dilated, tortuous, elongated, and saccular; moreover, they often collapse and cause transient blockages in the flow of oxygen (Hckel and Vaupel, 2001). Once the blockages disappear, oxygen is able to flow to reoxygenate the tumor area. The third type of tumor hypoxia is an *anemic hypoxia*. It is caused by a reduced O_2 transport capacity of the blood subsequent to tumor-associated or therapy-induced anemia (Vaupel and Harrison, 2004).

The three types of tumor hypoxia mentioned earlier are the ones that are important for tumors. With the exception of anemic hypoxia, both chronic and acute hypoxia have been widely demonstrated to be crucially important for cancer treat-

ment (Raleigh et al., 1996; Horsman, 1998). Nevertheless, there has not been any success in distinguishing between the two types of tumor hypoxia. Sequential imaging may provide some clues whether or not tumor hypoxia is acute; however, not all tumor sites are accessible by imaging techniques. Additionally, tumor vasculature is uncontrollably chaotic. Tumor microvessels are leaky, tortuous, elongated; moreover, they have incomplete endothelial linings and an increased vascular permeability. (For a more complete review regarding tumor vasculature and its effects on radiotherapy, the reader is kindly directed to (Vaupel, 2004).) This provides another challenge in determining the cause of tumor hypoxia. There are numerous factors affecting tumor hypoxia: to name a few, the host tissue of the tumor, the type of the tumor, and the grade of the tumor. These factors make the modeling of tumor hypoxia very challenging.

4.2 Tumor hypoxia measurements

Many studies of tumor hypoxia was based on direct pO_2 measurements using a polarographic oxygen electrode. A commonly used commercial system is the Eppendorf needle probe with an outer diameter of 300μm and a tip diameter of 120μm. The electrode is inserted into the tissue in steps of 1.0 mm; each is followed by a backward motion of 0.3 mm (Fyles et al., 2002). The inserted tip of the electrode is covered with teflon permeable only to oxygen molecules. As a result, the 'bumping' of these oxygen molecules to the tip of the needle is transformed into electrical signals which are then processed by a computer. What the electrode measures is the partial pressure of oxygen (pO_2) within a tumor. A complete review of the polarographic needle probe can be found in (Fatt, 1976). The Eppendorf needle electrode is considered the 'gold standard' in tumor hypoxia quantification.

Although this technique provides the most direct measurement, it is limited by its invasiveness. Some sites (kidney, urether, bladder, testicles) are considered to be too risky for oxygen electrode use (Raleigh et al., 1996). An attractive

alternative is immunohistochemical staining to detect proteins expressed by cells during hypoxia. Hypoxia is one of the adverse cell environments; consequently, cells are forced to adapt to it. The idea of immunohistochemistry is to identify chemicals responsible for cells' response. As the study of immunohistochemistry (protein markers) advances, scientists have been able to isolate chemicals that are specific to hypoxic cells. Raleigh and his group laid the basis for the hypoxia marker technique in tumor hypoxia quantification. They observed that 2-nitroimidazoles hypoxic cell radiosensitizers are activated and bound to hypoxic cells. This process shows an oxygen dependence close to that for the radiobiological oxygen effect (Raleigh et al., 1996; Horsman, 1998; Cline et al., 1990). The binding is a metabolic process. It involves endogeneous nitroreductases (enzymes) which convert the nitroheterocyclic compounds to binding intermediates in an oxygen dependence manner (Cline et al., 1990).

There are many 2-nitroimidazole compounds; however, CCI-103F (used in canine solid tumors), pimonidazole, and EF5 are the ones commonly used in tumor hypoxia research (Cline et al., 1997; Evans et al., 2006; Durand and Aquino-Parsons, 2006). While CCI-103F has been used solely in canine tumors, pimonidazole has been widely used as a hypoxia marker in a number of human malignancies. It has been designated as a novel hypoxia marker (Varia et al., 1998). The significant relationship between tumor blood vessels and EF5 staining makes this marker a "blood vessel marker", showing perfused blood vessels (Evans et al., 2001). Tumors with a high density of EF5 indicate that they have already begun or completed angiogenesis. Additionally, thioredoxin may serve as a hypoxia marker because of its role in redox mediators in biochemical pathways promoting cell survival under adverse conditions such as hypoxia (Hedley et al., 2004).

The hypoxia markers mentioned earlier are of *extrinsic* or *exogeneous* type. It means that there is a need for an injection of the marker prior to sampling (biopsy). Fortunately, there are other types of hypoxia markers that do not require any administration prior to biopsy: two such examples are HIF-1α and Carbonic Anhy-

drase IX (CAIX). This type of hypoxia marker is called an *intrinsic* or *endogeneous* marker.

The overexpression of hypoxia-inducible factor-1α (HIF-1α) indicates that tumor cells have already begun to adapt to adverse environments (such as low oxygenation); hence, an anti-angiogenic therapy, whose goal is to fail tumor angiogenesis, may be more suitable rather than conventional cancer therapies (chemotherapy or radiotherapy). For a more detail review about anti-angiogenic therapy, the reader is directed to (Jain, 2005). Carbonic anhydrase IX (CAIX) is an enzyme expressed on the cell membrane during hypoxia to balance the immediate extracellular microenvironment. It is widely regarded as a surrogate marker of chronic hypoxia in various cancers (Olive et al., 2001; Hoogsteen et al., 2005; Mseide et al., 2004; Hoskin et al., 2003); moreover, the overexpression of CAIX has been significantly associated with a poor prognosis (Hedley et al., 2003).

In an attempt to replace the direct pO_2 measurements with immunohistochemical assays of hypoxia protein markers, many researchers have claimed that hypoxia markers can assess tumor hypoxia as well as the in vivo Eppendorf pO_2 measurements. Nevertheless, there is no agreement in current literature; some believe that the two hypoxia quantification methods shall correlate with each other (Loncaster et al., 2001) but some disagree with that conclusion (Mayer et al., 2005; Hedley et al., 2003). Clarifying this confusion is one of the topics in this thesis; however, first we will take a step back and attempt to ask a question: what do we want to learn from the Eppendorf pO_2 measurements and the immunohistochemistry of CAIX respectively ? The next question is then how the two methods fare in quantifying tumor hypoxia, whether or not they are in concordance with each other.

4.3 Tumor hypoxia inferences

The first question is definitely an inference problem; we do not have a complete knowledge about the tumor. Our knowledge is limited to the data either obtained

from biopsies or direct Eppendorf measurements and by our theoretical knowledge about tumor dynamics. Unfortunately, unlike the laws of classical physics, the laws of tumor dynamics have not been solidly formulated yet. Nonetheless, physicians are called everyday to make decisions which affect patients directly, and data analysis in the presence of uncertainty is the best tool available.

As already mentioned, the hypoxia data generated by the PMH group have been analyzed with the classical orthodox statistical techniques, but some fundamental ambiguities and contradictions have not been resolved. It is therefore of practical importance to find out whether Bayesian probability methods are capable of offering sharper and more consistent inferences from the available data. First, we will proceed with tumor hypoxia quantification by the direct pO_2 measurements.

4.3.1 The direct pO_2 measurements

The pO_2 measurements by the Eppendorf needle can be performed in linear, random, or circular tracks. Each track consists of about 20 - 30 measurements each of which is recorded in mmHg. A typical track of pO_2 measurements is tabulated in Tab. 4.1. From all the tracks obtained, we are interested in inferring the hypoxic proportion, the proportion of measurements that is less than or equal to 5 mmHg (HP5). This is a measure used by clinicians to quantify tumor hypoxia. If HP5 is greater than or equal to 50%, the tumor is considered to be hypoxic based on the pO_2 measurements. HP5 is by no means the standard rule. Some prefer to use HP10 or HP2.5 to quantify tumor hypoxia.

Our collaborators at the Princess Margaret Hospital provided us with data sets of pO_2 measurements for 21 patients with invasive cervical carcinomas. From these, we want to obtain the most probable HP5 from each patient. Let us denote by X the proportion of the pO_2 measurements that are less than or equal to 5 mmHg. As it was explained earlier, the inferential machinery in Bayesian methodology is

Study number	Track number	Position	pO_2 (mmHg)
2136	1	1	3.5
2136	1	2	3.6
2136	1	3	3.8
2136	1	4	3.5
2136	1	5	3.8
2136	1	6	47.6
2136	1	7	154.7
2136	1	8	118.1
2136	1	9	82.1
2136	1	10	50.2
2136	1	11	29.5
2136	1	12	7.8
2136	1	13	2.9
2136	1	14	2.8
2136	1	15	2.7
2136	1	16	6.6
2136	1	17	18.1
2136	1	18	31.7
2136	1	19	40.7
2136	1	20	32.7
2136	1	21	24.8
2136	1	22	13.5
2136	1	23	5.2
2136	1	24	3.2
2136	1	25	3.2
2136	1	26	3.2
2136	1	27	2.8
2136	1	28	3.3
2136	1	29	3.2
2136	1	30	3.0
2136	1	31	2.8
2136	1	32	3.4
2136	1	33	3.1

Table 4.1: The first track of the pO_2 measurements taken from patient 2136

Bayes' theorem:

$$P(X^{(i)}|D^{(i)}, I) \propto P(D^{(i)}|X^{(i)}, I) \times P(X^{(i)}|I), \tag{4.1}$$

where the superscript i denotes the i^{th} track ($i = 1, 2, ..., 5$) and I is the background information such as each track and each measurement is independent of one another, respectively. By the application of the Bayes' theorem, we obtain the most probable value of the HP5 in light of the data and prior knowledge.

For the prior knowledge or the prior probability assignment, we assign a uniform or ignorance prior,

$$P(X^{(i)}|I) = \begin{cases} 1, & \text{for } 0 \leq X^{(i)} \leq 1 \\ 0 & \text{otherwise} \end{cases}, \tag{4.2}$$

because we have no reason to prefer any value for the HP5. The direct probability or the likelihood function can be thought of as obtaining the data 'r measurements that are less than or equal to 5 mmHg in N measurements' in a track, which is a binomial distribution,

$$P(D^{(i)}|X^{(i)}, I) \propto X^{(i)r} \, (1 - X^{(i)})^{N-r}. \tag{4.3}$$

The first application of the Bayes' theorem gives us a posterior probability density function (pdf) for HP5 based on one track of measurements; this posterior pdf can in turn serve as the prior knowledge for the second track and so on. Thus, one can view Bayes' theorem as an inferential tool that gradually incorporates additional information as they become available to improve our knowledge. The best estimate for the most probable HP5 value in a patient is given by the maximum of the posterior pdf in light of the last track's data and the uncertainty is simply the full width at half maximum (FWHM) of the posterior pdf. The results are summarized in Tab. 4.2. For completeness, we also present the other commonly-used hypoxic proportions: HP2.5 and HP10, summarized in Tab. 4.3 and 4.4 respectively. Based

on the criterion that a tumor is considered to be hypoxic if its HP5 is greater than or equal to 50% (Pitson et al., 2001), there are 12 out of 21 tumors can be classified as hypoxic (Tab. 4.2). In contrast, only nine of the same tumors would be considered as hypoxic marker under the HP2.5 rule, while that number goes up to 14 under the HP10 rule. Although these rules are arbitrary, most medical researchers consider the HP5, which falls in the middle of the range, as the most reliable criterion.

Patient	$HP5^{(1)}$	$HP5^{(2)}$	$HP5^{(3)}$	$HP5^{(4)}$	$HP5^{(5)}$	$HP5^{(6)}$	$HP5^{(7)}$	$HP5^{(8)}$
2136	55 ± 8.7	30 ± 5.8	20 ± 4.1	**24 ± 3.8**	–	–	–	–
2144	66 ± 8.4	67 ± 5.9	69 ± 4.7	**71 ± 4.0**	–	–	–	–
2149	28 ± 7.9	52 ± 6.2	55 ± 5.1	52 ± 4.4	**42 ± 3.9**	–	–	–
2152	66 ± 8.4	38 ± 6.1	58 ± 5.0	66 ± 4.2	53 ± 3.9	**51 ± 3.6**	–	–
2154	31 ± 8.2	28 ± 5.6	38 ± 4.9	**38 ± 4.2**	–	–	–	–
2156	58 ± 8.9	73 ± 5.6	61 ± 5.0	61 ± 4.3	**57 ± 3.9**	–	–	–
2157	3.1 ± 3.1	4.7 ± 2.6	14 ± 3.5	**10 ± 2.7**	–	–	–	–
2161	0	0	**0**	–	–	–	–	–
2165	0	25 ± 6.0	16 ± 4.1	13 ± 3.3	**33 ± 4.1**	–	–	–
2166	78 ± 7.3	53 ± 6.2	44 ± 5.1	**42 ± 4.4**	–	–	–	–
2167	34 ± 8.4	38 ± 6.1	54 ± 5.1	57 ± 4.4	64 ± 3.8	**66 ± 3.4**	–	–
2169	72 ± 7.9	86 ± 4.3	81 ± 4.0	84 ± 3.3	**85 ± 3.1**	–	–	–
2171	29 ± 8.2	29 ± 5.7	43 ± 5.1	56 ± 4.4	60 ± 4.0	**63 ± 3.5**	–	–
2172	78 ± 7.3	83 ± 4.9	74 ± 4.6	77 ± 3.8	**64 ± 3.9**	–	–	–
2174	75 ± 7.7	86 ± 4.3	85 ± 3.6	88 ± 2.9	76 ± 3.4	77 ± 3.0	70 ± 3.1	**61 ± 3.0**
2175	42 ± 8.9	57 ± 6.2	55 ± 5.1	62 ± 4.3	65 ± 3.8	**71 ± 3.3**	–	–
2176	0	22 ± 5.2	32 ± 4.8	39 ± 4.3	36 ± 3.8	**31 ± 3.4**	–	–
2177	22 ± 7.3	11 ± 3.9	18 ± 3.9	**13 ± 3.0**	–	–	–	–
2179	91 ± 5.2	92 ± 3.4	84 ± 3.7	85 ± 3.1	**88 ± 2.6**	–	–	–
2180	81 ± 6.9	91 ± 3.6	94 ± 2.5	80 ± 3.5	**83 ± 3.0**	–	–	–
2183	100	80 ± 5.4	63 ± 5.2	72 ± 4.1	75 ± 3.5	**80 ± 3.0**	–	–

Table 4.2: Summary of the most probable HP5 (%) estimate along with its uncertainty obtained from 21 patients as a track of measurements become available. The bold numbers denote the best HP5 estimate based on all available tracks.

Patient	HP2.5$^{(1)}$	HP2.5$^{(2)}$	HP2.5$^{(3)}$	HP2.5$^{(4)}$	HP2.5$^{(5)}$	HP2.5$^{(6)}$	HP2.5$^{(7)}$	HP2.5$^{(8)}$
2136	0	0	0	**7.1 ± 2.3**	–	–	–	–
2144	44 ± 8.8	53 ± 6.2	59 ± 5.0	**63 ± 4.3**	–	–	–	–
2149	0	22 ± 5.2	33 ± 4.8	34 ± 4.2	**27 ± 3.5**	–	–	–
2152	0	4.7 ± 2.6	36 ± 4.9	50 ± 4.4	40 ± 3.9	**39 ± 3.5**	–	–
2154	0	13 ± 4.1	24 ± 4.4	**27 ± 3.9**	–	–	–	–
2156	0	24 ± 5.4	23 ± 4.3	31 ± 4.1	**30 ± 3.6**	–	–	–
2157	0	0	3.1 ± 1.8	**2.3 ± 1.3**	–	–	–	–
2161	0	0	**0**	–	–	–	–	–
2165	0	23 ± 5.8	15 ± 4.0	12 ± 3.2	**33 ± 4.1**	–	–	–
2166	0	0	0	**0**	–	–	–	–
2167	31 ± 8.4	34 ± 5.9	49 ± 5.1	52 ± 4.4	60 ± 3.9	**61 ± 3.5**	–	–
2169	59 ± 8.7	78 ± 5.2	76 ± 4.4	79 ± 3.6	**80 ± 3.4**	–	–	–
2171	9.7 ± 5.3	13 ± 4.2	28 ± 4.6	42 ± 4.4	44 ± 4.0	**45 ± 3.6**	–	–
2172	56 ± 8.8	69 ± 6.0	64 ± 5.0	68 ± 4.2	**55 ± 4.0**	–	–	–
2174	53 ± 8.8	75 ± 5.4	78 ± 4.2	80 ± 3.6	68 ± 3.7	71 ± 3.3	64 ± 3.2	**56 ± 3.1**
2175	35 ± 8.6	54 ± 6.3	53 ± 5.1	60 ± 4.3	62 ± 3.8	**69 ± 3.3**	–	–
2176	0	21 ± 5.1	29 ± 4.7	37 ± 4.3	35 ± 3.8	**29 ± 3.3**	–	–
2177	16 ± 6.4	7.8 ± 3.4	14 ± 3.5	**10 ± 2.7**	–	–	–	–
2179	47 ± 8.8	64 ± 6.0	64 ± 4.9	70 ± 4.1	**73 ± 3.5**	–	–	–
2180	69 ± 8.2	83 ± 4.7	89 ± 3.3	75 ± 3.8	**79 ± 3.3**	–	–	–
2183	97± 3.1	71 ± 6.1	53 ± 5.4	65 ± 4.4	66 ± 3.9	**61 ± 3.0**	–	–

Table 4.3: Summary of the most probable HP2.5 (%) estimate along with its uncertainty obtained from 21 patients as a track of measurements become available. The bold numbers denote the best HP2.5 estimate based on all available tracks.

Patient	HP10$^{(1)}$	HP10$^{(2)}$	HP10$^{(3)}$	HP10$^{(4)}$	HP10$^{(5)}$	HP10$^{(6)}$	HP10$^{(7)}$	HP10$^{(8)}$
2136	64 ± 8.4	34 ± 6.1	25 ± 4.5	**30 ± 4.1**	–	–	–	–
2144	72 ± 7.9	73 ± 5.5	74 ± 4.5	**75 ± 3.8**	–	–	–	–
2149	100	91 ± 3.6	84 ± 3.7	77 ± 3.7	**68 ± 3.7**	–	–	–
2152	72 ± 7.9	41 ± 6.1	60 ± 5.0	69 ± 4.1	55 ± 3.9	**54 ± 3.6**	–	–
2154	47 ± 8.8	38 ± 6.1	45 ± 5.1	**45 ± 4.4**	–	–	–	–
2156	81 ± 7.1	87 ± 4.2	77 ± 4.3	73 ± 3.9	**69 ± 3.7**	–	–	–
2157	3.1 ± 3.1	13 ± 4.1	28 ± 4.6	**21 ± 3.6**	–	–	–	–
2161	0	0	**0**	–	–	–	–	–
2165	0	29 ± 6.3	19 ± 4.4	15 ± 3.5	**35 ± 4.1**	–	–	–
2166	91 ± 5.2	72 ± 5.6	61 ± 5.0	**59 ± 4.4**	–	–	–	–
2167	100	70 ± 5.7	78 ± 4.2	76 ± 3.8	79 ± 3.2	**80 ± 2.9**	–	–
2169	72 ± 7.9	86 ± 4.3	81 ± 4.0	84 ± 3.2	**85 ± 3.0**	–	–	–
2171	32 ± 8.4	32 ± 5.9	48 ± 5.1	61 ± 4.3	68 ± 3.8	**70 ± 3.4**	–	–
2172	78 ± 7.3	85 ± 4.7	76 ± 4.5	80 ± 3.6	**66 ± 3.8**	–	–	–
2174	78 ± 7.3	89 ± 3.9	90 ± 3.1	91 ± 2.5	81 ± 3.1	81 ± 2.8	73 ± 3.0	**64 ± 3.0**
2175	48 ± 9.0	68 ± 5.9	62 ± 5.0	68 ± 4.1	69 ± 3.7	**74 ± 3.2**	–	–
2176	0	25 ± 5.5	34 ± 4.8	42 ± 4.4	40 ± 3.9	**36 ± 3.5**	–	–
2177	38 ± 8.6	19 ± 4.9	24 ± 4.4	**18 ± 3.4**	–	–	–	–
2179	100	98 ± 1.6	93 ± 2.7	93 ± 2.3	**94 ± 1.9**	–	–	–
2180	91 ± 5.2	95 ± 2.6	97 ± 1.8	84 ± 3.3	**87 ± 2.7**	–	–	–
2183	100	85 ± 4.8	68 ± 5.0	76 ± 3.9	79 ± 3.3	**83 ± 2.8**	–	–

Table 4.4: Summary of the most probable HP10 (%) estimate along with its uncertainty obtained from 21 patients as a track of measurements become available. The bold numbers denote the best HP10 estimate based on all available tracks.

4.3.2 The immunohistochemical assay of CAIX

In addition to the direct pO_2 measurements, our collaborators at the Princess Margaret Hospital provided us with immunohistochemical assay of carbonic anhydrase IX (CAIX) data from the same patients with invasive cervical carcinomas. A typical size of biopsy is about five millimeters in diameter. The first data set consists of ten patients from whom three to five biopsies were obtained from each patients. Each biopsy is cut into four slices, except for three tumors (2144, 2148, and 2152) whose biopsies are fully sectioned. CAIX content was assessed as a percentage of area positive for the protein within each tumor slice (with protocol described in (Iakovlev et al., 2007)). The percentage of CAIX-positive pixels (area stained / total area) within the three tumors (2144, 2148, and 2152) of the first data set is shown in Fig. 4.1. The data generation for the first data sets is illustrated in Fig. 4.2. Based on each patient's data, we want to infer the proportion of CAIX-positive cells in the whole tumor. Additionally, we want to find the minimum number of biopsies which is sufficient to learn about the tumor hypoxia.

