universität wien

DISSERTATION

Bayesian Sequencing of Radiocarbon Dates - - Problems and Improvements

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr. rer.nat.)

Verfasser:	Mag. Franz Weninger
Matrikel-Nummer:	8501664
Dissertationsgebiet:	411 Physik
Betreuer:	emer. O. Univ.-Prof. Dr. Walter Kutschera

Wien, Juni 2011

KURZBESCHREIBUNG - (GERMAN ABSTRACT)

In den letzten zwei Jahrzehnten etablierte sich die Bayes'sche Sequenzierung als ein wichtiges Hilfsmittel in der Radiokarbondatierung. Sie ist eine effektive Antwort auf die Tatsache, dass die normale Kalibrierung von einzelnen Radiokarbon Messungen häufig auf Grund von Plateaus und Wellen in der Kalibrierkurve nur Alter mit hohen Unsicherheiten liefert. Datiert man aber eine ganze Serie von Proben, zum Beispiel solche aus ein und derselben archäologischen Grabung, dann liefert die Bayes'sche Sequenzierung durch Einbeziehung zusätzlicher, so genannter 'a priori'-Information über die Altersrelationen zwischen den verschieden Proben, welche man aus den stratigraphischen Gegebenheiten gewinnt, Ergebnisse mit kleineren Unsicherheiten. Allerdings krankt diese Methode auch selbst an einem für die Bayes'sche Statistik charakteristischen Problem: Die 'a priori' Information muss als Wahrscheinlichkeitsverteilung ausgedrückt werden um im Formalismus verwendet werden zu können. Üblicherweise ist die vorhandene Information aber nicht ausreichend um diese Verteilung eindeutig zu bestimmen, wodurch unterschiedliche Ergebnisse möglich werden. Das Hauptmotiv der vorliegenden Arbeit war es, diese intrinsische Mehrdeutigkeit der Bayes'schen Sequenzierung zu analysieren und eine Methode zu entwickeln, um dieses Problem zu verhindern oder zu reduzieren.

Nach einer allgemeinen Einführung im ersten Kapitel wird im zweiten die mathematische Basis der Bayes'schen Sequenzierung diskutiert. Darüber hinaus wird das Prinzip der 'Gibbs-sampling'-Prozedur, einer verwendeten Monte Carlo-Methode erklärt, und ein zur Durchführung aller erforderlichen Berechnungen entwickeltes Programm kurz beschrieben.
Um dem oben angesprochenen Problem der Prior-Mehrdeutigkeit zu begegnen, ist das Hauptanliegen dieser Dissertation eine anwendungsspezifische Realisierung einer bestimmten Variante der 'robusten Bayes'schen Analyse', eines in der Bayes'schen Statistik bereits bekannten Konzepts. Die Grundidee ist die simultane Verwendung aller möglichen unterschiedlichen Priorwahrscheinlichkeitsverteilungen, sofern diese mit der bekannten Priorinformation konsistent sind und das Gesamtresultat aus einer Art Vereinigung aller Einzelresultate zu gewinnen. Die Motivation für diesen doch eher extensiven Zugang kann man verstehen, wenn man die grundsätzliche Wirkung der Priorwahrscheinlichkeitsverteilung - kurz 'Prior' - analysiert, wie das im dritten Kapitel gezeigt wird.
Nach einer genauen Diskussion im vierten Kapitel von Maßen für die Verträglichkeit eines Priors mit den Radiokarbonmesswerten, welche benötigt werden um 'extreme' Priore auszuscheiden, die das Ergebnis unbrauchbar machen würden, werden im fünften Kapitel unterschiedliche Zugänge zur Realisierung der robusten Sequenzierung analysiert. Grob gesprochen verbleiben letztlich zwei grundsätzlich unterschiedliche Varianten: Eine Art gewichtete Summation der verschieden priorabhängigen Ergebnisse, welche durch eine kontinuierliche Variation eines parametrisierten Priors verwirklicht werden kann und andererseits eine nicht gewichtete Vereinigung der Resultate eines Satzes von diskreten Prioren. Obwohl die erste Methode aus mehreren Gründen vorteilhaft wäre zeigte sich aber, dass nur die zweite der ursprünglichen Idee der robusten Analyse wirklich nahe kommt.
Im letzten Kapitel werden die Eigenschaften der robusten Sequenzierung an Beispielen demonstriert. Unter anderem wird mit Hilfe von spezifisch konstruierten Beispielen die Fähigkeit der robusten Sequenzierung demonstriert 'Artefakte' zu vermeiden, welche bei der Verwendung eines einzelnen, üblichen Priors vorkommen. Abschließend wird eine umfangreiche reale Probensequenz mit der robusten Methode analysiert.