This is again an inference problem which shall be tackled by Bayes' theorem. Let us denote Y be the proportion of CAIX-positive cells within a tumor. In other words, we assume that the amount of CAIX positive staining is a constant in the tumor, which is equivalent to saying that the tumor hypoxia is supposed to be homogeneously distributed throughout the tumor. This is a simplifying assumption that is frequently made in the literature. Nonetheless, it is a very optimistic assumption which will be discussed in chapter six. Now our knowledge is encapsulated in the form of a posterior probability density function (pdf) of Y,

$$P(Y|D^{(i)}, I) \propto P(D^{(i)}|Y, I) \times P(Y|I), \tag{4.4}$$

which follows from the Bayes' theorem (i indicates the biopsy number). Since we have no information regarding the proportion of CAIX-positive cells, we assign a

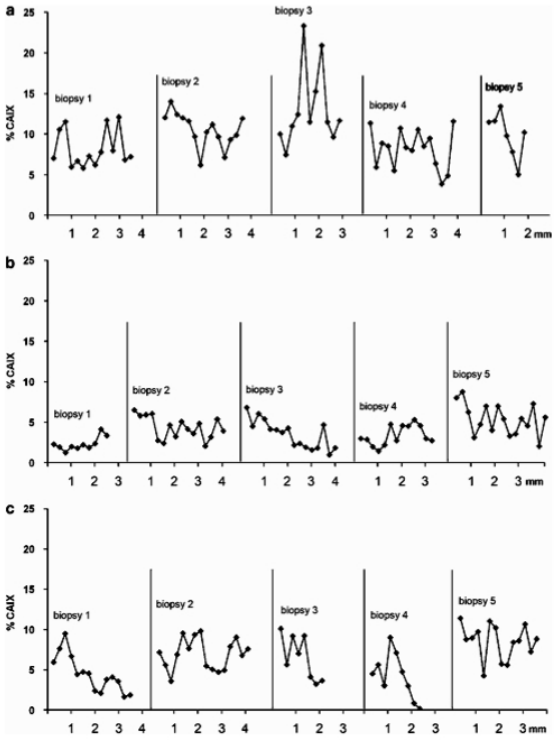

Figure 4.1: Percentage of CAIX-positive pixels within three (a, b, and c) tumors with fully sliced biopsies. Each (a, b, and c) panel is for one patient, five biopsies per patient, and each point gives CAIX value within an individual slice. The slices are shown in the sequence in which they were cut. Adapted from (Iakovlev et al., 2007)

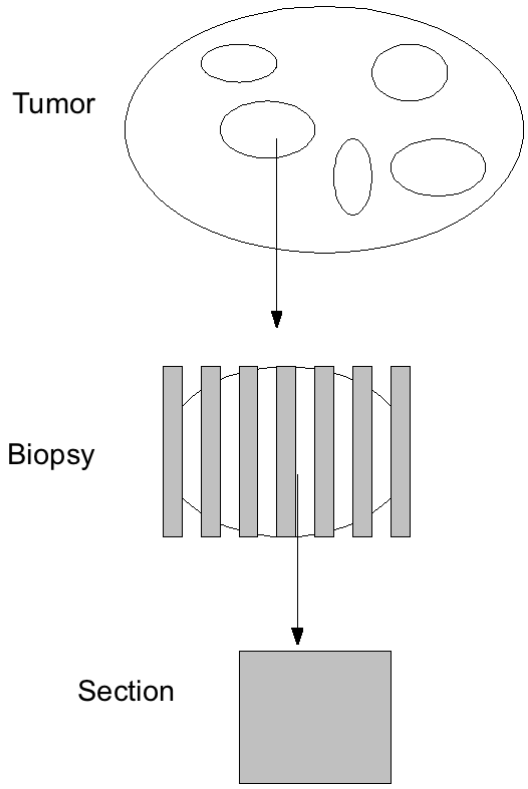

Figure 4.2: The data generation protocol for the first data sets

uniform prior,

$$P(Y|I) = \begin{cases} 1, & \text{for } 0 \leq Y \leq 1 \\ 0 & \text{otherwise} \end{cases} . \tag{4.5}$$

The direct probability however is different than in the case of pO_2 measurements. The stained tumor slices are imaged by fluorescence using a TISSUEscope laser scanning microscope (Iakovlev et al., 2007). A threshold for determining CAIX positivity within a tumor slice is set by a pathologist (Dr. Vladimir Iakovlev). Thus, there are occasions of over- and under-estimating the proportion of CAIX-positivite cells. This inaccuracy (noise) can be assumed to follow a Gaussian distribution with zero mean and s unknown standard deviations. The CAIX data within ith biopsy is modeled as

$$D_j^{(i)} = Y + \epsilon_j, \tag{4.6}$$

where $j = 1, 2, ..., n$ is the number of slices within a biopsy (i). Starting with the

first biopsy, the likelihood function is given by

$$\begin{aligned} P(D^{(1)}|Y,I) &\propto \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}s} \exp\left\{-\frac{1}{2s^2}(D_j^{(1)}-Y)^2\right\}, \\ &\propto \frac{1}{s^n} \exp\left\{-\frac{1}{2s^2}\sum_{j=1}^{n}(D_j^{(1)}-Y)^2\right\}, \end{aligned} \tag{4.7}$$

assuming each slice is independent of one another, which is reasonable because the slices were cut far enough from each other (Iakovlev et al., 2007). The reason why the likelihood function takes this form is that, as was mentioned earlier, that the noise follows a Gaussian distribution; hence, the probability of the noise $\epsilon_j^{(i)}$ is

$$\begin{aligned} P(\epsilon_j^{(1)}|Y,I) &= \frac{1}{\sqrt{2\pi}s} \exp\left\{-\frac{1}{2s^2}(\epsilon_j^{(1)})^2\right\}, \\ &= \frac{1}{\sqrt{2\pi}s} \exp\left\{-\frac{1}{2s^2}(D_j^{(1)}-Y)^2\right\}, \end{aligned} \tag{4.8}$$

which is exactly Eq. 4.7. Assuming that the noise is independent among the slices, the above equation simplifies to Eq. 4.7,

$$P(\epsilon^{(1)}|Y,I) \propto \frac{1}{s^n} \exp\left\{-\frac{1}{2s^2}\sum_{j=1}^{n}(D_j^{(1)}-Y)^2\right\}. \tag{4.9}$$

The posterior probability density function (pdf) for the proportion of CAIX-positive cells based on the first biopsy is given by a joint conditional probability density $P(Y,s|D_j,I)$ because s is also unknown. The presence of the nuisance parameter s can be handled easily by marginalization, namely

$$P(Y|D^{(1)},I) = \int_R P(Y,s|D^{(1)},I)ds, \tag{4.10}$$

where R is the range of s. The choice of R is not important in the case of a parameter estimation problem (however, it is crucial if we are doing a model selection).

Applying Bayes' theorem to the integrand of Eq. 4.10 we have

$$\begin{aligned} P(Y|D^{(1)},I) &= \int_R P(D^{(1)}|Y,s,I)\, P(Y,s|I)ds, \\ &= \int_R P(D^{(i)}|Y,s,I)\, P(Y|s,I)\, P(s|I)ds. \end{aligned} \tag{4.11}$$

Now the first factor is given by Eq. 4.7 and in the second factor we may use Eq. 4.5, since by hypothesis $P(Y|s,I) = P(Y|I)$. For the last factor in the integrand we use Jeffrey's prior,

$$P(s|I) = \begin{cases} \frac{1}{s} & \text{for } 0 < s < \infty \\ 0 & \text{otherwise} \end{cases}, \tag{4.12}$$

because s is a scale parameter. Hence, the posterior pdf is (after integrating with a change of variable $t = \frac{1}{s}$)

$$\begin{aligned} P(Y|D^{(1)},I) &= \int_R \frac{1}{s^{n+1}}\, \exp\left\{-\frac{1}{2s^2}\sum_{j=1}^{n}(D_j^{(1)} - Y)^2\right\} ds, \\ &\propto \frac{1}{\left(\sum_{j=1}^{n}(D_j^{(1)} - Y)^2\right)^{\frac{n}{2}}}, \end{aligned} \tag{4.13}$$

which can be put in the form of a Student-t distribution (see Appendix B). The best estimate of the proportion of CAIX-positive cells is obtained by maximizing Eq. 4.13 and the uncertainty of the estimate is the full width at the half maximum of the posterior pdf. For the cases where the posterior is unimodal and symmetric with respect to the maximum, the best estimate of the proportion of CAIX-positive cells within the tumor can be simply encapsulated in the form

$$Y = \hat{Y} \pm \sigma, \tag{4.14}$$

where, as usual, σ is proportional to the FWHM provided that the posterior pdf is very well approximated by Gaussian density. However, as already mentioned, for the cases where the posterior pdf is multimodal (patient 2148), the full posterior pdf

Biopsy	Patient 2144		Patient 2148		Patient 2152	
	$\hat{Y}$ (%)	σ (%)	$\hat{Y}$ (%)	σ (%)	$\hat{Y}$ (%)	σ (%)
1	8.2	0.6	2.3	0.3	4.5	0.6
2	9.5	0.6	2.9	0.6*	6.2	0.6
3	9.9	0.6	3.4	0.4	6.3	0.5
4	9.1	0.5	3.4	0.3	5.9	0.5
5	9.2	0.5	3.7	0.3	6.6	0.4

Table 4.5: Summary of the proportion of CAIX-positive staining from the first data sets where the biopsies are fully-sectioned. The $*$ denotes cases where the posterior pdf is not symmetric.

is the best way to present the inference about the proportion of CAIX-positive cells. The inferences about the tumor hypoxia obtained from the first biopsy of the three patients constituting the first data set are given by the first row of Tab. 4.5. Next, we apply the same algorithm to the second biopsy. This time, however, we are not completely ignorant about the value of our parameter Y, and so the uninformative prior of Eq. 4.5 is no longer appropriate. In fact, since what we have learned about Y from the first biopsy is given by Eq. 4.13, we use this posterior as the new prior in Bayes' theorem, and as a result we obtain the new posterior $P(Y|D^{(2)}, I)$ which incorporates information from both the first and the second biopsies. This sequential analysis is repeated for the third, fourth, and fifth biopsies, and the final posterior $P(Y|D^{(5)}, I)$ will contain all the information that can be extracted from all five biopsies (under the assumption of homogeneous distribution of hypoxia). The results of the first data sets analysis are summarized in Tab. 4.5. For the biopsies that are fully sectioned, we show the posterior probability density functions (pdfs) in Fig. 4.3. As it is evident from Fig. 4.3 and Tab. 4.5, from the first data set (patient 2144, 2148, and 2152) whose biopsies are fully sectioned, two biopsies seem sufficient to learn about the oxygenation status in the whole tumor. The importance of this result will be discussed in chapter six.

The second data set was generated by a different protocol (Iakovlev et al., 2007). Shortly, two biopsies from each of the 24 patients are sectioned three levels: the first and the second level are 250 μm apart and the third level is taken from

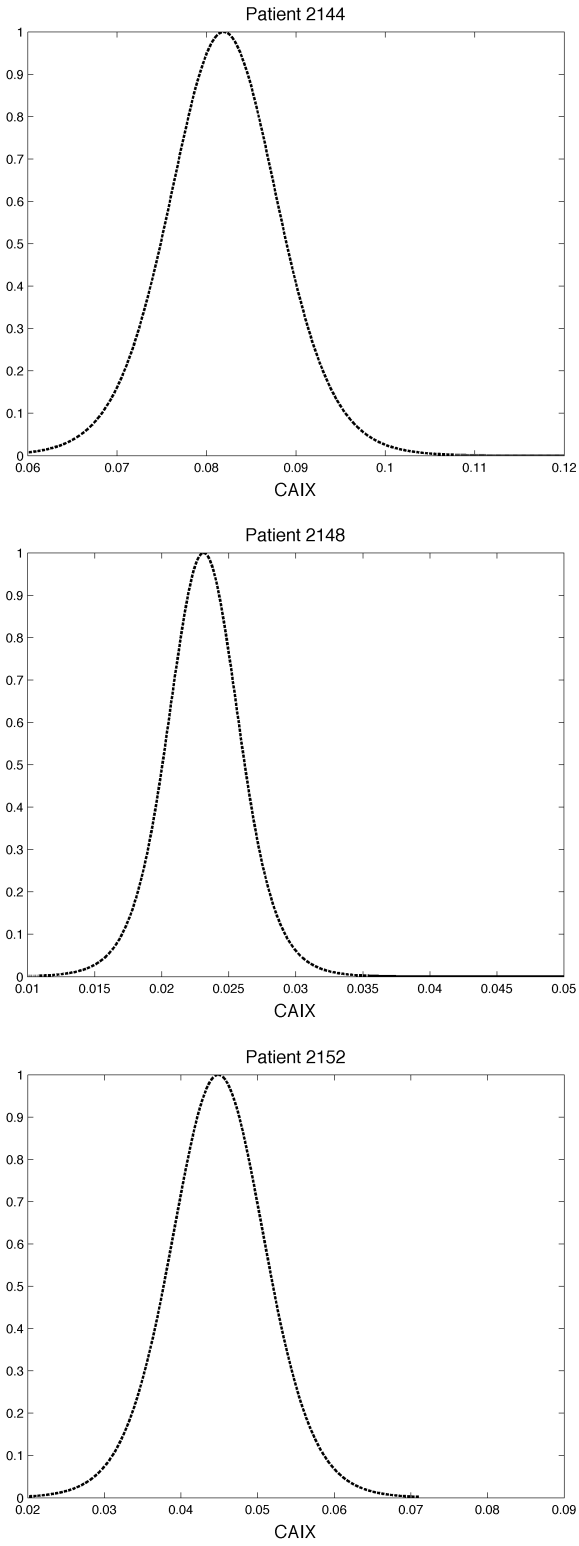

Figure 4.3: The posteriors of the first biopsy obtained from three patients of the first data set

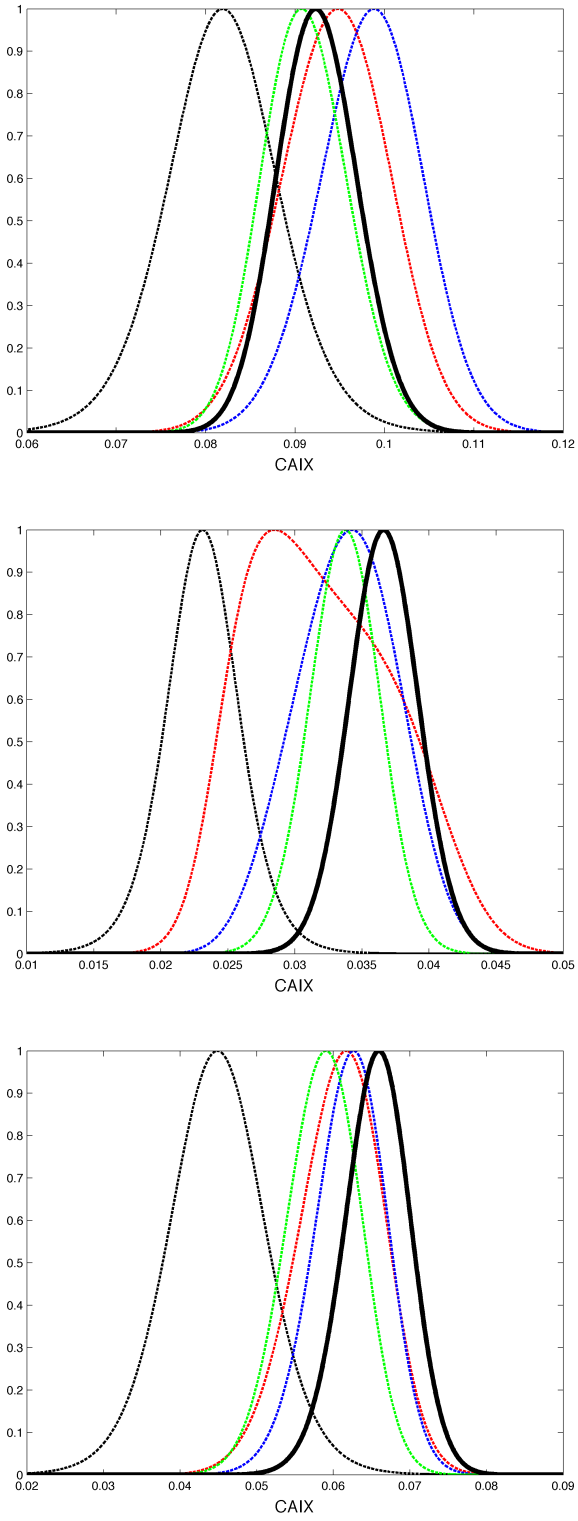

Figure 4.4: The posteriors for each of the five biopsies (black dashed: first biopsy; red dashed: second biopsy; blue dashed: third biopsy; green dashed: fourth biopsy; bold solid: fifth biopsy)

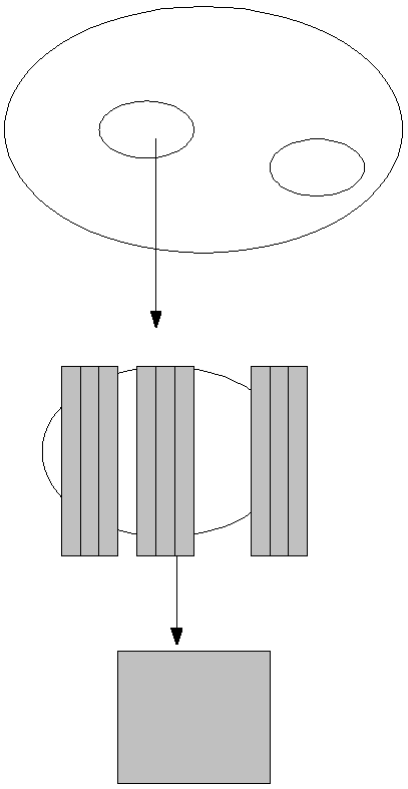

Figure 4.5: The data generation protocol for the second data sets

the opposite side of the biopsy (Fig. 4.5). CAIX content was also assessed as a percentage of area positive for the protein within the tumor tissue. The Bayesian analysis of this second data set proceeds in exactly the same manner as in the first case. The difference, of course, is that we have only two biopsies per tumor; hence, the knowledge we can infer from the data will be encapsulated by the posterior $P(Y|D^{(2)}, I)$, and the most plausible values of Y obtained are shown in Tab. 4.6. Given the relatively large number of patients, we are not showing the graphs of all posteriors, but only a few representative samples in Fig. 4.7. The top three panels show typical examples of symmetric posteriors in which the contribution in the increase of our knowledge from the second biopsy is either minimal or clearly visible. Overall, 17 out of 24 cases have a similar behavior. The bottom two panels of Fig. 4.7, however, show two of the several cases where the result or our inferences cannot be summarized simply by the two number $\hat{Y}$ and σ. Rather than illustrating all the pdfs, we show a subset of the pdfs; some of the cases where two biopsies

Patient	Biopsy 1 $\hat{Y}$ (%)	Biopsy 2 $\hat{Y}$ (%)
2142	9.3 ± 0.6	**7.1 ± 0.4**
2144	14.5 ± 0.3	**14.6 ± 0.3**
∗ 2149	13.3 ± 1.8	**10.0 ± 1.4**
215	6.1 ±0.8	**6.6 ± 0.8**
2152	8.6 ± 0.6	**8.7 ± 0.6**
2154	3.9 ± 0.7	**4.1 ± 0.7**
2156	2.6 ± 0.3	**2.7 ± 0.3**
2157	2.4 ± 0.5	**2.8 ± 0.6**
2161	17.3 ± 0.8	**17.5 ± 0.8**
2163	1.7 ± 0.2	**1.7 ± 0.2**
2165	0.41 ± 0.02	**0.42 ± 0.02**
2166	2.9 ± 0.2	**3.0 ± 0.2**
2167	1.0 ± 0.2	**1.1 ± 0.2**
∗ 2169	17.2 ± 2.6	**5.7 ± 0.6**
∗ 2170	9.4 ± 1.8	**0.38 ± 0.1**
2171	21.8 ± 0.7	**8.8 ± 0.5**
∗ 2172	35.8 ± 1.5	**36.9 ± 1.3**
2174	12.7 ± 0.4	**12.9 ± 0.4**
∗ 2175	27.8 ± 1.7	**26.2 ± 1.5**
2176	16.2 ± 0.5	**15.9 ± 0.6**
2177	0.49 ± 0.04	**0.49 ± 0.04**
2179	7.1 ± 2.0	**6.4 ± 0.8**
∗ 2180	51.8 ± 1.6	**13.8 ± 1.4**
∗ 2183	26.0 ± 1.3	**28.7 ± 1.1**

Table 4.6: Summary for the best estimate of the overall CAIX percentage within a cervical carcinoma along with the reliability of the estimate based on three-level sampling protocol. The ∗ denotes cases in which the inference $Y = \hat{Y} \pm \sigma$ is not reliable.

suffice and a couple of cases where two biopsies are not sufficient (Fig. 4.7).

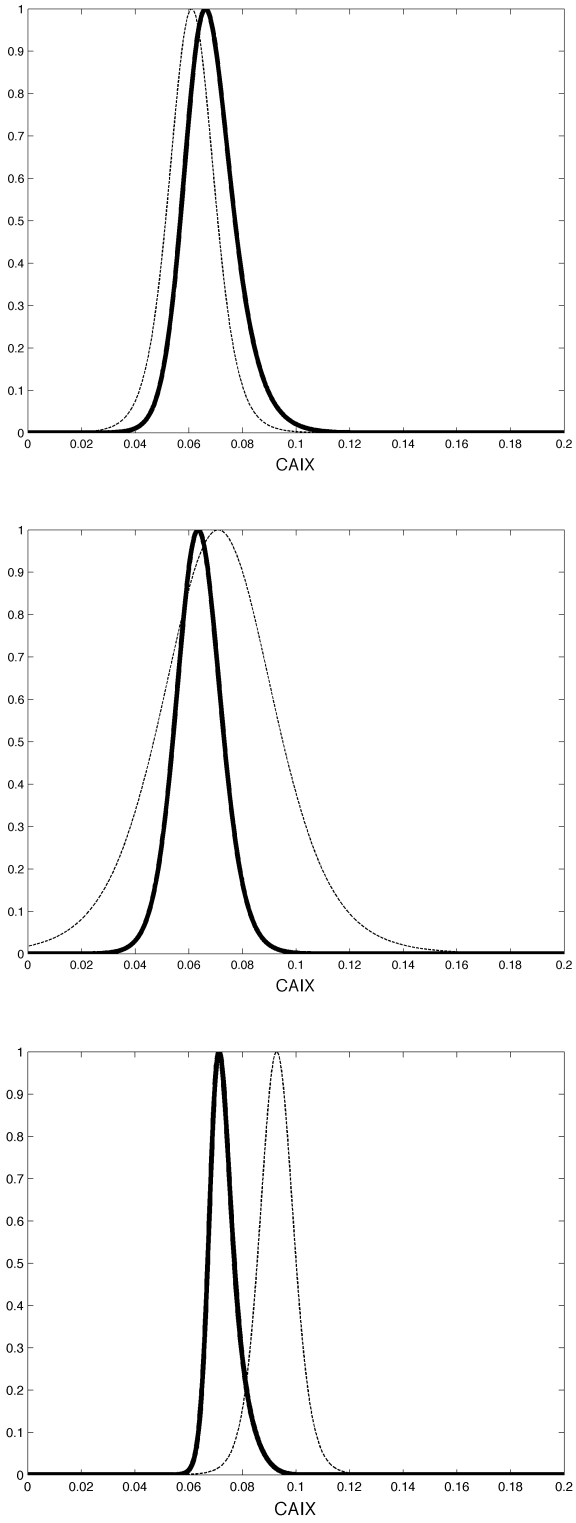

Figure 4.6: Examples of $P(Y|D^{(2)})$ from the second data set which results in symmetric and unimodal pdf (thin dashed: first biopsy; solid: second biopsy).

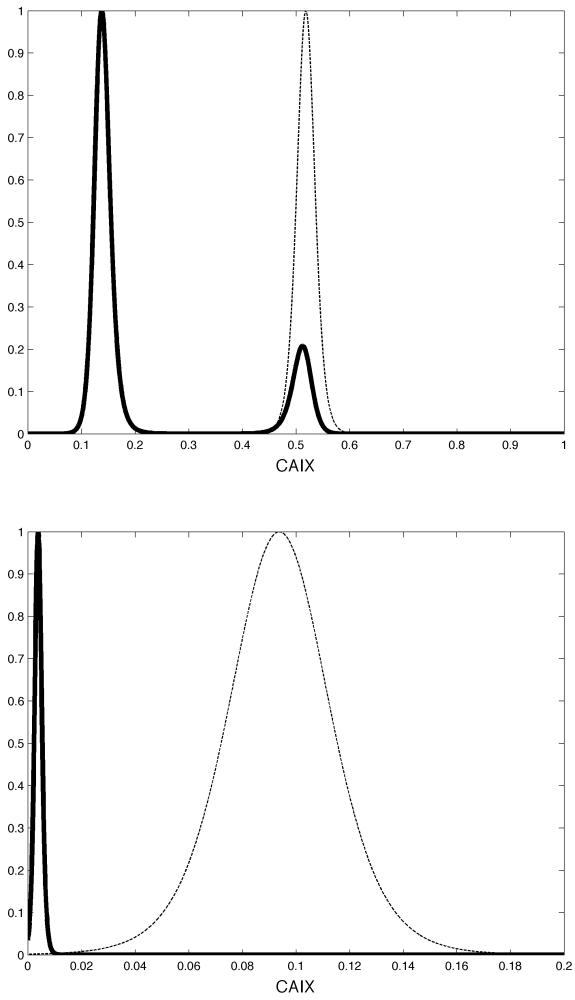

Figure 4.7: Examples of $P(Y|D^{(2)})$ which results in bimodal or truncated pdf (thin dashed: first biopsy; solid: second biopsy).

4.4 Correlation and concordance

The next important step in tumor hypoxia research is to find out whether or not the two tumor hypoxia quantifications tell the same story about hypoxia within a tumor. In other words, we need to ascertain whether or not the indirect immunohistochemical assays of CAIX protein marker is correlated with the direct pO_2 measurements. This problem has been tackled by numerous researchers, but so far no consensus has been reached.

Loncaster and her groups was the first to demonstrate that CAIX was correlated with tumor oxygen measurements in cervical carcinomas, based on a prospective study of 68 patients (Loncaster et al., 2001); however, other research groups were not able to reproduce the conclusion. Some research groups (Hedley et al., 2003; Mayer et al., 2005) showed that CAIX expression and tumor oxygenation status do not correlate in cervical carcinomas. In addition to the heterogeneity of the tumors, the statistical analyses performed by each research group might be different. This provides challenges in order to confirm other studies.

A Bayesian approach to the problem does not require any *ad-hoc* procedures. Assuming that both HP5 and CAIX staining measure the amount of hypoxia in a tissue, we denote by X the former and by Y the latter. Next we introduce the joint pdf $P(X, Y|D_x, D_y, I)$, where D_x and D_y are the available data for HP5 and CAIX respectively, and I is the background information. With this pdf the concept of error-bar in the one-dimensional case can be easily extended by the introduction of the *variances* of X and Y, namely

$$\sigma_X^2 = \int\int_R (X - \hat{X})^2 \, P(X, Y|D_x, D_y, I) dX dY, \tag{4.15}$$

$$\sigma_Y^2 = \int\int_R (Y - \hat{Y})^2 \, P(X, Y|D_x, D_y, I) dY dX, \tag{4.16}$$

and the idea of variance can, in turn, be extended by introduction of the *covariance*,

$$\sigma_{XY}^2 = \int\int_R (X - \hat{X})(Y - \hat{Y}) \, P(X, Y|D_x, D_y, I) dX dY, \tag{4.17}$$

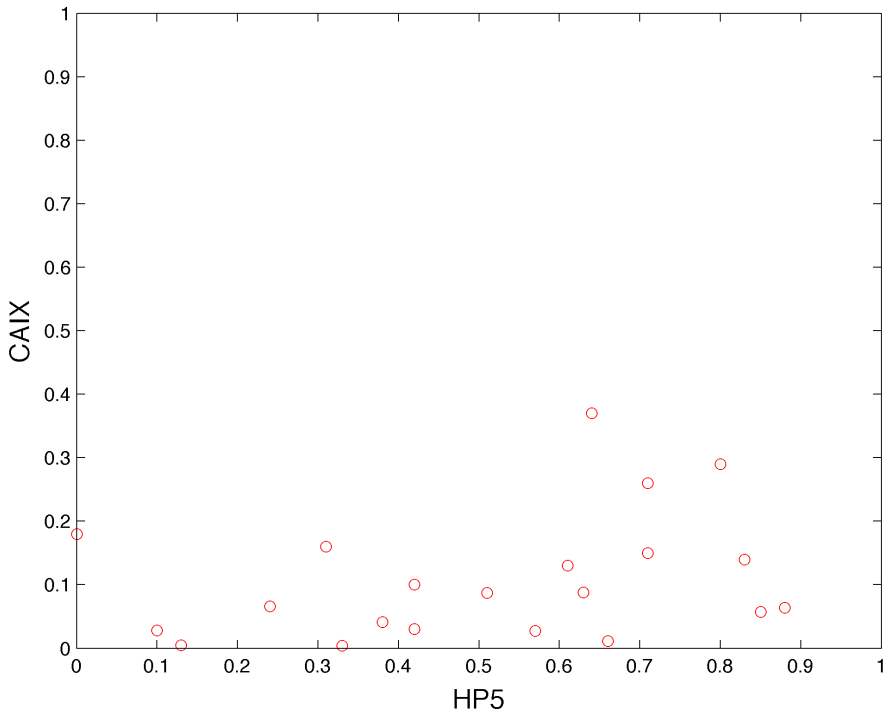

Figure 4.8: A scatter-plot for the HP5 versus CAIX in 21 patients

where R is the range of X and Y and $\hat{X}$ and $\hat{Y}$ are the best estimates of the hypoxic proportion by HP5 and the CAIX-positive cells respectively. If this were applied to the two hypoxia measurements of each patient, we would get $\sigma^2_{XY} = 0$. This is due to the fact that the joint posterior pdf, $P(X, Y|D_x, D_y, I)$ factors into the separate pdf for X and Y, since the two hypoxia measurements are independent.

In other words, we would be asking the wrong question. A more appropriate question is to ask whether a scatter-plot of X and Y for *all* patients shows any correlation. In Fig. 4.8, we display the best estimate of the percentage of CAIX-positive staining against the best estimate of the HP5. The error bars are neglected for a reason that will be explained later on. Having pooled all the data from 21 patients, we would like to observe the correlation between the HP5 and the CAIX. We can then fit an ellipse on the data and then investigate whether the major axis of the ellipse has a positive or negative slope (Sivia, 2006). To begin with, we use

the general equation of a conic,

$$aX^2 + bXY + cY^2 + dX + eY + f = 0. \tag{4.18}$$

If the discriminant,

- $b^2 - 4ac < 0$, we have an ellipse, or circle;
- $b^2 - 4ac = 0$, we have a parabola;
- $b^2 - 4ac > 0$, we have a hyperbola.

Furthermore, one may rewrite Eq. 4.18 in the following way:

$$\mathbf{x}^T \, A_Q \, \mathbf{x} = 0, \tag{4.19}$$

where $\mathbf{x}^T = (x \quad y \quad 1)$ and

$$A_Q = \begin{pmatrix} a & b/2 & d/2 \\ b/2 & c & e/2 \\ d/2 & e/2 & f \end{pmatrix}; \tag{4.20}$$

the subscript Q stands for quadratic.

We can proceed with the similar procedures as in the case of parameter estimation (Section 3.1). We can model the data X_i in general as

$$X_i^2 = [AY_i^2 + BX_iY_i + CX_i + DY_i + E] + \epsilon_i, \qquad i = 1, 2, ..., 21 \tag{4.21}$$

where ϵ_i is the noise. Eq. 4.21 is related to Eq. 4.18: $a = 1$, $A = -c$, $B = -b$, $C = -d$, $D = -e$, and $E = -f$. As it was explained earlier, assuming the noise has an unknown second moment and using uninformative prior for the parameters but Jeffrey's prior for the standard deviation of the noise, we have the following

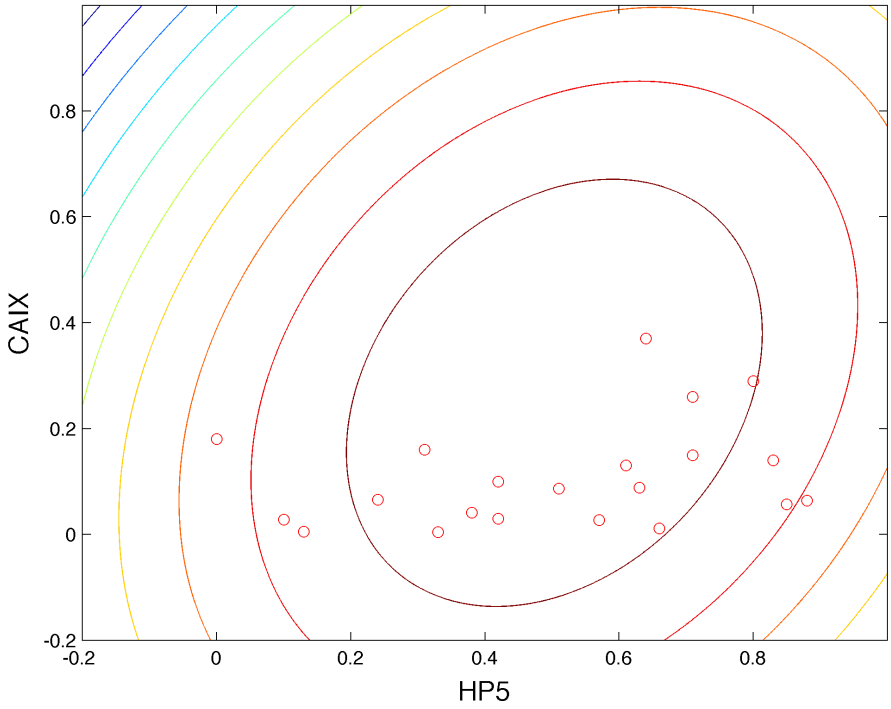

Figure 4.9: Ellipse-fitting to the data

posterior pdf,

$$P(\{A,...,E\}|D,I) \propto \frac{1}{\left[\sum_{i=1}^{21}\left(X_i^2 - \{AY_i^2 + BX_iY_i + CX_i + DY_i + E\}\right)^2\right]^{\frac{21}{2}}}. \tag{4.22}$$

Using the optimization technique explained in Section 3.1, we can obtain the most plausible value for each estimate. Fig. 4.9 illustrates the fitting according to our proposed model. Having obtained the most plausible values for each parameter, we can write the matrix representation of the conic section,

$$A_Q = \begin{pmatrix} 1 & -0.22 & 0.037 \\ -0.22 & 0.59 & -0.11 \\ 0.037 & -0.11 & -0.068 \end{pmatrix}. \tag{4.23}$$

In order to analyze the nature of the ellipse, we can investigate the first minor of

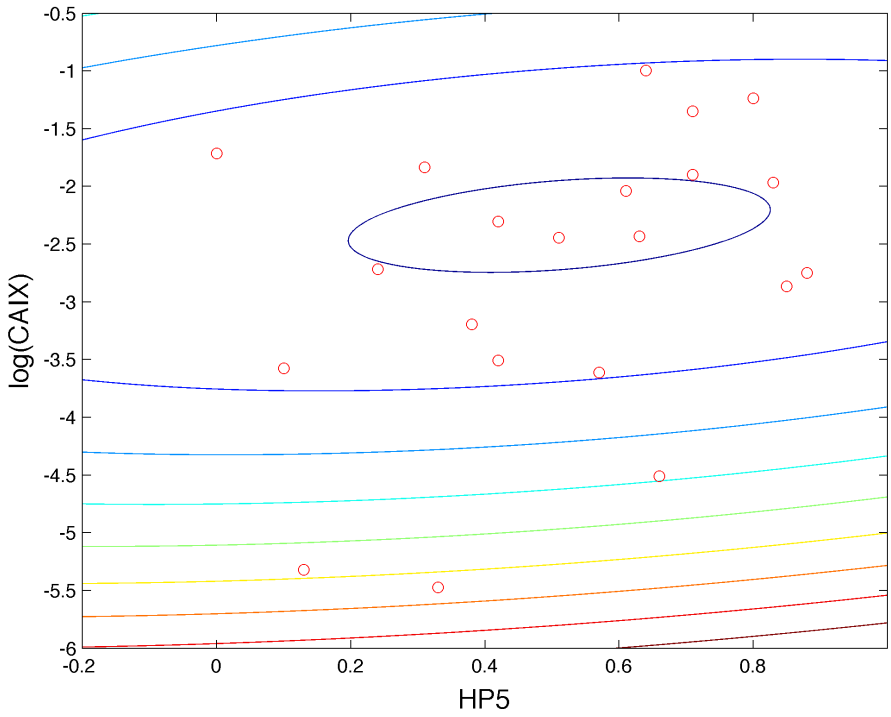

Figure 4.10: Ellipse-fitting to the data with log(CAIX)

A_Q:

$$A_1 = \begin{pmatrix} 1 & -0.22 \\ -0.22 & 0.59 \end{pmatrix}. \tag{4.24}$$

The principal axes of the ellipse can be determined from the eigenvectors of A_1. Furthermore, the inverse of A_1 is related to the covariance matrix, whose off-diagonal entries indicate the correlation coefficient between X and Y (Sivia, 2006). From these pooled data, we have a correlation coefficient of 0.40, which is similar to the result obtained by the PMH group (Iakovlev et al., 2007).

For better visualization purposes, one can transform the CAIX values into their logarithm space and then proceed with the similar conical fitting. In doing so, we obtain Fig. 4.10 and a slightly stronger correlation coefficient of 0.47. One has to keep in mind that the axes in Fig. 4.10 are not in the same scale.

Let us refer back to the statement that "the error-bars on each datum can be neglected". Had we included the error-bars on each datum, we would have had a

very much smaller ellipse, whose axes are perpendicular, around each datum. One of the possible approaches is then to perform a convolution between each small ellipse and a much larger one encompassing all the data. Nonetheless, we realize that a convolution between two functions, one of which is much larger than the other, would only give us the larger function (Gregory, 2005). Our conclusion would not have changed, had we included the added complexity of the error-bars around each datum.

In addition to correlation statistic, many medical researchers have used a concordance statistic to measure the degree of agreements between two raters (Feinstein, 2002). This technique can be implemented in tumor hypoxia research. We begin by setting a threshold for which the tumor is classified as hypoxic. For example, according to the PMH group a reasonable choice is $\geq 8\%$ for the CAIX measurement and $\geq 50\%$ for the HP5. Given this much, we can proceed with our Bayesian approach as follows. Let $H \in [0, 1]$ denote the concordance measure. If $H = 0$, then the two techniques are in perfect discordance; whereas, if $H = 1$, then the two techniques are in perfect concordance. Thus, by the application of Bayes' theorem, we have

$$P(H|\{\text{data}\}, I) \propto P(\{\text{data}\}|H, I) \times P(H|I), \tag{4.25}$$

where I denotes our background information, for instance, the assumption of homogeneous tumor can be included in I. Since we have no additional information for the concordance, we use the uniform prior probability for H,

$$P(H|I) = \begin{cases} 1, & \text{for } 0 \leq H \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.26}$$

The direct probability of obtaining the data 'r agreements; i.e., both analyses indicate hypoxic tissue (when the CAIX percentage is $\geq 8\%$ and the HP5 is $\geq 50\%$) and non-hypoxic tissue (when the CAIX percentage is $< 8\%$ and the HP5 is $< 50\%$)

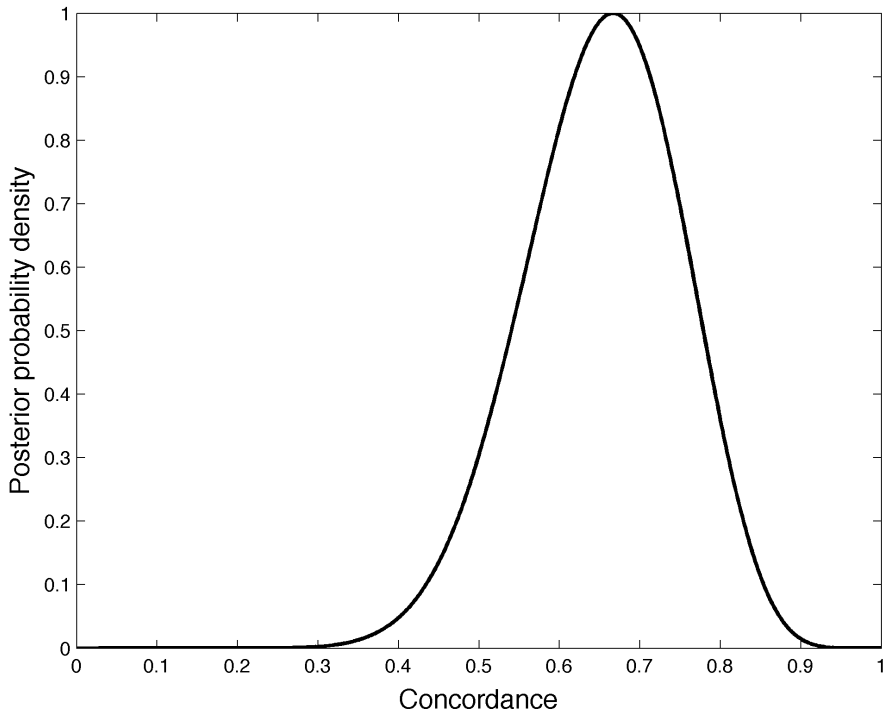

Figure 4.11: The posterior pdfs for the concordance between the estimate of the proportion of CAIX-positive cells and the HP5 in 21 patients.

in N patients' follows a binomial distribution:

$$P(\{\text{data}\}|H, I) \propto H^r \, (1 - H)^{N-r}. \tag{4.27}$$

Finally, substituting Eg. 4.26 and 4.27 into the Bayes' theorem (Eq. 4.25) gives us a posterior probability density function (pdf) for the concordance and the best estimate of the concordance,

$$H = 0.67 \pm 0.10. \tag{4.28}$$

Thus, our updated state-of-knowledge about this type of concordance between the CAIX immunohistochemical analysis and the direct pO_2 measurements in quantifying tumor hypoxia is depicted in Fig. 4.11. Of course, if the thresholds are chosen differently, then this estimate will change. Nonetheless, this definition of

concordance appears to be quite useful in a medical setting.