Insgesamt zeigt die vorliegend Arbeit bereits, dass die robuste Sequenzierung ein vielversprechender Weg zur Erhöhung der Zuverlässigkeit der Bayes'schen Sequenzierung ist, obgleich noch weitere Verbesserungen der Methodik denkbar sind.

ABSTRACT

In the last two decades Bayesian sequencing has been established as a powerful tool in radiocarbon dating. It is an efficient answer to the fact, that the basic single-sample calibration procedure of radiocarbon dates frequently generates results with large uncertainties, caused by plateaus and wiggles in the calibration curve. In case of dating a whole set of samples, as for example such excavated together at one particular archaeological site, Bayesian sequencing can reduce the uncertainties by considering additional, so-called 'a priori' information on the age relations of the individual samples deduced from the stratigraphic evidence. However, as the method is Bayesian, it suffers from a fundamental problem of Bayesian statistics: The 'a priori' information has to be expressed as a probability distribution to be used within the formalism. Unfortunately, the available information is usually not detailed enough that this can be done in an unambiguous way, which allows for various different outcomes. The main motivation for the current work was to analyse this intrinsic arbitrariness of Bayesian sequencing and to develop a method to avoid or reduce the problem.

After a general introduction in the first chapter, the second chapter gives a precise discussion of the mathematical framework of the Bayesian sequencing. Additionally the principles of the 'Gibbs-sampling' procedure, a Monte Carlo method for the numerical realisation, and a brief description of the developed program code that carries out all needed calculations, are given.
In order to overcome the prior-ambiguity problem mentioned above, the major approach within this thesis is to introduce a specific realisation of a particular form of 'robust Bayesian analysis', which is a concept already known in Bayesian statistics. The basic idea is to use simultaneously all possible, differently shaped prior probability distributions that are consistent with the known prior information, and generate the final result as a kind of unification of all individual results. The motivation for this extensive approach can be understood by analysing the impact of the prior probability distribution in principle, as illustrated in the third chapter.
After a detailed discussion in the fourth chapter of measures for the agreement between prior distribution and radiocarbon data, which are needed to discard somehow 'extreme' priors that would destroy the result of the method, different approaches to realise robust sequencing are analysed in the fifth chapter. Roughly speaking, there remain two fundamental different realisations: A kind of weighted summation of the different prior-dependent results, which can be realised by a continuous variation of a parametric prior on one hand, and a non-weighted unification of the results of a set of discrete priors on the other hand. Although the first method is advantageous for different reasons, it turned out that just the latter is really close to the originally idea of robust analyses.
In the last chapter the characteristics of robust sequencing are illustrated by examples. Amongst others, specific artificial examples are used that demonstrate the ability of robust sequencing to eliminate 'artefacts' that occur when using a common single prior. Finally, the robust approach is applied to a large real-world sequence.

All in all, the current work shows that robust sequencing is a promising way to improve the reliability of Bayesian sequencing, however leaving room for further refinements.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 RADIOCARBON DATING

Radiocarbon dating has become the most powerful scientific dating method for archaeological applications, since it had been developed by Willard Libby in the nineteenfourties. A great benefit of this dating method is that the underlying principle is very strait forward and reliable. Unfortunately there are effects that disturb this ideal principle, so that much effort is needed to recover the full potential of the method. 'Bayesian sequencing', which is the topic of this work, can be an important contribution to this intention.