Chapter 5

Data analysis of phase II clinical trials of relapsed ovarian cancers

Before we proceed with the analysis, it is useful to provide some background knowledge for this clinical study. We will begin with a brief overview of clinical trials and ovarian cancers. For readers who are familiar with these concepts, the next two sections may be skipped and proceed directly to the third section of this chapter.

5.1 A brief review of clinical trials

According to some scholars, the origin of clinical trials dates as early as the year 980 when Avicenna, an Arabian physician and philosopher, proposed some rules to evaluate the effects of drugs on diseases. In his *Canon of Medicine*, he suggested that

> "a remedy should be used in its natural state, with uncomplicated disease, and should be observed in two 'contrary type of disease' ... the time of action and reproducibility of the treatment effect should be studied" (Meinert, 1986)

The formal definition and constitution of a clinical trial however have not been proposed until centuries later. There are several definitions of a clinical trial; however,

Everitt and Pickles (Everitt and Pickles, 2004) define a clinical trial as

> "any form of planned experiment designed to assess the most appropriate treatment of patients with a particular medical condition, where the outcome in a group of patients treated with the test treatment are compared with those observed in a similar group of patients receiving control treatment, and patients in both groups are enrolled, treated, and followed over the same period."

Consequently, studies involving animals or carried out in vitro using human biological substances cannot be regarded as clinical trials. The first recorded clinical trial was conducted by James Lind in 1747 while at sea on board the *Salisbury*. For the full historical account of clinical trials, interested readers are directed to (Meinert, 1986). Tab. 1.1 of (Everitt and Pickles, 2004) briefly summarizes historical events in the development of clinical trials.

Early clinical trials allocated their patients into test and control groups (also known as treatment arms) in an arbitrary and non-systematic scheme. Fisher then introduced the concept of *randomization* in clinical trials and the first properly randomized group was for the clinical trial of streptomycin in the treatment of pulmonary tubercolosis (Everitt and Pickles, 2004). Since then randomized clinical trials have become a standard procedure in the evaluation of new drugs or treatments. The justification for randomized clinical trials is that "[They] nicely illustrate the desire of modern democratic society to justify [their] medical choices on the basis of the objectivity inherent in statistical and quantitative data" (Everitt and Pickles, 2004).

Randomization is regarded as the primary requirement in any acceptable clinical trial. Another necessary requirement for an acceptable clinical trial is *blinding*. The fundamental idea of blinding is that trial participants, physicians, and those who are involved with their management and data collection, shall not be influenced by knowledge of the assigned treatment. Blinding is a way to minimize the possibility of bias (Everitt and Pickles, 2004). There are two types of blinding processes:

1. Single-blinding, in which only the patients who are unaware of which treatment(s) they are receiving;
2. Double-blinding, when both the patients and the investigators are unaware of the patients' treatment(s).

The most common design for a clinical trial is the *fixed sample size parallel groups design* with random allocation of patients to treatment(s). This design is easy to implement; however, a large number of patients are needed to estimate reliably the magnitude of any treatment difference. We shall keep in mind that the clinical trial design has been formulated under the frequentist methodology; as a result, a large number of sample size is necessary to minimize the effect of 'chance'. In the parallel group design, each clinical trial group is administered one treatment. A modification of the parallel group design is the *cross-over design*, in which each patient receives two treatments; for example, in the 2×2 cross-over design, one group receives two treatments in the order AB and the other group receives in the order BA, with patients randomly allocated to the two groups (Senn, 1993). The second treatment must be administered once the effect of the first treatment has disappeared. This requirement can be difficult to satisfy; hence, a cross-over design is rarely used. A more general clinical trial design is the *factorial design*, proposed by the late David Byar in 1992. He stated that "such designs may offer impressive gains in efficiency compared with series of trials studying one treatment at a time. This is especially true when treatements do not interact with one another" (Byar et al., 1993). In a factorial design, several treatments are considered at the same time. The number of clinical trial groups (treatment arms) is 2^n, where n is the number of test treatments. Lubsen and Pocock conducted a clinical trial with a factorial design for three test treatments, ISIS-4 trial, whose protocol is summarized as follows:

> "Patients were simultaneously randomized to each of the three active treatments or their respective controls (captopril or its placebo, mononitrate or its placebo, magnesium or no magnesium) in ... a complete $2 \times 2 \times 2$ factorial (sometimes called a three-way factorial). Note that

> there are, clinically speaking $2 \times 2 \times 2 = 8$ treatment groups: captopril plus mononitrate plus magnesium, captopril plus mononitrate, captopril plus magnesium, mononitrate plus magnesium, captopril alone, mononitrate alone, magnesium alone, and none of these (placebo)." (Lubsen and Pocock, 1994)

Those three clinical trial designs are the ones commonly conducted; randomization and blinding can easily be implemented in those designs as long as no surgery is involved. There are however other clinical trial designs such as *open design* and *orphan design*, for which randomization and blinding cannot be guaranteed. The open design is usually for any clinical trial involving a surgical procedure; whereas, the orphan design is conducted for testing drugs or treatments to treat diseases affecting fewer than 200,000 peoples worldwide or rare genetic diseases (Health, 2000). Due to the small sample size, randomization is difficult for this case.

Whichever clinical trial the design investigators choose, it has four well-known phases (Everitt and Pickles, 2004):

1. Phase I: Clinical pharmacology and toxicity;
2. Phase II: Initial clinical investigation for treatment effect;
3. Phase III: Full-scale evaluation of treatment; and
4. Phase IV: Post-marketing surveillance.

The first phase of clinical trials deal primarily with drug safety and toxicity. One of the objectives is to determine the maximum tolerated dose (MTD), obtained from dose-escalation experiments. Additionally, investigators also conduct studies of drug metabolism and bioavailability. Typically, this phase consists of about 20 - 80 participants who are in a healthy condition. The second phase of clinical trials continues to monitor the safety of the drugs or treatments and their effectiveness in a larger sample size (typically about 100 - 200 patients). In this phase, drugs or treatments are screened; those which are effective are selected to proceed to the next phase. The primary objectives in the phase II of clinical trials are to

- identify the group of patients that can benefit from the drug or treatment and to

- verify and estimate the effectiveness of the drug dosage determined in the previous phase (Everitt and Pickles, 2004).

The third phase of clinical trials is the one which is usually made public. This phase consists of a substantial number of patients (usually greater than 200 patients from various hospitals) and it is the most expensive phase of clinical trials. In this phase, a new drug or treatment is evaluated against the current standard treatment(s). Under the standard clinical trial protocol, investigators cannot modify the clinical trial once it has already begun this phase. It has to continue until the prescribed endpoint is reached. The reason for this stiffness in clinical trials is to ensure that randomization and blinding processes are preserved for quantitative analysis purposes. Once the phase III is completed, the drug or treatment is determined whether or not it can be marketed. If so, the final phase of clinical trials is to closely monitor the patients using the drug or treatment for any possible adverse effects. Further research may still be required to study the long-term morbidity or mortality caused by the new drug or treatment.

5.2 A brief introduction to ovarian cancer

According to the National Cancer Institute (NCI, 2009), ovarian cancer is defined as

> "cancer that forms in tissues of the ovary (one of a pair of female reproductive glands in which the ova, or eggs, are formed). Most ovarian cancers are either ovarian epithelial carcinomas (cancer that begins in the cells on the surface of the ovary) or malignant germ cell tumors (cancer that begins in egg cells)."

The estimated new cases and deaths from ovarian cancer in Canada in 2008 were 2,500 and 1,700 respectively (CCS, 2009) and in the United States for the same year, the estimated new cases and deaths were 21,650 and 15,520 respectively.

The typical treatment for ovarian cancer includes surgery, chemotherapy, radiation therapy, immunotherapy, or combination of therapies. The current first-line treatment for patients with ovarian cancer is a combination of carboplatin and paclitaxel. Unfortunately, despite advances in therapy, ovarian cancer still remains one of the deadliest gynecological cancers. Here is a statistical summary for ovarian cancers:

> "Less than 30% of women with advanced stage disease survive long-term. When diagnosed in stage I, 90% of patients can be cured with conventional surgery and chemotherapy. At present, only 25% of ovarian cancers are detected in stage I due, in part, to the absence of specific symptoms and to lack of an effective screening strategy" (Badgwell and Bast, 2007) [and] ... 70% of stage IIA (metastases to the uterus or Fallopian tubes) tumors are curable if detected, 70% of women who are diagnosed with stage III and IV tumors have widespread intra-abdominal disease or distant metastases at diagnosis. For these patients, the cure rate plummets to $< 30\%$. ... Ovarian cancer has an overall cure rate of approximately 45%, a relatively discouraging prognosis" (Rustin et al., 2004).

Despite conventional treatments, the mortality rate remains greater than 50% of the diagnosed patients and 80% of the patients will relapse (Rustin et al., 2004). Recent breakthrough in biomedical research have resulted in the development of numerous molecularly-targeted agents that inhibit signal transduction, angiogenesis, and other cellular pathways. Some of these agents, such as bevacizumab, trastuzumab (Herceptin) or Gleevec, are undergoing clinical trials in the hope of continuing to improve the treatment of ovarian cancers.

In addition to advances in cancer treatments, it is evident from the statistics that early detection of ovarian cancers is crucial. Currently, there is no single effective early diagnostic tool for ovarian cancers. The most commonly reported symptoms prior to ovarian cancer diagnosis are "abdominal or pelvic pain, bloating, gastrointestinal distress, and abdominal swelling" (Smith et al., 2005), which can be symptoms for other diseases; hence, misdiagnosis or late diagnosis often occurs in ovarian cancer patients.

The most direct detection tool for early diagnosis is *transvaginal sonography (TVS)*, but TVS alone is incapable to detect ovarian cancer effectively; some tumor locations prohibit the use of such imaging techniques. Additionally, obtaining high resolution biological images is still expensive and an active area of research. Furthermore, one may not be able to distinguish whether tumor masses are benign or malignant solely from their images.

One can determine the characteristic of tumor masses by performing biomarker assays, identifying certain proteins common in cancer cells. Cancer antigen 125 (CA-125), "a heavily glycosylated high-molecular-weight mucin (MUC 16)" (O'Brien et al., 1991; Yin et al., 2002), has been widely used as a serum marker for ovarian cancers. Unlike CAIX, which is a protein present in the extracellular cell membrane, CA-125 is a larger protein that can be easily separated from blood plasma. The discovery of CA-125 serum has been promising and many have been convinced that it is one of the best early indicators for ovarian cancers because

> "CA-125 levels are elevated in $50 - 60\%$ of patients with early stage ovarian cancer and in 90% of patients diagnosed with late stage ovarian cancer (Bast et al., 2005). Overall, significant expression of CA-125 is observed in 80% of ovarian cancers at a tissue level" (Rosen et al., 2005).

The interpretation of serum marker analysis depends heavily on statistical methodologies for the analysis of serum marker data. Two important and commonly used measures in the studies of reliability of diagnostic tools are *sensitivity* and *specificity*. Sensitivity measures the proportion of actual positives correctly identified as positives (true positive). Specificity measures the proportion of negatives correctly identified (true negative). These two measures (Tab. 5.1) are closely related to the concepts of type I (α) and type II (β) errors in orthodox statistics (Armitage et al., 2002),

$$\begin{aligned} \text{sensitivity} &= \frac{\text{number of true positive}}{\text{number of true positive} + \text{number of false negative}}, \\ \text{specificity} &= \frac{\text{number of true negative}}{\text{number of true negative} + \text{number of false positive}}, \end{aligned} \tag{5.1}$$

	Test	
True	+	-
+	$1-\beta$ (sensitivity)	β
-	α	$1-\alpha$ (specificity)
$\downarrow$	$\downarrow$ positive predictive value	$\downarrow$ negative predictive value

Table 5.1: The binary classification for diagnostic tests

In addition to sensitivity and specificity, *positive predictive value* is also a commonly used measure in diagnostic studies; it measures the proportion of true positives,

$$\text{PPV} = \frac{\text{number of true positive}}{\text{number of true positive} + \text{number of false positive}}. \tag{5.2}$$

Thus, a reliable diagnostic tool shall have a high sensitivity, specificity and in turn, a high positive predictive value.

In order to be a reliable diagnostic tool, CA-125 has to satisfy the above requirements: high sensitivity, high specificity, and high positive predictive value. For general population of women (pre-menopausal and post-menopausal), CA-125 has limited sensitivity; moreover, individual values of CA-125 are not sufficiently specific for early detection of ovarian cancers (Badgwell and Bast, 2007). CA-125 level is elevated in response to any irritation of the surfaces of body cavity, the peritoneal cavity, and the pleural cavity. Elevation of CA-125 level may also be caused by other non-malignant diseases such as cirrhosis of the liver, pneumonia, and heart failure. A variety of benign conditions including menstruation, first trimester of pregnancy, and endometriosis can increase CA-125 level. All of these provide challenges for using CA-125 as an early diagnostic tool for ovarian cancers. To enhance both sensitivity and specificity of CA-125, many have studied and performed clinical trials on the combination of CA-125 assay and transvaginal sonography (TVS) concurrently and sequentially. The Prostate, Lung, Colon, and Ovary (PLCO) Screening Trial has conducted ovarian cancer screening clinical trial for post-menopausal women between 55 and 74 years of age: 37,000 patients randomized to the screening group of the trial and another 37,000 patients participating as non-screened controls (Bast

et al., 2005; Hayes et al., 2000). The preliminary report of this study suggested a combined positive predictive value (PPV) of TVS and CA-125 was 23.5%; as opposed to individual PPV of 1% and 3.7% for TVS and CA-125 respectively.

In an effort to improve sensitivity and specificity of early diagnostic tools for ovarian cancers, many have also considered the use of multiple serum biomarkers. There have been many novel biomarkers discovered for the past decade. Badgewell and Bast provide a clear and brief summary for the commonly used and recently discovered biomarkers for ovarian cancers (Badgwell and Bast, 2007). To name a few of those: *mesothelin*, a protein attached to the cell surface thought to have a role in cell adhesion and possibly in cell-to-cell recognition and signalling; *kallikrein*, enzyme that has peptide bond in protein; *osteopontin*, a glycoprotein responsible for cell attachment and wound healing; *vascular endothelial growth factor (VEGF)*, sub-family of growth factors which is important in angiogenesis; *interleukin*, a group of signalling molecules crucial in immune system, and many others. While advances in biomedical research lead to the discovery of many novel biomarkers, the cost of administering them in patients is still very high. Consequently, CA-125 is still actively used as a serum marker for ovarian cancers despite its relatively low specificity.

In the next section, we illustrate how Bayesian methods contribute to analyze the study of phase II clinical trials of relapsed ovarian cancer patients involving molecularly-targeted agents (MTAs) conducted by the Princess Margaret Hospital.

5.3 Clinical trials' data classifications

We have clinical trials' data from four phase II relapsed ovarian cancer clinical trials involving molecularly-targeted agents (MTAs). Each clinical trial consists of eligible eighteen to sixty-five patients. For each clinical trial group, one or a combination of MTAs is administered (Welsch, Wang, Gunawan, Mackay, Nathwani, Yau, Tsang, MacAlpine, Sivaloganathan, and Oza, Welsch et al.). The level of CA-125 serum

marker is monitored regularly according to the Gynocologic Cancer InterGroup (GCIG) guidelines for a relapsed clinical trial. In a relapsed clinical trial, a patient is classified to have a *response* according to a CA-125 serum level if there is "at least a 50% reduction in CA-125 levels from a pretreatment sample and the response must be confirmed and maintained for at least 28 days" (GCIG, 2009).

Additionally, the longest diameter of the tumor is regularly monitored during each follow-up according to the Response Criteria in Solid Tumors (RECIST) guidelines. Without going into details, we summarize some of the important concepts. First, the eligibility criteria for clinical trials' patients are described in the following (RECIST, 2009):

- Only patients with *measurable disease* at baseline (course index 0) should be included in protocols. Measurable disease means that there is at least one measurable lesion;
- all measurements should be taken and recorded in metric notation. All baseline evaluations should be performed as closely as possible to the beginning of treatment and never more than 4 weeks prior to the treatment;
- the same method of assessment and the same technique must be used at baseline and during follow-up; and finally,
- tumor markers alone cannot be used to assess response. If markers are initially above the upper normal limit, they must [be in a normal range] for a patient to be considered in complete clinical response when all lesions have disappeared.

The baseline documentation of "target" lesions shall be performed on lesions with the longest diameter (LD) and lesions which are suitable for accurate repeated measurements. The baseline LD is used as reference to characterize the objective tumor. The eligible clinical trials' patients are classified in the following way:

- **Complete Response (CR):** disappearance of all target lesions
- **Partial Response (PR):** At least 30% decrease in the sum of the LD of target lesions, taking as reference the baseline sum LD
- **Progressive Disease (PD):** At least 20% increase in the sum of LD of target lesions, taking as reference the smallest sum LD recorded since the treatment started or the appearance of one or more new lesions

- **Stable Disease (SD):** Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD, taking as reference the smallest sum LD since the treatment started.

Note that to be assigned a status of PR or CR, changes in tumor measurements must be confirmed by repeat assessments that should be performed no less than 4 weeks after the criteria for response are first met. In the case of SD, follow-up measurements must have met the SD criteria at least once after study entry at a minimum interval (in general, not less than 6-8 weeks) that is defined in the study protocol.

Each eligible patient is classified either 'responsive' or 'non-responsive' according to CA-125 serum level and the RECIST criteria (only CR and PR are classified as 'responsive'). Tab. 5.2 on the next page summarizes the clinical trial responses as assessed by the PMH group (Welsch, Wang, Gunawan, Mackay, Nathwani, Yau, Tsang, MacAlpine, Sivaloganathan, and Oza, Welsch et al.). These are the data which we shall use in our Bayesian analysis.

Code	CA-125	RECIST
19002	N	SD
19003	N	SD
19004	Y	SD
19005	N	SD
19006	N	SD
19007	N	PD
19008	N	SD
19012	N	PD
19013	N	PD
19014	N	SD
19015	N	SD
19016	N	SD
19018	N	SD

Code	CA-125	RECIST
25001	N	SD
25003	Y	SD
25004	N	PD
25005	N	PR
25006	N	SD
25009	N	SD
25010	N	SD
25012	Y	SD
25015	Y	SD
25016	Y	SD
25018	Y	SD
25020	Y	SD
25021	N	SD
25024	N	PD
25025	N	SD
25026	N	PD
25028	Y	SD
25029	Y	SD
25033	N	SD
25034	N	PD
25035	Y	SD
25036	N	PD
25037	N	SD
25038	N	SD
25039	N	SD
25040	N	SD
25042	Y	PR

Code	CA-125	RECIST
37002	Y	SD
37003	N	SD
37004	N	PD
37006	N	SD
37007	N	SD
37008	N	PD
37011	Y	SD
37013	N	SD
37014	N	SD
37015	N	PD
37019	Y	SD
37020	Y	SD
37021	Y	SD
37022	N	SD
37024	Y	PR
37025	N	SD
37027	N	SD
37030	N	SD
37033	Y	SD
37035	Y	SD
37037	N	SD
37039	Y	SD
37040	N	SD
37042	N	SD
37043	N	SD
37048	Y	SD
37050	N	SD
37053	N	SD
37059	N	SD
37065	N	SD

Code	CA-125	RECIST
41902	N	PD
41904	N	SD
41905	N	SD
41906	N	PD
41907	N	SD
41911	N	SD
41912	N	PD
41913	N	SD
41916	N	SD
41917	N	SD

Table 5.2: Four clinical trials' results obtained from the Princess Margaret Hospital. 'Y' denotes 'response' according to CA-125 serum level and 'N' denotes non-responsive. RECIST column shows the patient's classification following the tumor's longest diameter (LD).

5.4 Concordance analysis

The purpose of this section is to analyze whether or not CA-125 responses are in concordance with the objective tumor measurements of the longest diameter in four clinical trial groups. The two classifications are in concordance if CA-125 serum level indicates 'response' and the RECIST criteria based on the longest diameter indicate 'complete response (CR)' or 'partial response (PR)'. Thus, the concordance we are talking about here is not the straight-forward measure given by the covariance of the two parameters (which involves the entire data set), but rather, it is what we called "medical concordance" in the last section of chapter four, namely the one which only uses the subset of the data satisfying prescribed "thresholds".

As we did in the case of CAIX staining and the direct pO_2 measurements, we hypothesize that the plausibility that the concordance between the RECIST response and the CA-125 response is quantified by $H \in [0, 1]$. If $H = 0$, then both CA-125 and the objective tumor measure are in perfect discordance; whereas, if $H = 1$, then they are in perfect concordance. As usual, by the application of the Bayes' theorem, we have

$$P(H|\{\text{data}\}, I) \propto P(\{\text{data}\}|H, I)\ P(H|I), \tag{5.3}$$

where I denotes our background information, for instance, the two responses (CA-125 and RECIST) are independent of each other and the outcome of one patient shall not influence that of another. The direct probability of obtaining the data 'r agreements; i.e., both criteria indicate response due to the administered MTA(s) in a clinical trial group of a size N' follows a binomial distribution:

$$P(\{\text{data}\}|H, I) \propto H^r\ (1 - H)^{N-r}. \tag{5.4}$$

When we have no prior knowledge about the concordance, we may use the

uniform prior probability for H. For illustration purpose, we also include two other prior probabilities: One prior is peaked at $H = 0.5$, reflecting an approximately 50% concordance, and the other alternative prior is sharply peaked at $H = 0$ and $H = 1$, indicating that our inference about the concordance is heavily biased toward the ends (Fig. 5.1). Next we must assign the prior $P(H|I)$. We have discussed at

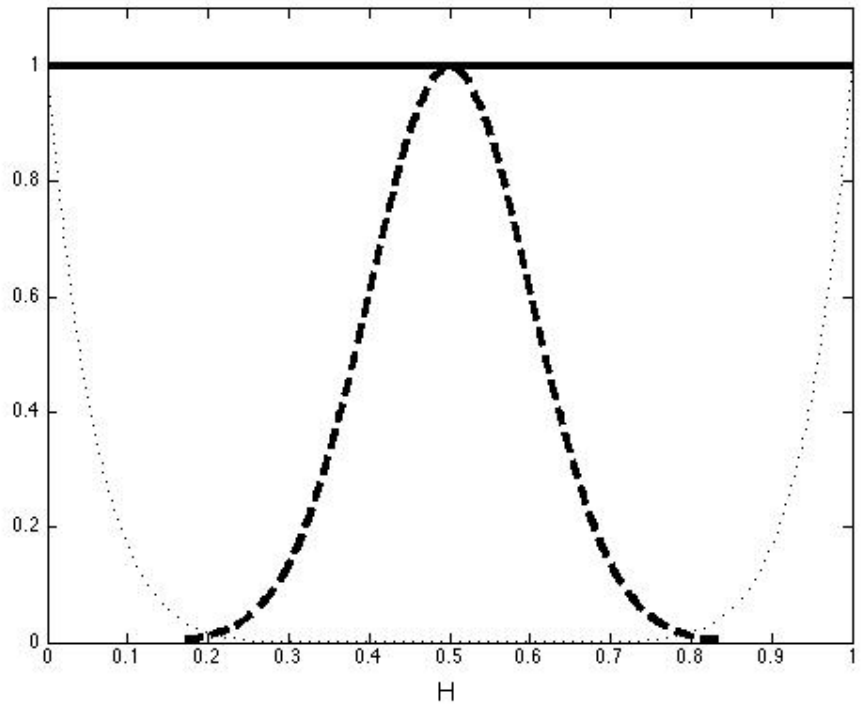

Figure 5.1: Various prior probability assignments: Uniform prior (solid), unimodal prior at $H = 0.5$ (dashed), and biased prior toward the ends (dotted)

length the importance of this assignment in the first three chapters of this Thesis. However, some readers interested in this problem might not have the time, or the desire, to do that extra reading; consequently, we will briefly review here the effect of different priors for the present problem.

Consider first the case where *we have no information* prior to seeing the data of Tab. 5.2. Then the most honest way of describing our ignorance is to assign a uniform prior, i.e, to set $P(H|I) =$ constant. Graphically, this distribution is shown in Fig. 5.1 by the horizontal solid line for the case constant = 1. Next,

we consider the case where *we do have information* about the values of H. For illustration purposes, Fig. 5.1 shows two distributions for $P(H|I)$: A Gaussian distribution peaked at $H = 0.5$ (dashed line), and a biased distribution showing great preference for the values $H = 0$ and $H = 1$ (dotted line).

We are now ready to apply the Bayesian parameter estimation algorithm with these three priors to the data of Tab. 5.2, starting with group 19 (the leftmost table of the figure). The number 19 is just the first two digits of the code in the first column, and the groups representing the other three clinical trials will be identified in the same manner.

As we did before, we will apply the sequential procedure; that is, after the assigned priors and the likelihood function, Eq. 5.4, are used in Bayes' theorem Eq. 5.3 we get the posterior pdf, and this will now be used as the prior for the second case of group 19, and so on. The evolutions of the posteriors as more data are used shown in Fig 5.2, and the final posterior pdf's are reproduced in the top left panel of Fig. 5.3. The other three panels in this figure represent the final posterior pdf's for group 25, 37, and 41.

As it is evident in Fig. 5.2, the uniform (flat) prior and the biased priors adjust quickly once a concordance and a discordance have been observed; these two priors encode a large degree of ignorance about the nature of the concordance between the two classifications. In contrast, the second prior assignment (a unimodal prior or Gaussian prior peaked at $H = 0.5$) claims to be quite well-informed about the nature of the concordance between the two classifications. As Sivia mentioned in (Sivia, 2006), this is a 'fair-minded' prior; hence, it takes much more to be convinced that the nature of the concordance is not fair compared to ignorant priors.

The biased prior yields a bimodal probability density function (Tab. 5.3); nevertheless, one of the modes is much smaller than the other (at least a factor 1,000). The bimodality of the resulting pdf can be explained in the following way: Initially, the biased prior gives a high preference to perfect concordance and perfect discordance respectively. As more concordances accumulate from the data, the biased

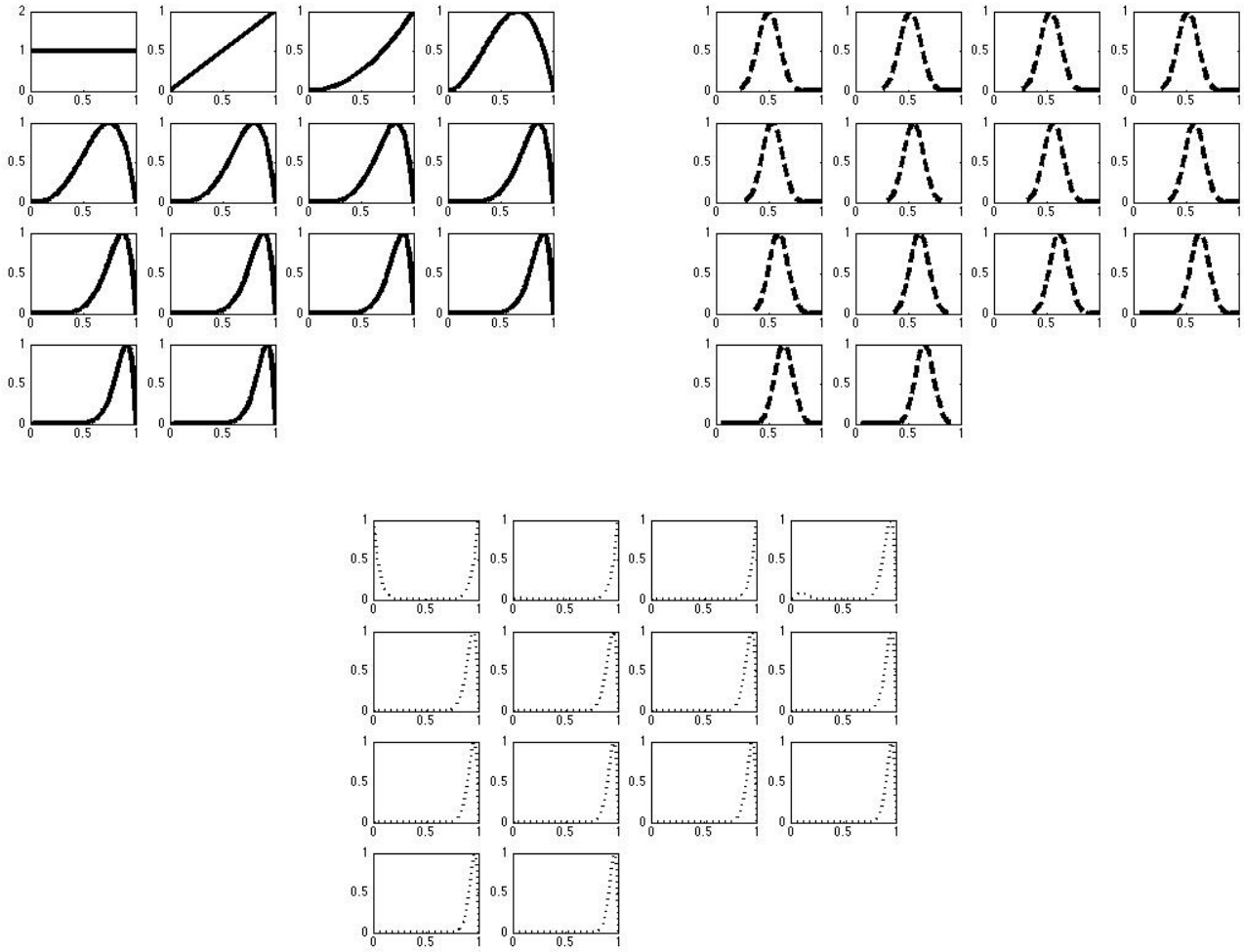

Figure 5.2: The evolution of the posterior as more data are used sequentially for various prior probability assignments: Uniform prior (solid), Gaussian prior (dashed), biased priors (dotted) to the group 19

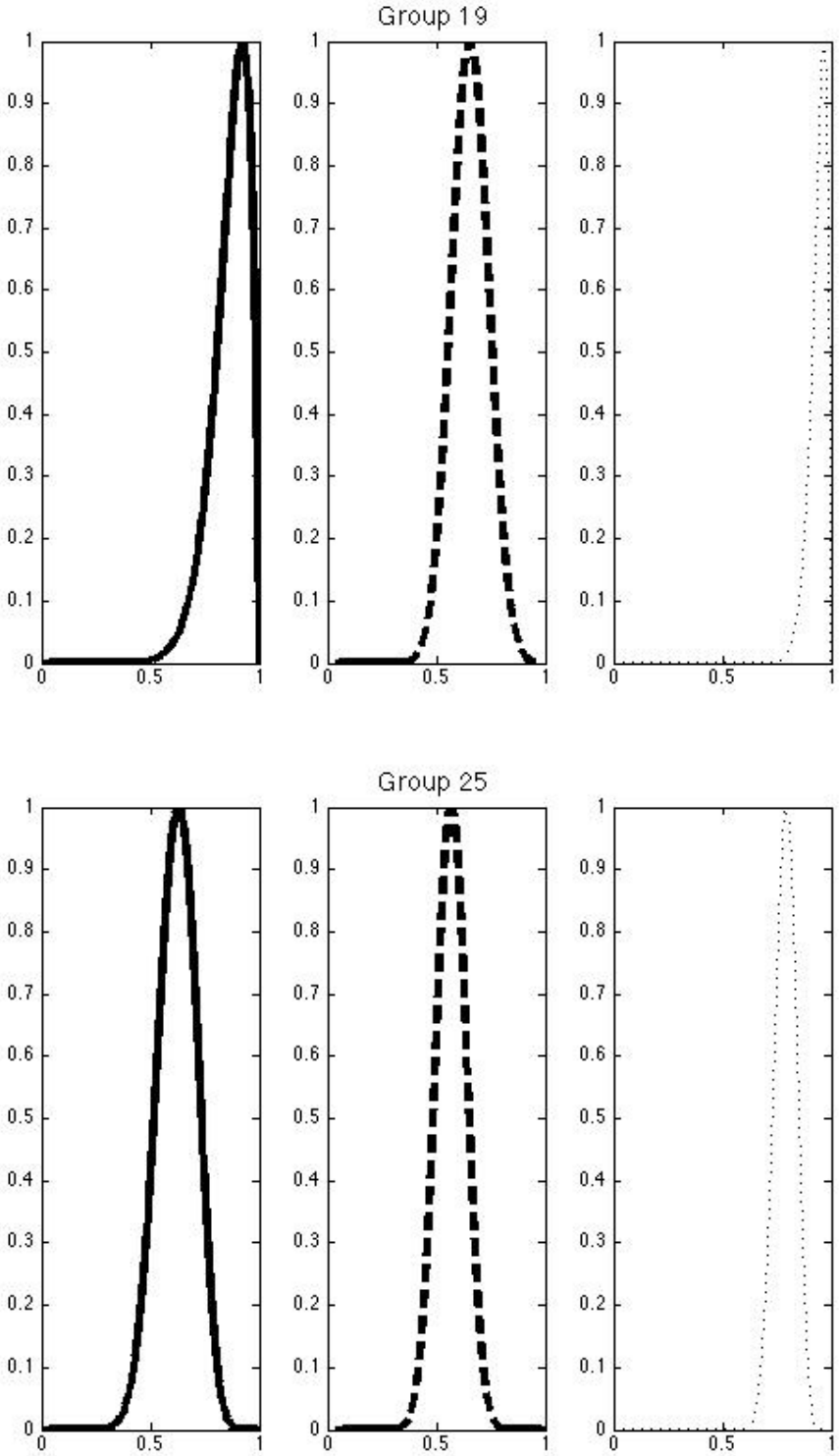

Figure 5.3: The posterior probability density function for group 19 and 25 respectively, as a result of different prior assignments.

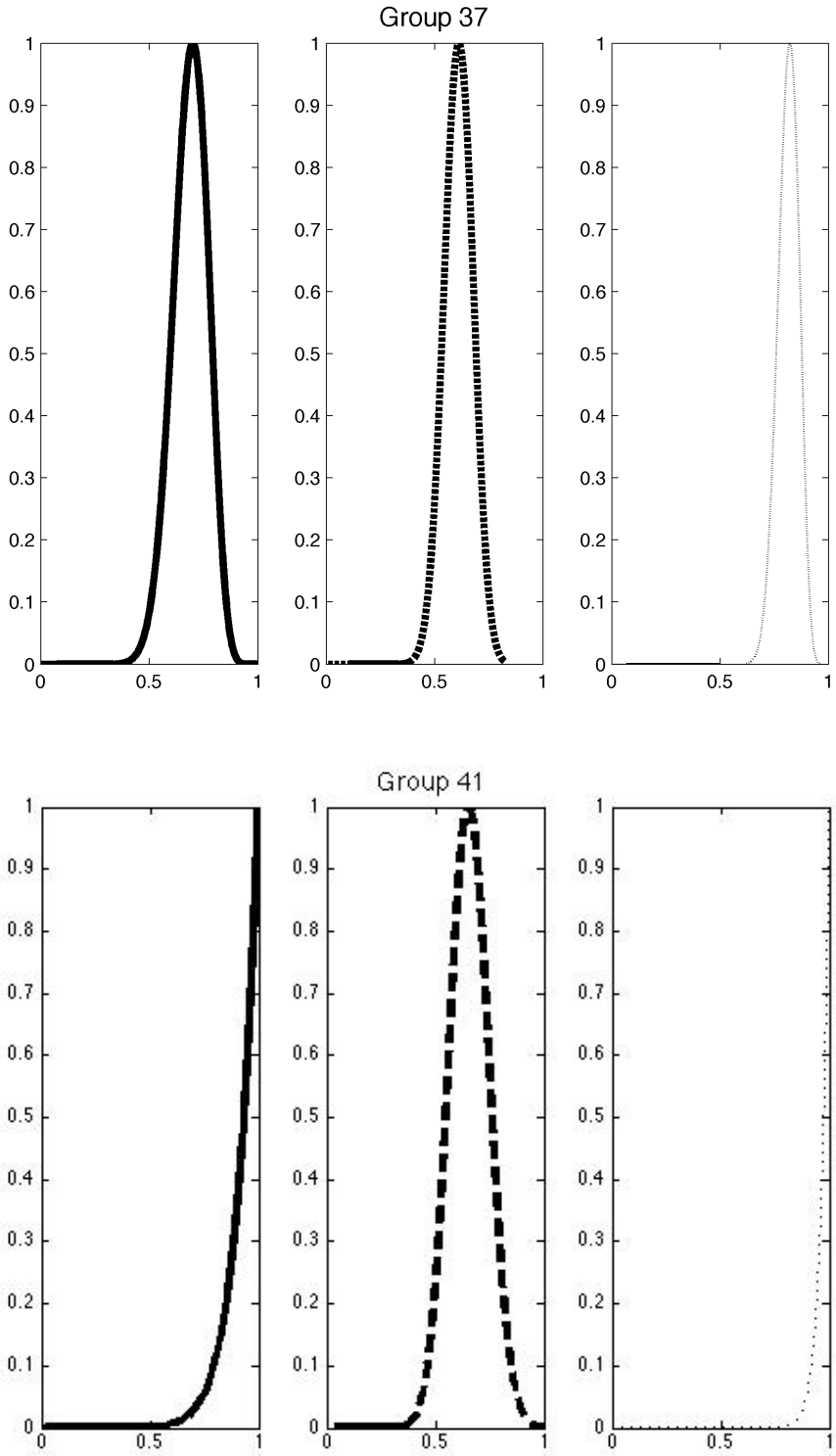

Figure 5.4: The posterior probability density function for group 37 and 41, respectively, as a result of different prior assignments.

Group	H_{uniform}	H_{gaussian}	H_{biased}
19	0.92 ± 0.07	0.65 ± 0.09	$\mathbf{0.97 \pm 0.03}$ and 0.30 ± 0.05
25	0.63 ± 0.09	0.57 ± 0.07	$\mathbf{0.79 \pm 0.05}$ and 0.31 ± 0.05
37	0.67 ± 0.09	0.59 ± 0.07	$\mathbf{0.80 \pm 0.05}$ and 0.33 ± 0.05
41	1	0.65 ± 0.09	**1** and 0.28 ± 0.06

Table 5.3: Summary for the best estimate of the concordances as a result of various priors. The bold fonts denote the ones with higher probability.

prior becomes more convinced that the concordance is more preferred, pulling away from discordance. With the uniform prior, however, we are open-minded about the nature of the concordance; as a result, our inference about the concordance between the two classifications is not as extreme as the biased prior. In group 41, the data show no discordance between the two classifications; that is why, the uniform prior and the biased prior have a maximum value of probability at $H = 1$. The fair-minded prior however only shifts a little to the right. Since we only have ten eligible patients in this group, they are not large enough to significantly convince this fair-minded prior.

The three cases above are just examples of the dependence of the posterior pdf from the assigned prior. In the actual case, our collaborators from PMH made available to us data from earlier clinical trials conducted in other institutions under similar procedures and MTAs. So we were able to repeat the previous analysis for this set of data before looking at the PMH data. Thus, we were not ignorant about the concordance parameter, and it was possible for us to use this information and use this pdf as the prior for the analysis of the PMH data.

With the informative prior probability assignment, our inference about the concordance in each group, based on the current clinical trials changes accordingly. Unlike in the previous cases when we do not have any information prior to observing the clinical trials' data, informative prior assignment means that we have some knowledge about the concordance before looking at the current data. The results of the concordance are tabulated in Tab. 5.4 and the final posterior pdfs are displayed in Fig. 5.6.

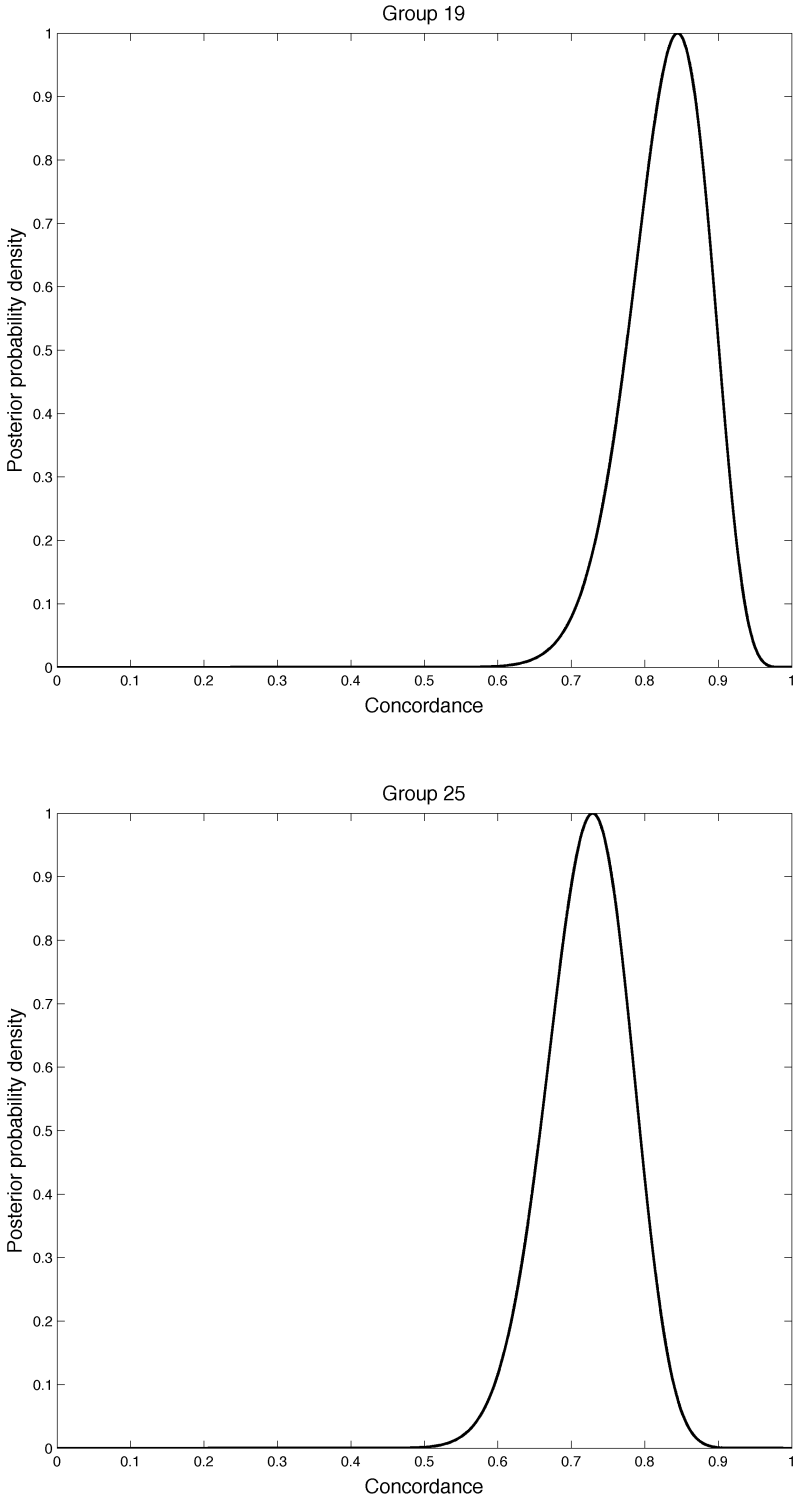

Figure 5.5: The posterior probability density function for group 19 and 25 respectively, as a result of informative prior assignment from past experiments.