1.1.1 Basic priciple

Radiocarbon dating rests upon the fact that plants accumulate carbon from the carbon dioxide available in the air. The main isotope of the atmospheric carbon is ^{12}C, but there is also a tiny fraction of the radioactive ^{14}C isotope that is produced by cosmic rays in the atmosphere through the interaction of secondary neutrons with ^{14}N. Therefore the plants, and via the food chain also the animals, show this particular $^{14}C/^{12}C$ ratio, which is roughly $1.2 \cdot 10^{-12}$ (comparing the number of atoms). After the death of a living organism there is no further incorporation of carbon, and from this moment the $^{14}C/^{12}C$ ratio will decrease due to the radioactive decay of the ^{14}C nuclei. Knowing the initial $^{14}C/^{12}C$ ratio at the lifetime of an organic sample, which is equal to the atmospheric level (denoted with 'modern' below), and measuring the present ratio within the sample (denoted with 'sample'), the sample age x can be determined easily from the decay law:

Equation 1.1:
$$x = \frac{t_{1/2}}{\ln(2)} \cdot \ln\left(\frac{(^{14}C/^{12}C)_{\text{sample}}}{(^{14}C/^{12}C)_{\text{modern}}}\right)$$

Where $t_{1/2}$ is the half life of the radioactive ^{14}C. Libby used a value of 5568 a; in our days the half-life is determined to be 5730 ±40 a (GODWIN, 1962). A $^{14}C/^{12}C$ ratio can be measured by counting the β-decay events of a sample with known (via weight) ^{12}C content, or by measuring the ratio in an accelerator mass spectrometer directly. The latter has the advantage that much less sample mass is needed (factor of ~1000).
Except for very rough estimates, the formalism above cannot be used directly, because of the slightly different behaviour of the two carbon isotopes during the formation of the sample, called isotopic fractionation, and also because of temporal variations of the atmospheric $^{14}C/^{12}C$ ratio. Fortunately there are solutions for both problems, which are briefly described in the next two sections.

1.1.2 Correction of the isotopic fractionation

If the fractionation during formation of the dated material is the same than that of the reference material from which the $(^{14}C/^{12}C)_{modern}$ ratio is derived - Libby based the system on wood - no error would occur. But how can samples be dated exactly, from which the rate of fractionation is unknown, e.g. from an animal whose food consisted of unknown amounts of different unknown plants? The solution is provided by the minor stable isotope of carbon, ^{13}C, whose abundance is 1.1 %. Treating the fractionation process as mass-proportional, which is often a good approach for chemical reactions, the rate of fractionation between ^{13}C and ^{12}C can be thought to be the same as that between ^{14}C and ^{13}C, because of their similar ratios of atomic mass. Believing this, it can be shown easily that the term

$$\left(\frac{(^{13}C/^{12}C)_{wood}}{(^{13}C/^{12}C)_{sample}} \right)^2$$

corrects the ^{14}C concentration of a sample to that level that would have occurred with a wooden sample. Thus, the calculated sample age from above is modified to:

Equation 1.2:
$$x = \frac{t_{½}}{\ln(2)} \cdot \ln \left(\frac{(^{14}C/^{12}C)_{sample}}{(^{14}C/^{12}C)_{modern}} \cdot \left(\frac{(^{13}C/^{12}C)_{wood}}{(^{13}C/^{12}C)_{sample}} \right)^2 \right)$$