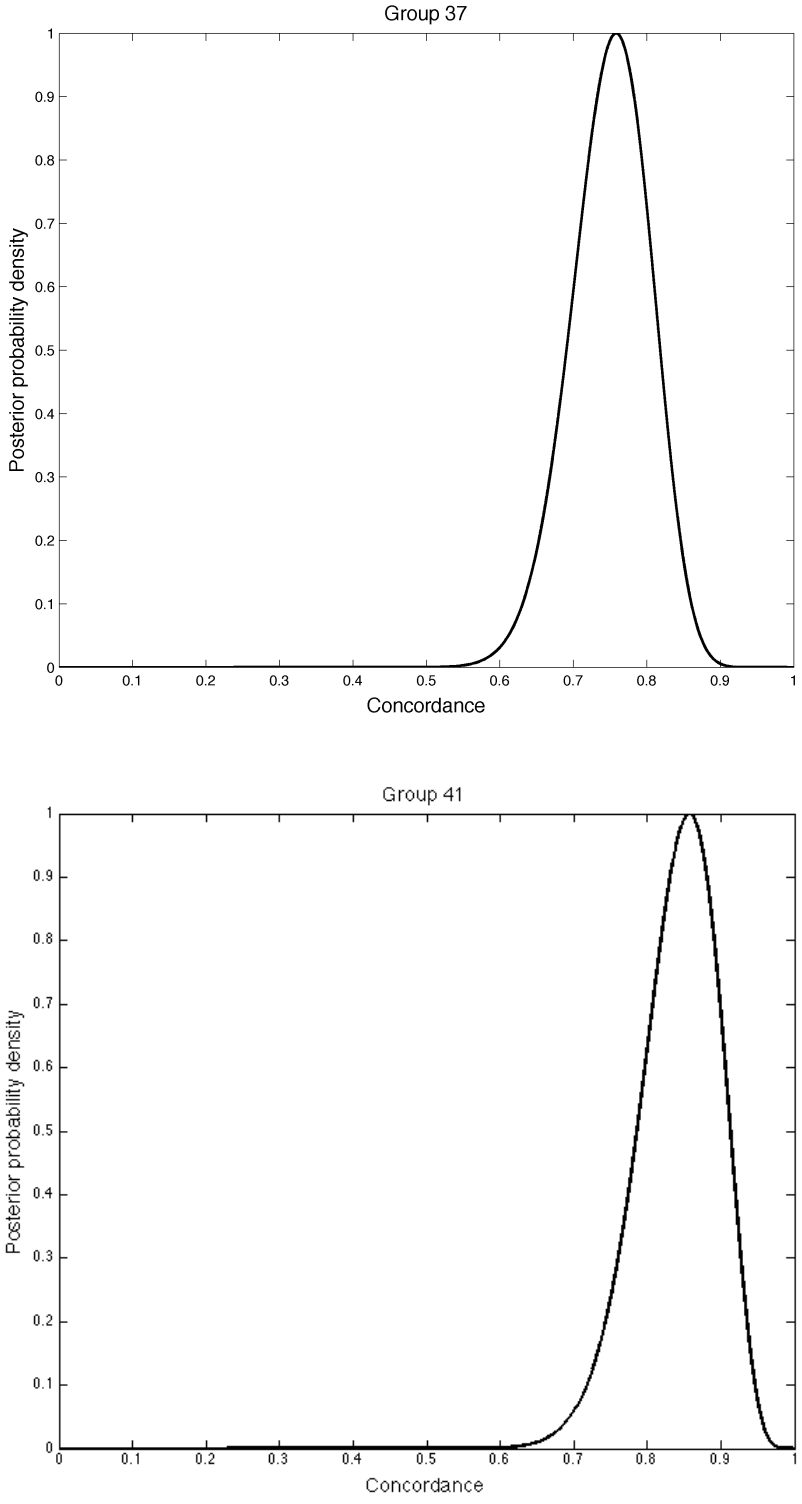

Figure 5.6: The posterior probability density function for group 37 and 41, respectively, as a result of informative prior assignment from past experiments.

Group	H
19	0.84 ± 0.05
25	0.73 ± 0.06
37	0.76 ± 0.06
41	0.86 ± 0.05

Table 5.4: The inferences about the concordance with informative prior from previous experiments.

5.5 A survival analysis for SD classified patients

According to the standard clinical trial's paradigm for a relapsed study (RECIST), SD classified patients are considered 'non-responsive'. Many oncologists and clinicians have hypothesized that the administered MTA(s) may have a cytostatic effect rather than a cytotoxic one; as a result, the effect of the MTA(s) may not be as obvious as eradicating a large number of tumor cells. In this section, we want to analyze further on the importance of patients classified as 'stable disease' (SD). In order to validate the hypothesis, we dichotomize all SD patients into two groups: CA-125 responders (19 patients) and CA-125 non-responders (49 patients). A common approach used in the statistical community is computing the survival probability in each group of patients using the Kaplan-Meier estimates.

The Kaplan-Meier statistic is a function used to define a survival probability for a period of time. Specific to the Kaplan-Meier statistic is that the intervals are defined by the observed events (in our case, when death occurs); a new interval would be demarcated each time a mortality occurs.The Kaplan-Meier statistic is a generalization of the fixed time interval method, which is also known as Berkson-Gage method (Feinstein, 2002). Tab. 5.5 and 5.6 summarize the Kaplan-Meier estimates for each of the two groups of SD patients. Fig. 5.7 is a common feature for illustrating the survival probabilities in CA-125 responders and CA-125 non-responders. Based on Fig. 5.7, it is evident that CA-125 responders (the blue line) have a survival advantage compared to CA-125 non-responders. This lends the support that the administered MTAs have a cytostatic effect rather than cytotoxic

Number of interval	Cumulative survival rate before death(s)	Time of death(s) that end(s) interval (month)	Number alive before death(s)	Number of deaths	Interval survival rate	Censored before next death
1	1	8	19	1	0.9474	4
2	0.9474	9	14	1	0.9286	0
3	0.8798	10	13	1	0.9231	0
4	0.8121	11	12	1	0.9167	0
5	0.7445	13	11	1	0.9091	2
6	0.6768	14	8	2	0.7500	0
7	0.5076	15	6	1	0.8333	0
8	0.4213	17	5	1	0.8000	0
9	0.3371	18	4	1	0.7500	0
10	0.2528	21	3	1	0.6667	1
11	0.1685	27	1	0	1	0

Table 5.5: The Kaplan-Meier survival rates for CA-125 responders

one.

Number of interval	Cumulative survival rate before death(s)	Time of death(s) that end(s) interval (month)	Number alive before death(s)	Number of deaths	Interval survival rate	Censored before next death
1	1	3	44	2	0.9545	9
2	0.9545	4	33	2	0.9394	3
3	0.8967	5	28	5	0.8214	1
4	0.7365	6	22	1	0.9545	1
5	0.7030	7	20	1	0.9500	0
6	0.6679	19	2	2	0.8950	1
7	0.5978	9	16	1	0.9375	0
8	0.5604	10	15	2	0.8667	1
9	0.4857	11	12	1	0.9167	0
10	0.4552	12	11	1	0.9091	0
11	0.4047	13	10	1	0.9000	1
12	0.3642	14	8	1	0.8750	1
13	0.3187	15	6	2	0.6667	0
14	0.2125	16	4	1	0.7500	0
15	0.1594	20	3	1	0.6667	1
16	0.1063	24	1	1	0.9583	0

Table 5.6: The Kaplan-Meier survival rates for CA-125 non-responders

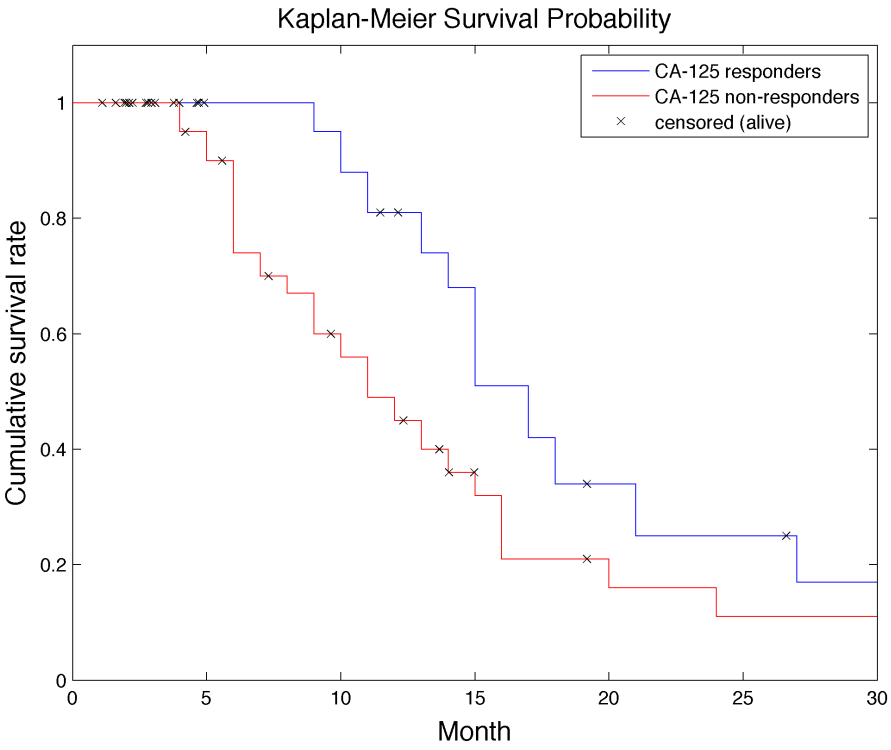

Figure 5.7: The survival rates for SD classified patients

Chapter 6

Discussion and conclusion

Data analysis is a crucial component of Science and Engineering, but the collection of reliable data is in turn importantly dependent on advanced technology. That is, perhaps, the main reason why for several millennia, Medicine has been considered more of an art than a science. In the last century, however, advances in technology have resulted in quite sophisticated instruments which have allowed medical researchers to collect reams of data in all areas of medicine. Two examples of such instruments, important in the area of oncology, are the Eppendorf polarographic electrode and the CAIX immunohistochemical assay, which are used in the assessment of hypoxia levels in tumors as explained in chapter four.

The second aspect of data analysis has to do with the methods used, which are necessarily highly mathematical. This has traditionally represented an obstacle for medical researchers who have only a passing knowledge of higher mathematics. The standard solution of the conundrum is for the doctor to pass on the data set to the statistician, who does the actual analysis and passes on the results back to the doctor. Although this "team-work" is a practical necessity, given the extremely complicated systems one deals with in biological settings, it is also fraught with danger for two reasons. On the one hand, the statistician (usually a conventional one) may not be knowledgeable enough to appreciate the medical background, and on the other, the doctor may misinterpret the statistical results and make clinical

decisions that may be harmful to the patient.

The danger of misinterpretation is pointed out with force by Steven Goodman when doctors have to deal with a P-value:

> "In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with P=0.05, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect. This is an understandable but categorically wrong interpretation because P-value is calculated *on the assumption that the null hypothesis is correct*. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false." (Goodman, 1999)

This logical error reinforces the mistaken belief that the data alone can tell us the probability that a hypothesis is true. In other words, this type of interpretation of a P-value is an attempt to make inferences from the data without using Bayes' theorem, because the latter requires the use of a prior probability as we already saw in chapter one of this Thesis. Numerous authors have tried to correct this misunderstanding (Browner and Newman, 1987; Diamond and Forrester, 1983; Lilford and Braunholtz, 1996) but the error is still widespread in the medical literature, so much that almost no papers reporting the analysis of medical data are accepted for publication without a P-value.

The misunderstanding and the confusion are further increased by the joining of the P-value with Hypothesis Testing. The latter method was introduced long ago by (Neyman and Pearson, 1933) as an answer to the inappropriateness of Fisher's P-value to the development of an inferential method without Bayes' theorem. In their hypothesis test, one poses two hypotheses about nature: a null hypothesis H_o and an alternative hypothesis H_a. *The outcome of the test is a behavior, not an inference,* namely to reject one hypothesis and accept the other exclusively on the basis of the data. This puts the researcher at risk for two types of errors; that is - in the case of comparing two therapies, for example - behaving as though the two therapies differ when they are actually the same (*false positive* result, a *type I error*

or an α *error*), or concluding that they are the same when in fact, they differ (*false negative* result, a *type II error* or a β *error*).

This approach is appealing to the mathematician because if we assume an underlying truth, then the chances of these errors can be calculated deductively, and therefore, "objectively". Nevertheless, as a model for scientific practice, it is problematic, to say the least. These authors recognized that their method represented a drastic change in the way scientific conclusions must be drawn. In their words:

> " ... no test based upon a theory of probability can itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of a test from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long-run of experience, we shall not often be wrong." (Neyman and Pearson, 1933)

As pointed out by Goodman (1999), this passage clearly states the price that must be paid for "objectivity": We must abandon our ability to measure evidence, or judge truth, in an individual experiment.

> "Hypothesis tests are equivalent to a system of justice that is not concerned with which individual defendant is found guilty or innocent (that is, whether each separate hypothesis is true or false) but tries instead to control the overall number of incorrect verdicts (that is, in the long-run of experience, we shall not often be wrong)." (Goodman, 1999)

In practice, the Neyman and Pearson method means that the researcher's duty is to report only whether or not the results are statistically significant and act in accordance with that verdict. This is often held up as a paradigm of the scientific method.

Notwithstanding the explicit words of Neyman and Pearson cited above, it appeared to many that if the P-value were used in the hypothesis testing procedure then the drawbacks of the method would be eliminated. Thus, in the medical area (and not only there), the researchers proceed as follows: An experiment is designed

to control the probabilities of two types of "error", namely type I error (α, usually equal to 0.05) and type II error (β, usually less than 0.20). When the data are obtained, a P-value is calculated and used as a quantitative measure of evidence against H_o. If P$< \alpha$, then the result is declared "significant" and H_o is regarded as unlikely to be true.

The reason why this procedure appears appealing is due to the superficial similarity of the P-value and the false positive error rate α. Both are tail-area probabilities under the null hypothesis H_o. Because of this similarity, it is easy to identify P as a special case of α, specific to the data at hand. Additionally, using Fisher's contention that P-value measures how severely H_o is contradicted by the data - that is, it could serve as a measure against H_o - we have an index (P) that does double duty: It seems to be a Neyman-Pearson data specific, false positive error rate and a Fisher measure of evidence against H_o (Royall, 1997; Goodman, 1993; Fisher, 1973).

Thus, this approach is based on the fallacy that an event can be viewed simultaneously both from a long-run and a short-run perspective.

> "In the 'long-run' perspective, which is error-based and deductive, we group the observed result together with other outcomes that might have occurred in hypothetical repetitions of the experiments. In the 'short-run' perspective, which is evidential and inductive, we try to evaluate the meaning of the observed result from a single experiment. If we could combine these perspectives, it would mean that inductive ends (drawing scientific conclusions) could be served with purely deductive methods (objective probability calculations)." (Goodman, 1999)

But these views *cannot* be reconciled because a given result (the short-run) can legitimately be included in many different long-runs. Equivalently, a result cannot be at the same time be an anonymous member of a group of results (the long-run view) and an identifiable (i.e., unique) member (the short-run view). A real-world example helps clarifying this. Suppose we examine the meaning of the statement that "a result with P=0.05 is in a group of outcomes that has 5% chance

of occurring under H_o". Although that is literally the case, we know that the result is not just *in* that group (that is, anonymous); we know where it is (that is, it is identifiable). As an analogy, we can consider the following statement,

> "... a student who ranks 10 out of 100 is *in* the top 10% of the class ... Although literally true, [this statement is] deceptive because [it suggests] that a student could be anywhere in a top fraction when we know he or she is at the lowest level of that group." (Goodman, 1999)

In contrast to this state of affairs, the Bayesian approach adopted in this Thesis is completely straight-forward. The philosophical differences have been pointed out in chapter one, and the contributions of the 'founding fathers' have been reviewed in chapter two. We recognize from the start that scientific reasoning is inductive, which means that we must make appropriate assessments of the *uncertainty* necessarily produced by the measuring apparatus. It follows that measurements and inferences have to be described as probability distributions.

Consider a measuring apparatus which is designed to let the experimenter investigate some unknown features $X = \{X_1, X_2, ...\}$ of observed samples. Regrettably, equipment seldom measures X directly or in full. Instead, it produces data $D = \{d_1, d_2, ...\}$ which may depend on X in a complicated way. Also, repeated observations even on the same sample usually produce different data because of noise. Supposing there exists a functional form (or a model of it) $D = R(x)$, which we can call the "response function" of the equipment, these data will vary according to some probability distribution

$$P(D|X, I) = \text{likelihood function},$$

where, as usual, the vertical bar denotes conditioning and I stands for "background information". Usually, we suppose that noise (denoted by σ) is additive so that we can write

$$D = R(x) \pm \sigma,$$

and that the additive noise has a Gaussian pdf, in which case we have the explicit formula (for the likelihood function)

$$P(D|X, I) = \frac{1}{Z} \exp\left\{-\frac{\chi^2}{2}\right\},$$

where χ^2 is the usual misfit between the mock data and the actual data, namely

$$\chi^2(X) = \sum \frac{[R(X_k) - d_k]^2}{\sigma_k^2}.$$

Once we have the likelihood function, we know what the data mean, and so we can realistically aim to infer the values of the unknown quantities X. Such inference will be uncertain; hence, we can at best aim for the posterior pdf

$$P(X|D, I) = \text{inference},$$

This involves inverting the assigned likelihood, but the inversion requires us to assign a prior pdf

$$P(X|I) = \text{prior},$$

which represents our guess, or preconception, about the range of the unknown values that might be present *before* we see the data. As seen in chapter two and three, it is the Bayes' theorem which effects the inversion.

We have applied in chapter four this methodology to the analysis of CAIX staining and pO_2 measurements on the data collected by the PMH group, and analyzed by them with the methods of conventional (frequentist) statistics. Throughout the analysis, the simplifying assumption of tumor homogeneity was made, and as a result we were able to estimate the minimum number of biopsies capable of giving an estimate of the amount of hypoxia in the entire tumor - something not easily done by conventional methods.

But even more interesting are the results reported in Section 4.4 concerning the question of whether or not a positive correlation exists between the CAIX assess-

ment of hypoxia and the Eppendorf probe assessment. Finding such a correlation is an important goal for oncological research, in that it would replace the Eppendorf polarographic electrode which is far more cumbersome and subject to large fluctuations in the measurement of oxygen tension. Calculations of this potential correlation is published so far were done using only conventional techniques, and little, or at most moderate, correlation was found. We have estimated this correlation in Section 4.4 *using the entire data set*, and *found zero correlation*. We then repeated the calculation with a manipulated set of data, which was the same set used by our collaborators at PMH. They used conventional statistics and reported their results in the following form: "Pearson = 0.30, Spearman = 0.35, P = 0.091" (Iakovlev et al., 2007). With our Bayesian approach we found a much higher value, namely 0.67.

In our opinion, this result should be taken with great caution because of the manipulation of the data. One of the merits of Bayesian methods is that it gives the most honest inference from a given data set, provided the full data set and only monotonic transformations are used. This is not always the case in the frequentist approach; for example, in (Iakovlev et al., 2007), they used the transformation

$$\text{caix} = \arcsin\sqrt{\text{CAIX}},$$

where CAIX is the original proportion and caix is the transformed value in order to "stabilize the variance of the residuals".

Finally, in chapter five, we report on the Bayesian analysis of an unrelated set of data, namely data from phase II clinical trials of relapsed ovarian cancers. Here as well the underlying assumptions are such that the Bayesian calculations are completely straight-forward, and the results are currently being incorporated in a paper jointly with our medical collaborators (Welsch, Wang, Gunawan, Mackay, Nathwani, Yau, Tsang, MacAlpine, Sivaloganathan, and Oza, Welsch et al.).

Appendix A

A brief overview of Monte Carlo integration

Monte Carlo simulation is a vast area of research; here, we provide a brief overview for crude Monte Carlo integration technique. For simplicity, let us consider a one-dimensional definite integral,

$$\theta = \int_D f(x)\, dx. \tag{A.1}$$

Suppose we can find a random variable y with support in D such that

$$f(y) = g(y) \underbrace{p(y)}_{\text{pdf of } y}. \tag{A.2}$$

Since

$$\int_D g(y)\, p(y)\, dy =< g >, \tag{A.3}$$

where $< g >$ denotes the expectation of g, then assuming the random variable is *uniformly* distributed (i.e. $p(y) = \text{const} = 1$ without lost of generality), allows us to write

$$\int_D f(x)\, \mathrm{d}x = \int_D g(y)\, p(y)\, dy = \theta. \tag{A.4}$$

k	$f(\xi_k)$
1	0.4651
2	0.3616
3	0.4169
4	0.4684
5	0.2925

For illustration purposes, let us consider the following example,

$$\theta = \int_0^1 \frac{e^x - 1}{e - 1} dx,$$

where $\theta = 0.418$ (Hammersley and Handscomb, 1964). For five iterations, the crude Monte Carlo results are : Hence the expectation of $f(\xi)$ is

$$\hat{\theta} = \frac{1}{5} \sum_{k=1}^{5} f(\xi_k), \tag{A.5}$$

therefore, $\theta \approx \hat{\theta}$. This is an example of crude Monte Carlo method; there are various other Monte Carlo methods for definite integration, namely, *hit-and-miss Monte Carlo*, or Monte Carlo with importance sampling, leading to the development of more sophisticated techniques such as Markov Chain Monte Carlo (MCMC). We should keep in mind that unlike deterministic numerical integrations, there is no rule of convergence in Monte Carlo methods and they may have long 'burn-in' time for a high dimensional integration; consequently, we must be cautious in interpreting the results. There are numerous review resources for Monte Carlo methods; however, (Hammersley and Handscomb, 1964) is a good resource for the fundamentals of Monte Carlo methods and (Gentle, 2003) is detailed resource about various types of random number generators.

Appendix B

Student-t distribution

When we use a Gaussian distribution with an unknown standard deviation s for our prior probability of the noise, we obtain the following form for our posterior pdf (after integration and uniform prior assignment)

$$P(Y|D,I) \propto \frac{1}{\left(\sum_{i=1}^{N}(d_i - Y)^2\right)^{\frac{N}{2}}}, \tag{B.1}$$

where Y is our parameter of interest and N is the total data points. It is the well-known Student-t distribution after some simple simplifications in the denominator,

$$\begin{aligned}
\sum_{i=1}^{N}(d_i - Y)^2 &= \sum_{i=1}^{N}(d_i^2 - 2d_iY + Y^2), \\
&= \sum_{i=1}^{N} d_i^2 - 2Y\sum_{i=1}^{N} d_i + NY^2, \\
&= N(Y^2 - 2Y\bar{d}) + \sum_{i=1}^{N} d_i^2,
\end{aligned}$$

where $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$, the sample mean. We proceed by completing the square in the first factor on the right-hand side, and adding and subtracting $N\bar{d}^2$,

$$\begin{aligned}
\sum_{i=1}^{N}(d_i - Y)^2 &= N(Y-\bar{d})^2 - N\bar{d}^2 + \sum_{i=1}^{N} d_i^2, \\
&= N(Y-\bar{d})^2 + N\bar{d}^2 - 2N\bar{d}^2 + \sum_{i=1}^{N} d_i^2, \\
&= N(Y-\bar{d})^2 + N\bar{d}^2 - 2N\bar{d}\frac{\sum_{i=1}^{N} d_i}{N} + \sum_{i=1}^{N} d_i^2, \\
&= N(Y-\bar{d})^2 + N\bar{d}^2 - 2\bar{d}\sum_{i=1}^{N} d_i + \sum_{i=1}^{N} d_i^2.
\end{aligned}$$

We complete the square again on the last two factors on the right-hand side,

$$\begin{aligned}
\sum_{i=1}^{N}(d_i - Y)^2 &= N(Y-\bar{d})^2 + N\bar{d}^2 + \sum_{i=1}^{N}(d_i-\bar{d})^2 - N\bar{d}^2, \\
&= N(Y-\bar{d})^2 + \sum_{i=1}^{N}(d_i-\bar{d})^2,
\end{aligned}$$

where often we write $V = \sum_{i=1}^{N}(d_i - \bar{d})^2$. Finally, the posterior pdf has the following form

$$P(Y|D,I) \propto \frac{1}{\left(N(Y-\bar{d})^2 + V\right)^{\frac{N}{2}}}, \tag{B.2}$$

which is known as a Student-t distribution with $N-1$ degrees of freedom. When $N = 2$, this probability density function is also known as the *Cauchy* distribution. It has a maximum is at $Y = \bar{d}$, a full width at half maximum (FWHM) proportional to $\sqrt{V}$, and a very long tail.

Appendix C

Bayesian non-parametric analysis

In chapter four and five of this Thesis, we saw some applications of Bayesian data analysis in cancer research. The analyses presented in those chapters fall under the category parameter estimation because we estimate the parameter based on an assumed functional model (CAIX-positive cells are *constant* within a tumor, in chapter 4). In cases where we do not know the exact functional model but we may have some reasons to restrict the functional model into a class of functions, we may use a model selection procedure, i.e, when we want to determine the number of exponential models present in the data as in (Sivia, 2006; Bretthorst et al., 2005).

A much harder case is where we cannot characterize the functional model at all; we do not know about the physics of the problem and have no confidence to assume the distribution has a simple form (Sivia, 2006). This type of problems falls under the class of "non-parametric" or "free-form" estimation. In many experiments, we may need to compare samples of data from a probability whose density function (pdf) is unknown. For example, estimating the CAIX-positive cells' distribution in a tumor from a sample. As we mentioned earlier, we do not yet know about the physics of CAIX protein to be able to predict the theoretical form of the distribution and we have no reason to assume the distribution has some simple form, such as Gaussian.

Although this type of problem is also encountered in the classical method (Ar-

mitage et al., 2002; Wadsworth, 1997), we will proceed with the Bayesian approach as first proposed by Gull and Fielden (1984). In the Bayesian framework, we make only minimal assumptions about the nature of the underlying pdf (Gull and Fielden, 1984). For illustration purposes, we follow the procedures for estimating the position of a quantile (i.e., the median of a pdf *not* the sample) is outlined as follows (Gull and Fielden, 1984).

Suppose that we have N samples, $\{x_i\}$, taken from a certain pdf $P(x)$, where $x \in [a, b]$. Let us then denote by M_x the proposition that the median lies at x:

$$M_x : \qquad \int_a^x P(s)ds = \frac{1}{2}, \tag{C.1}$$

because median position at x means that the area under the probability density function to the left of x and to the right of x is equal. Then, using the Bayes' theorem, we obtain the posterior probability density function (pdf) for M_x,

$$P(M_x|\{x_i\}, I) \propto P(M_x|I) \times P(\{x_i\}|M_x, I). \tag{C.2}$$

The proposition M_x (for different x) are *exclusive* and *exhaustive*, since the median certainly lies in $[a, b]$. We therefore take the prior $P(M_x|I)$ to be uniform for all x in this interval. Unlike in the earlier cases, the likelihood $P(\{x_i\}|M_x, I)$ is not uniquely determined by the position of the median.The proposition M_x however constitutes *testable* information about $P(x)$; i.e, given any $P(x)$, we decide immediately whether it is consistent with information M_x. Furthermore, if the samples are *exchangeable*, then M_x is testable information about the joint pdf $P(\{x_i\})$; thus, we shall seek a pdf that incorporates information available, yet is maximally non-committal about other parameters of the distribution. We shall use Maximum Entropy (MaxEnt) principle under the constraint M_x; we maximize

$$S = -\int d^N x \;\; P(\{x_i\}) \log\left[\frac{P(\{x_i\})}{m(\{x_i\})}\right], \tag{C.3}$$

subject to normalization condition,

$$\int d^N x \;\; P(\{x_i\}) = 1, \tag{C.4}$$

and the position of the median,

$$\int_a^x d^N s \;\; P(\{s_i\}) = \frac{1}{2}, \tag{C.5}$$

because the integral is N-dimensional, rather than writing many integral notations, we use the notation commonly used in the physics literature. For one sample, x', and a uniform prior in $[a, b]$, the solution has a simple form:

$$P(x'|M_x, I) = \begin{cases} \frac{1}{2(x-a)}, & \text{for } x' < x \\ \frac{1}{2(b-x)}, & \text{for } x' > x. \end{cases} \tag{C.6}$$

As an example consider $x' = 0.75$, $a = 0$, and $b = 1$, the likelihood is

$$P(0.75|M_x, I) = \begin{cases} \frac{1}{2(x-0)}, & \text{for } 0.75 < x \\ \frac{1}{2(1-x)}, & \text{for } 0.75 > x. \end{cases} \tag{C.7}$$

Using Bayes' theorem, $P(M_x|0.75, I) \propto P(0.75|M_x, I)$; hence, for the posterior pdf, we can view Eq. C.7 as a function of x. Fig C.1 displays the result for one-sample posterior pdf for the median position M_x.

For multiple, exchangeable samples, the MaxEnt distribution is necessarily independent because the constraints all take the form of separate equalities on each of the marginal distributions. We find the likelihood to have the following form:

$$P(\{x_i\}|M_x, I) = \left[\frac{1}{2(x-a)}\right]^{N<} \quad \left[\frac{1}{2(b-x)}\right]^{N>}, \tag{C.8}$$

where $N <$ and $N >$ are the number of data points $x_i < x$ amd $x_i > x$ respectively. The posterior pdf is proportional to C.8 but viewed as a function of x. In this case, we wish to estimate, from a certain sample of size N, the position of the

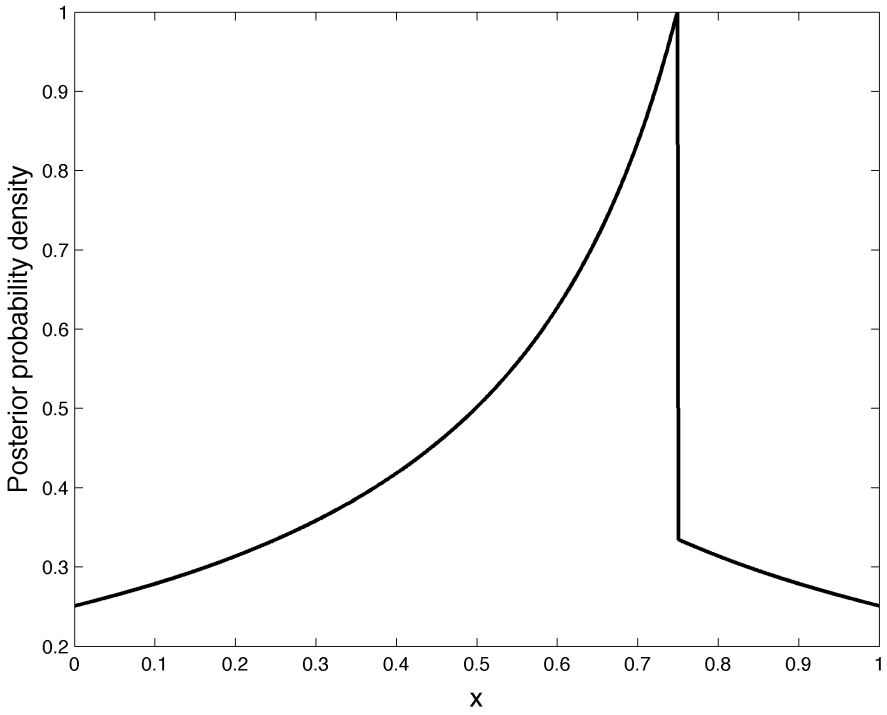

Figure C.1: The posterior pdfs for the position of median for one-sample set. The pdf is normalized vertically so that the maximum height is unity.

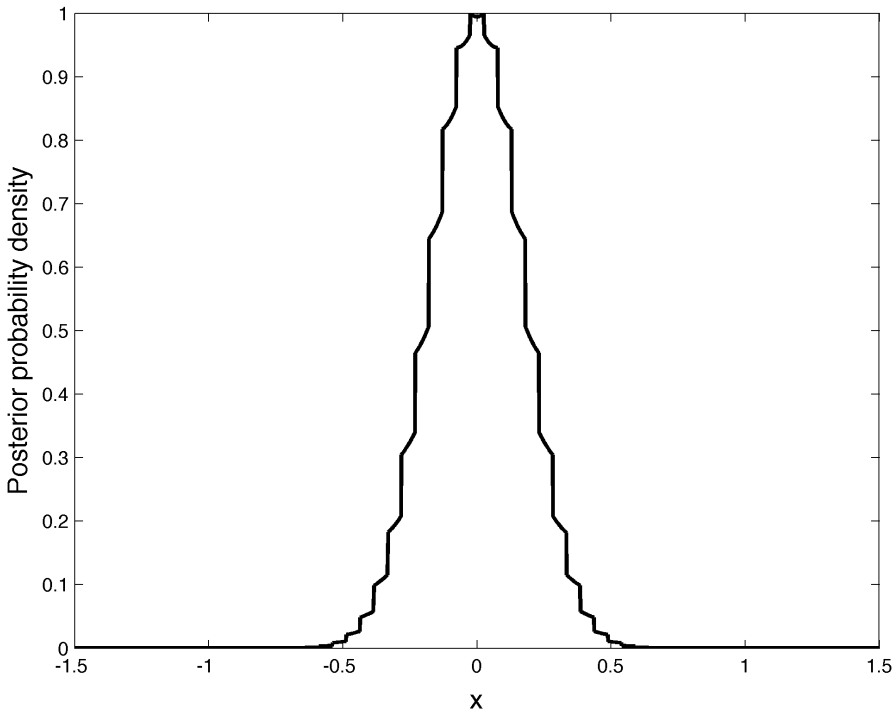

Figure C.2: The posterior pdfs for the position of median for 39 samples uniformly distributed between -1 and 1. The pdf is normalized vertically so that the maximum height is unity.

median making only minimal assumptions about the form of the distribution. Note that we are *not* claiming that the MaxEnt distribution for some x is in fact the underlying pdf. To illustrate this in practice, let us follow the example provided in (Gull and Fielden, 1984): determining the position of the median of the pdf based on 39 samples uniformly spaced between -1 and 1, with a prior range of $[-1.5, 1.5]$. Fig. C.2 illustrates the posterior pdf of M_x, the position of the median.

An interesting feature about this problem is that the effect of the prior range is still apparent. Fig. C.3 shows the posterior pdf for the same sample with a prior range $[-1.5, 1.5]$, $[-3, 3]$, and $[-5, 5]$ respectively. As it is evident, the distribution gets wider as the prior range is increased, but not proportionally. This is reasonable when we consider how little has been assumed about the distribution (Gull and Fielden, 1984). In this example, we specifically designed the sample to be uniformly spaced between $[-1, 1]$; thus, the resulting pdf is symmetric and unimodal, aside

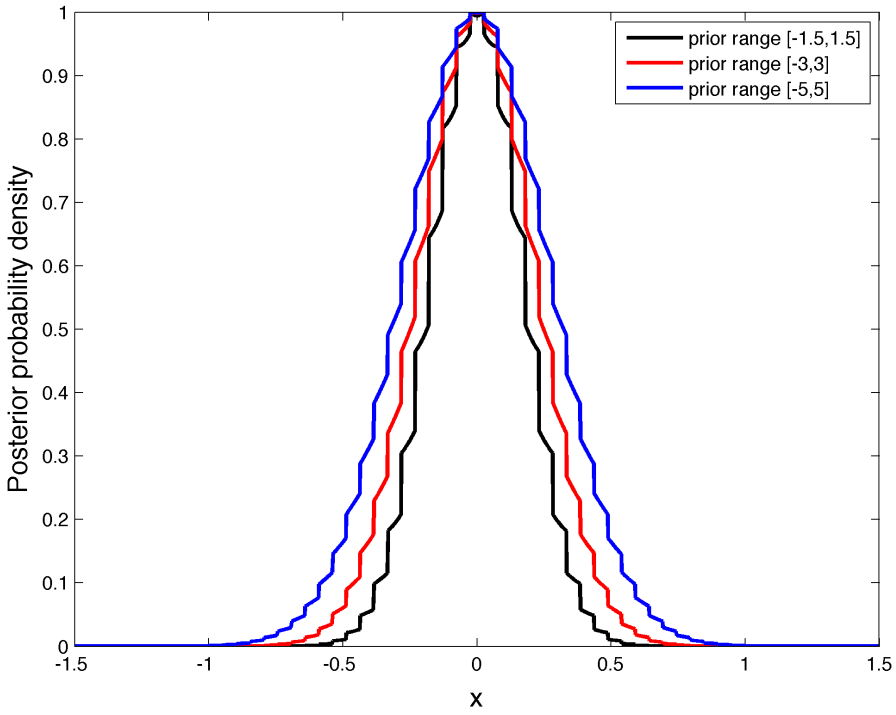

Figure C.3: The posterior pdfs for the position of median for 39 samples uniformly distributed between -1 and 1 with different prior ranges.

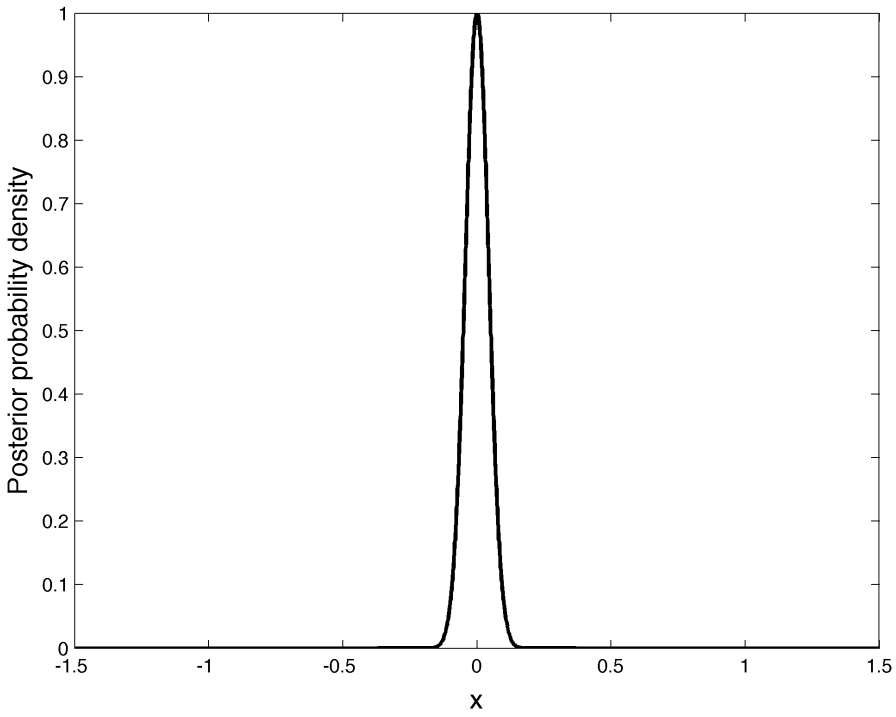

Figure C.4: The posterior pdfs for the position of median for 600 samples uniformly distributed between -1 and 1.

from the discontinuities at the sample points. The best estimate of the position of the median is the value which yields the maximum and the uncertainty of the estimate is simply the full width at half maximum of pdf.

As we notice from the figures, the posterior pdf has an unusual characteristic, namely discontinuities at the sample points. If we have many more sample points (i.e, 600 points uniformly spaced within $[-1, 1]$), then the discontinuities will be unnoticeable (Fig. C.4). Although many have attempted in considerable length to explain this phenomenon, it is not necessary; as mentioned in (Gull and Fielden, 1984), "it is the job of a pdf to tell us how much probability falls into any interval x, hence it must certainly be integrable". There is no requirement that pdf has to be a continuous one.

In real experimental data however, there is no guarantee for the symmetric and unimodal pdf. As an example, let us consider the CAIX data from the first and

second set, respectively (Fig. C.5). The best estimate of the median in patient 2144's data can be inferred quite reliably (Fig. C.5, left). The same conclusion cannot be said for patient 2149's data due to their multimodality.

As a final note, in the CA-125 project, the PMH group is interested in determining the median survival times for two different classes of 'SD' classified patients: CA-125 responders and CA-125 non-responders in order to investigate the importance of CA-125 serum marker. As it is mentioned in (Welsch, Wang, Gunawan, Mackay, Nathwani, Yau, Tsang, MacAlpine, Sivaloganathan, and Oza, Welsch et al.), the molecularly-targeted agents (MTAs) may have cytostatic effects rather than cytotoxic ones; consequently, the administered MTAs may be effective in SD classified patients and CA-125 serum marker is the diagnostic tool to show the effectiveness of the MTAs in these clinical trials. Fig. C.6 shows the posterior pdf for the median survival times for CA-125 responders (black line) and CA-125 non-responders (red line). Even though we see that the posterior pdf of CA-125 responders is located to the right of the posterior pdf of CA-125 non-responders, the width of the pdf of CA-125 responders and the multimodality of the pdf of CA-125 non-responders warn us that we should not jump to conclusions based on these data.