Using internationally agreed assumptions given below, this formula gives the standardised (uncalibrated) so-called 'radiocarbon age'. The assumptions are the following: For the half-life the value of 5568 a is used, which was introduced by LIBBY (1952). Although there is a value with higher accuracy available now, the use of Libby's value still makes sense, because today the radiocarbon age is not interpreted directly, it is more or less an artificial base for an additional calibration process, as shown next. The values for $(^{14}C/^{12}C)_{modern}$ and $(^{13}C/^{12}C)_{wood}$ are fixed by relating them to primary standards. In the measurement process, both the $^{14}C/^{12}C$ and the $^{13}C/^{12}C$ ratios of the samples are often determined relative to secondary standards, which are based on the former. The $^{14}C/^{12}C$ ratio of the standards decreases at the same rate as the one of the samples, so that relative measurements give, independently from the time of the measurement, ages related to the year 1950 AD, the year on which the system is based. Ages determined in this way are conventionally termed as 'before present' (BP), where present means always 1950.
It should be mentioned that the notation used in the formulas above is the simplest one to see the principles, but it is not the commonly used in dating practice. There, the pMC values (percentage modern carbon) or the $F^{14}C$ values (fraction modern) are used, that contain the fractionation correction in an implicit way.

1.1.3 Correction of the temporal variations of the atmospheric $^{14}C/^{12}C$ ratio

The method described above is based on the assumption, that there is a constant atmospheric $^{14}C/^{12}C$ ratio that determines the initial ratio in a living organism. In reality the ratio shows considerable temporal changes, mainly caused by the variation of the field strength of the geomagnetic field, which shields the cosmic rays of

charged particles, and therefore influences the production rate of radiocarbon (^{14}C). In addition, changes in solar activity and in ocean-atmosphere coupling (CO_2 exchange) also influences the atmospheric $^{14}C/^{12}C$ ratio to some extend. To solve this problem, a source to determine the $^{14}C/^{12}C$ ratio of the past is needed. Most important therefore are fossil trees that are dated by dendrochronology, which is a method that strings together tree-ring sequences by comparing characteristic variations of the ring thicknesses. Based on this fossil wood (and other sources), a calibration curve had been constructed that relates the (uncalibrated) radiocarbon age from the formula above to a real age (calendar date). The commonly used calibration curve is given by REIMER *et al.* (2009), and is called IntCal09. For real ages up to 12000 years, covering most archaeological applications, the curve is based on well-established tree-ring sequences and remained equal to the former version (IntCal04; REIMER *et al.*, 2004). Figure 1.1. shows the calibration curve for the last 5000 years.

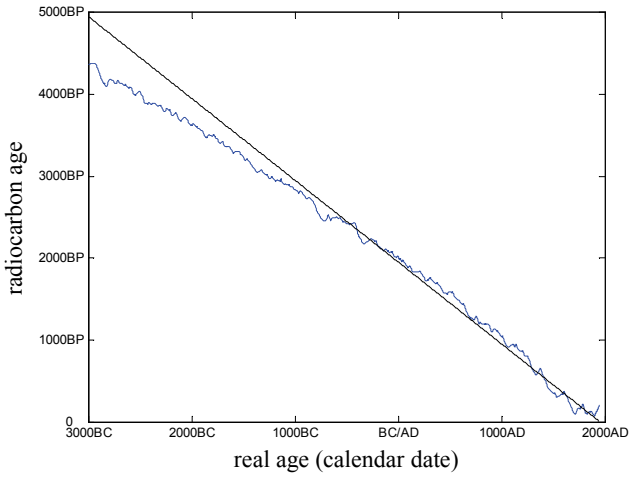

Figure 1.1: The calibration curve (IntCal09) gives the radiocarbon age as a function of the real age. The straight line would be the relation of the radiocarbon age and real age without calibration, reflecting 1950 as 'present'. Note that a particular radiocarbon age results frequently in more than one possible real ages.

With the help of the calibration curve, the real sample age can be deduced from the radiocarbon age in principle. Unfortunately the calibration curve is often ambiguous, so that even a radiocarbon age without error, would result in more than one possible real age. This is the fundamental problem of the calibration procedure for radiocarbon dates (see e.g. GUILDERSON *et al.*, 2005). Furthermore, the real calibration process includes the measurement error of the radiocarbon age. Therefore the radiocarbon age is put into the calibration as a Gaussian probability density distribution due to the measurement error, which then is transformed by the calibration curve to the real age axis, resulting (after normalisation and under the assumption that any real age is previously equal probable) in a probability density distribution of the real age. Now, ambiguous parts or 'wiggles' in the calibration curve will produce various local maxima within a spread-out probability density.
The mathematical formulation of the calibration process is given in Equation 1.3. Additionally a graphical illustration is shown in Figure 1.2:

Equation 1.3:
$$\mathrm{pdf}(t) \propto \exp\left(-\frac{(x-c(t))^2}{2\sigma^2}\right)$$

Where pdf(t) is the probability density distribution (or function) of the real age t, x is the determined radiocarbon age, σ its uncertainty and $c(t)$ the calibration curve. It should be mentioned that the calibration curve is treated here without considering its

own measurement error, which will be introduced later. (A further short remark: The calibration of radiocarbon dates with the help of Equation 1.3 is the usual convention but not the only possible way in general; see e.g. DEHLING and VAN DER PLICHT, 1993. In the Bayesian description it reflects the use of a constant prior, as pointed out in section 2.2.2.)
The formalism described so far, gives the basics of radiocarbon dating, including the correction of the most important irregularities, which are the isotopic fractionation and the variation of the atmospheric ^{14}C concentration. Naturally, when looking at the method in more detail, various additional sources of systematic errors (including sample preparation and measurements) have to be avoided or corrected as far as possible. However, this will not be discussed here, because the Bayesian method, which is the topic of this work, does not influence the procedures to get the (uncalibrated) radiocarbon age, it focuses only on the last step, the calibration process as described above.

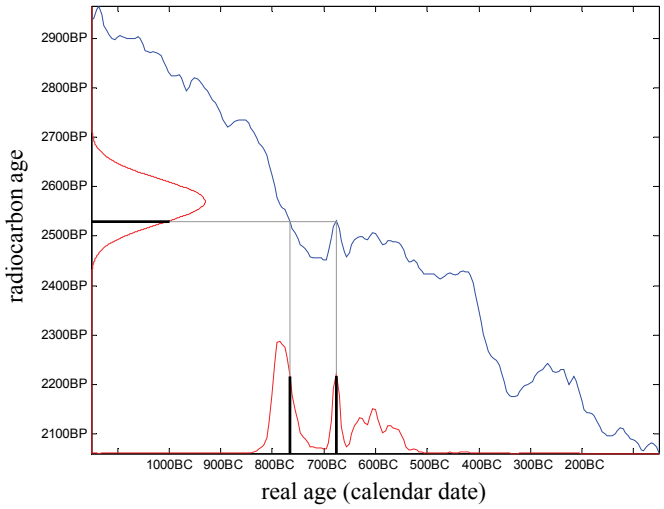

Figure 1.2: Graphical demonstration of the calibration process. The Gaussian shaped curve on the vertical axis gives the probability density of the determined radiocarbon age due to the accuracy of the measurement (2570±45 years BP for this example). The curve on the horizontal axis gives the probability density of the real age[*)], constructed as indicated by the use of the calibration curve (blue). The probability densities are given in arbitrary units; the density axes are not plotted.
([*)] after normalisation and under the assumption that any real age is primarily equal probable)

For a more detailed view on the radiocarbon dating procedure see e.g. BOWMAN (1990). An exact formulation of standardisation and fractionation correction of ^{14}C concentrations (but not especially for radiocarbon dating) is given by MOOK and VAN DER PLICHT (1999). Details to the calibration curve and calibration process are given by REIMER *et al.* (2004 and 2009), mentioned already above.

1.2 ARCHAEOLOGICAL KNOWLEDGE ABOUT SAMPLES

Sometimes only a single sample is available for radiocarbon dating. However, this is rather unusual for archaeological excavations. In many cases a whole set of samples can be taken from an excavation to be dated. One can imagine that the location of the samples within the specific structure of the excavation can tell a lot about the relations of their ages (real ages). Archaeologists mostly characterise the structure of an excavation by so called stratigraphies. A stratigraphy describes a sequence of strati, and a stratum is simply spoken an observable confined layer within the excavation that represents a short time period of the site, well defined by particular changes. An example, easy to visualise, would be the repeated rebuilding