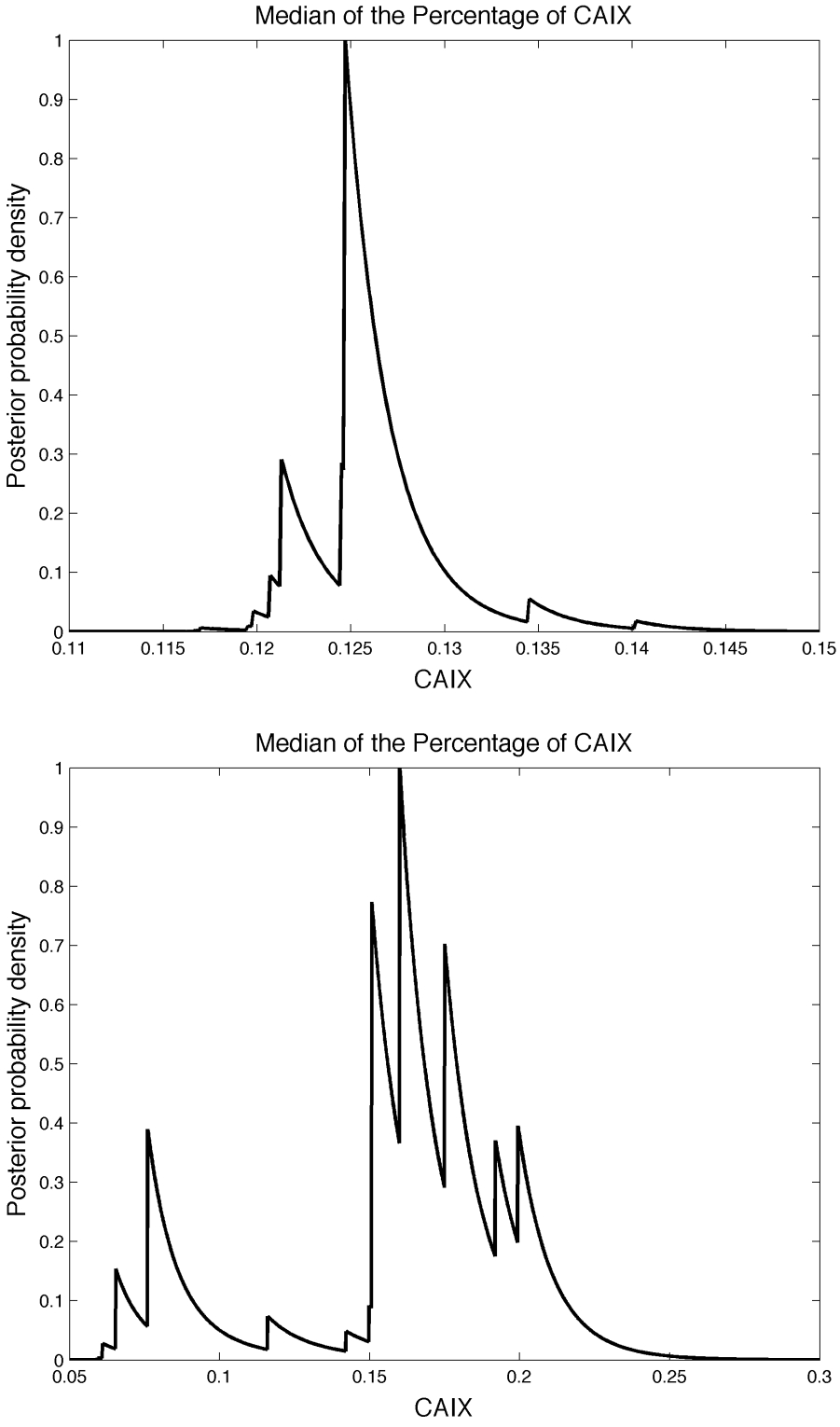

Figure C.5: The posterior pdfs for the position of median. *Left*: patient 2144 (from the first data set). *Right*: patient 2149 (from the second data set)

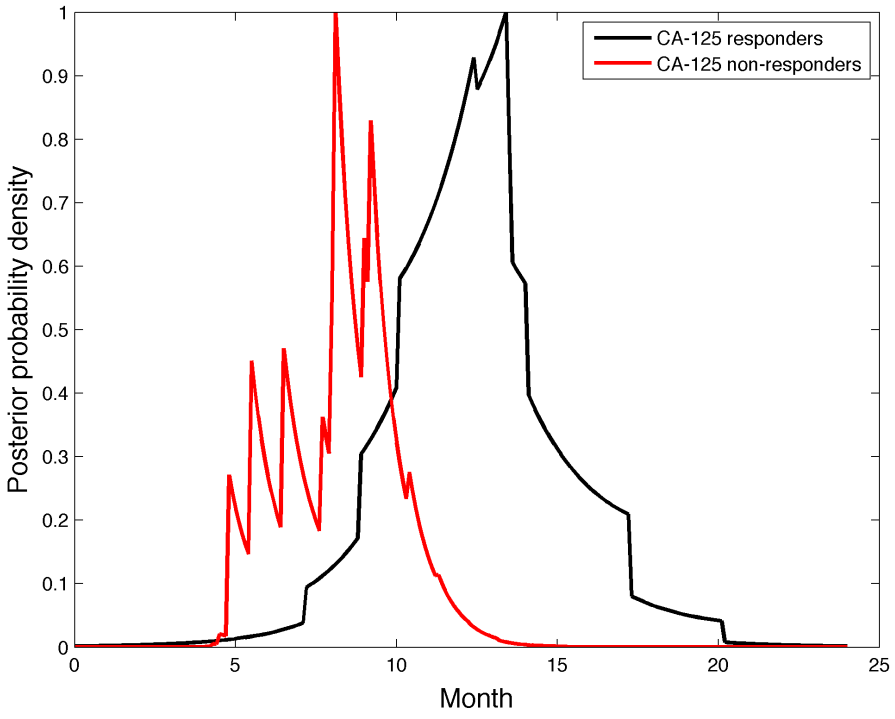

Figure C.6: The posterior pdfs for median survival times in SD classified patients.

References

Armitage, P., G. Berry, and J. N. S. Matthews (2002). *Statistical Methods in Medical Research*. Blackwell Science.

Badgwell, D. and R. C. Bast (2007). Early detection of ovarian cancer. *Disease Markers 23*, 397–410.

Bast, R. C., D. Badgwell, Z. Lu, R. Marquez, D. Rosen, J. Liu, K. A. Baggerly, E. N. Atkinson, S. Skates, Z. Zhang, A. Lokshin, U. Menon, I. Jacobs, and K. Lu (2005). New tumor markers: CA125 and beyond. *International Journal of Gynecological Cancer 15 Suppl 3*, 274–281.

Berry, D. A. (2005). Introduction to bayesian methods iii: use and interpretation of bayesian tools in design and analysis. *Clin. Trials 2*, 295–300.

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery 5*, 27–36.

Bretthorst, G. L. (1988). *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag.

Bretthorst, G. L., W. C. Hutton, J. R. Garbow, and J. J. H. Ackerman (2005). Exponential model selection (in nmr) using bayesian probability theory'. *Concepts in Magnetic Resonance 27A*, 64–72.

Browner, W. S. and T. B. Newman (1987). Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *Journal of the American Medical Association 257*, 2459–2463.

Byar, D. P., A. M. Herzberg, and W. Y. Tan (1993). Incomplete factorial designs for randomized clinical trials. *Stat Med 12*, 1629–1641.

CCS (2009). Canadian. This is an electronic document. Date retrieved: April 28, 2009.

Cline, J. M., G. L. Rosner, J. A. Raleigh, and D. E. Thrall (1997). Quantification of CCI-103F labeling heterogeneity in canine solid tumors. *International Journal of Radiation Oncology, Biology, and Physics 37*, 655–662.

Cline, J. M., D. E. Thrall, R. L. Page, A. J. Franko, and J. A. Raleigh (1990). Immunohistochemical detection of a hypoxia marker in spontaneous canine tumours. *British Journal of Cancer 62*, 925–931.

Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics 14*, 1–13.

Cox, R. T. (1961). *The Algebra of Probable Inference*. The Johns Hopkins Press.

Dewhirst, M. W. (1998). Concepts of oxygen transport at the microcirculatory level. *Seminar in Radiation Oncology 8*, 143–150.

Diamond, G. A. and J. S. Forrester (1983). Metadiagnosis. An epistemologic model of clinical judgment. *American Journal of Medicine 75*, 129–137.

Durand, R. E. and C. Aquino-Parsons (2006). The fate of hypoxic (pimonidazole-labelled) cells in human cervix tumours undergoing chemo-radiotherapy. *Radiotherapy and Oncology 80*, 138–142.

Evans, S. M., D. Fraker, S. M. Hahn, K. Gleason, W. T. Jenkins, K. Jenkins, W. T. Hwang, P. Zhang, R. Mick, and C. J. Koch (2006). EF5 binding and clinical outcome in human soft tissue sarcomas. *International Journal of Radiation Oncology, Biology, and Physics 64*, 922–927.

Evans, S. M., S. M. Hahn, D. P. Magarelli, and C. J. Koch (2001). Hypoxic heterogeneity in human tumors: EF5 binding, vasculature, necrosis, and proliferation. *American Journal of Clinical Oncology 24*, 467–472.

Everitt, B. S. and A. Pickles (2004). *Statistical Aspects of the Design and Analysis of Clinical Trials*. Imperial College Press.

Fatt, I. (1976). *Polarographic Oxygen Sensor : Its Theory of Operation and Its Application in Biology, Medicine, and Technology*. CRC Press.

Feinstein, A. R. (2002). *Principles of Medical Statistics*. Chapman & Hall/CRC.

Fisher, R. (1973). *Statistical Methods and Scientific Inference*. Macmillan.

Fyles, A., M. Milosevic, D. Hedley, M. Pintilie, W. Levin, L. Manchul, and R. P. Hill (2002). Tumor hypoxia has independent predictor impact only in patients with node-negative cervix cancer. *Journal of Clinical Oncology 20*, 680–687.

GCIG (2009). Gynecologic cancer intergroup. This is an electronic document. Date retrieved: May 1, 2009.

Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods*. Springer-Verlag.

Goodman, S. N. (1993). p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology 137*, 485–496.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine 130*, 995–1004.

Goodman, S. N. (2005). Introduction to bayesian methods i: measuring the strength of evidence. *Clinical Trials 2*, 282–290.

Goodman, S. N. and J. T. Sladky (2005). A Bayesian approach to randomized controlled trials in children utilizing information from adults: the case of Guillain-Barre syndrome. *Clinical Trials 2*(4), 305–310.

Gregory, P. (2005). *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press.

Gull, S. F. (1988). Bayesian data analysis - straight line fitting. In *Maximum Entropy and Bayesian Methods*.

Gull, S. F. and J. Fielden (1984). Bayesian non-parametric analysis. In *Maximum Entropy and Bayesian Methods in Applied Statistics*.

Hall, E. (1994). *Radiobiology for The Radiologist*. J.B. Lippincott.

Hamilton, W. C. (1964). *Statistics in Physical Science: Estimation, Hypothesis Testing and Least Squares*. The Ronald Press Company.

Hammersley, J. and D. Handscomb (1964). *Monte Carlo Methods*. Methuen.

Harrison, L. B., M. Chadha, R. J. Hill, K. Hu, and D. Shasha (2002). Impact of tumor hypoxia and anemia on radiation therapy outcomes. *Oncologist 7*, 492–508.

Hayes, R. B., D. Reding, W. Kopp, A. F. Subar, N. Bhat, N. Rothman, N. Caporaso, R. G. Ziegler, C. C. Johnson, J. L. Weissfeld, R. N. Hoover, P. Hartge, C. Palace, and J. K. Gohagan (2000). Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control Clinical Trials 21*, 349S–355S.

Health, G. (2000). Types of clinical trials. This is an electronic document. Date of publication: October, 2000. Date retrieved: April 28, 2009.

Hedley, D., M. Pintilie, J. Woo, A. Morrison, D. Birle, A. Fyles, M. Milosevic, and R. Hill (2003). Carbonic anhydrase IX expression, hypoxia, and prognosis in patients with uterine cervical carcinomas. *Clinical Cancer Research 9*, 5666–5674.

Hedley, D., M. Pintilie, J. Woo, T. Nicklee, A. Morrison, D. Birle, A. Fyles, M. Milosevic, and R. Hill (2004). Up-regulation of the redox mediators thioredoxin and apurinic/apyrimidinic excision (APE)/Ref-1 in hypoxic microregions of invasive

cervical carcinomas, mapped using multispectral, wide-field fluorescence image analysis. *American Journal of Pathology 164*, 557–565.

Hoogsteen, I. J., H. A. Marres, K. I. Wijffels, P. F. Rijken, J. P. Peters, F. J. van den Hoogen, E. Oosterwijk, A. J. van der Kogel, and J. H. Kaanders (2005). Colocalization of carbonic anhydrase 9 expression and cell proliferation in human head and neck squamous cell carcinoma. *Clinical Cancer Research 11*, 97–106.

Horsman, M. R. (1998). Measurement of tumor oxygenation. *International Journal of Radiation Oncology, Biology, and Physics 42*, 701–704.

Hoskin, P. J., A. Sibtain, F. M. Daley, and G. D. Wilson (2003). GLUT1 and CAIX as intrinsic markers of hypoxia in bladder cancer: relationship with vascularity and proliferation as predictors of outcome of ARCON. *British Journal of Cancer 89*, 1290–1297.

Hckel, M. and P. Vaupel (2001). Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects. *Journal of National Cancer Institute 93*, 266–276.

Iakovlev, V. V., M. Pintilie, A. Morrison, A. W. Fyles, R. P. Hill, and D. W. Hedley (2007). Effect of distributional heterogeneity on the analysis of tumor hypoxia based on carbonic anhydrase IX. *Laboratory Investigation 87*, 1206–1217.

Jain, R. K. (2005). Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. *Science 307*, 58–62.

Jaynes, E. T. (1978). Where do we stand on MaxEnt ? Maximum Entropy Formalism Conference.

Jaynes, E. T. (1985). Bayesian methods: General background. In *Maximum Entropy and Bayesian Methods*.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1958). *Theory of Probability*. Oxford University Press.

Laplace, P. S. (1774). *Probability of Causes*. John Wiley and Sons.

Lilford, R. J. and D. Braunholtz (1996). The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal 313*, 603–607.

Loncaster, J. A., A. L. Harris, S. E. Davidson, J. P. Logue, R. D. Hunter, C. C. Wycoff, J. Pastorek, P. J. Ratcliffe, I. J. Stratford, and C. M. West (2001). Carbonic anhydrase (CA IX) expression, a potential new intrinsic marker of hypoxia: correlations with tumor oxygen measurements and prognosis in locally advanced carcinoma of the cervix. *Cancer Research 61*, 6394–6399.

Lubsen, J. and S. J. Pocock (1994). Factorial trials in cardiology: pros and cons. *European Heart Journal 15*, 585–588.

Mandel, J. (1984). *The Statistical Analysis of Experimental Data*. Dover.

Mayer, A., M. Hckel, and P. Vaupel (2005). Carbonic anhydrase IX expression and tumor oxygenation status do not correlate at the microregional level in locally advanced cancers of the uterine cervix. *Clinical Cancer Research 11*, 7220–7225.

Meinert, C. L. (1986). *Clinical Trials – Design, Conduct, and Analysis*. Oxford University Press.

Milosevic, M., A. Fyles, D. Hedley, and R. Hill (2004). The human tumor microenvironment: invasive (needle) measurement of oxygen and interstitial fluid pressure. *Seminar in Radiation Oncology 14*, 249–258.

Mseide, K., R. A. Kandel, R. S. Bell, C. N. Catton, B. O'Sullivan, J. S. Wunder, M. Pintilie, D. Hedley, and R. P. Hill (2004). Carbonic anhydrase IX as a marker for poor prognosis in soft tissue sarcoma. *Clinical Cancer Research 10*, 4464–4471.

NCI (2009). National cancer institute. This is an electronic document. Date retrieved: April 28, 2009.

Neyman, J. and E. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transaction of the Royal Society, Series A 231*, 289–337.

O'Brien, T. J., L. M. Raymond, G. A. Bannon, D. H. Ford, H. Hardardottir, F. C. Miller, and J. G. Quirk (1991). New monoclonal antibodies identify the glycoprotein carrying the CA 125 epitope. *American Journal of Obstetrics and Gynecology 165*, 1857–1864.

Olive, P. L., C. Aquino-Parsons, S. H. MacPhail, S. Y. Liao, J. A. Raleigh, M. I. Lerman, and E. J. Stanbridge (2001). Carbonic anhydrase 9 as an endogenous marker for hypoxic cells in cervical cancer. *Cancer Research 61*, 8924–8929.

Pitson, G., A. Fyles, M. Milosevic, J. Wylie, M. Pintilie, and R. Hill (2001). Tumor size and oxygenation are independent predictors of nodal diseases in patients with cervix cancer. *International Journal of Radiation Oncology, Biology, and Physics 51*, 699–703.

Raleigh, J. A., M. W. Dewhirst, and D. E. Thrall (1996). Measuring Tumor Hypoxia. *Seminar in Radiation Oncology 6*, 37–45.

RECIST (2009). Response criteria in solid tumors. This is an electronic document. Date retrieved: May 1, 2009.

Rosen, D. G., L. Wang, J. N. Atkinson, Y. Yu, K. H. Lu, E. P. Diamandis, I. Hellstrom, S. C. Mok, J. Liu, and R. C. Bast (2005). Potential markers that complement expression of CA125 in epithelial ovarian cancer. *Gynecology Oncology 99*, 267–277.

Royall, R. (1997). *Statistical Evidence: A Likelihood Primer. Monographs on Statistics and Applied Probability 71.* Chapman and Hall.

Rustin, G. J., R. C. Bast, G. J. Kelloff, J. C. Barrett, S. K. Carter, P. D. Nisen, C. C. Sigman, D. R. Parkinson, and R. W. Ruddon (2004). Use of CA-125 in

clinical trial evaluation of new therapeutic drugs for ovarian cancer. *Clinical Cancer Research 10*, 3919–3926.

Senn, S. (1993). *Cross-over Trials in Clinical Research*. John Wiley and Sons.

Shannon, C. E. (1971). *Theory of Information*. John Wiley and Sons.

Sivia, D. S. (2006). *Data Analysis: A Bayesian Tutorial*. Oxford University Press.

Smith, L. H., C. R. Morris, S. Yasmeen, A. Parikh-Patel, R. D. Cress, and P. S. Romano (2005). Ovarian cancer: can we make the clinical diagnosis earlier? *Cancer 104*, 1398–1407.

Tatum, J. L., G. J. Kelloff, R. J. Gillies, J. M. Arbeit, J. M. Brown, K. S. Chao, J. D. Chapman, W. C. Eckelman, A. W. Fyles, A. J. Giaccia, R. P. Hill, C. J. Koch, M. C. Krishna, K. A. Krohn, J. S. Lewis, R. P. Mason, G. Melillo, A. R. Padhani, G. Powis, J. G. Rajendran, R. Reba, S. P. Robinson, G. L. Semenza, H. M. Swartz, P. Vaupel, D. Yang, B. Croft, J. Hoffman, G. Liu, H. Stone, and D. Sullivan (2006). Hypoxia: importance in tumor biology, noninvasive measurement by imaging, and value of its measurement in the management of cancer therapy. *International Journal of Radiation Oncology, Biology, and Physics 82*, 699–757.

Varia, M. A., D. P. Calkins-Adams, L. H. Rinker, A. S. Kennedy, D. B. Novotny, W. C. Fowler, and J. A. Raleigh (1998). Pimonidazole: a novel hypoxia marker for complementary study of tumor hypoxia and cell proliferation in cervical carcinoma. *Gynecology Oncology 71*, 270–277.

Vaupel, P. (2004). Tumor microenvironmental physiology and its implications for radiation oncology. *Seminar in Radiation Oncology 14*, 198–206.

Vaupel, P. and L. Harrison (2004). Tumor hypoxia: causative factors, compensatory mechanisms, and cellular response. *Oncologist 9 Suppl 5*, 4–9.

Vaupel, P. and A. Mayer (2007). Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Reviews 26*, 225–239.

Wadsworth, H. M. (1997). *Handbook of Statistical Methods for Engineers and Scientists*. McGraw-Hill.

Welsch, S., L. Wang, R. Gunawan, H. Mackay, K. Nathwani, K. Yau, J. Tsang, K. MacAlpine, S. Sivaloganathan, and A. Oza. Ca125 response assessment in patients with recurrent ovarian cancer undergoing clinical trials of molecularly targeted agents. *Gynecologic Oncology (to be submitted)*.

Yin, B. W., A. Dnistrian, and K. O. Lloyd (2002). Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. *International Journal of Cancer 98*, 737–740.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons.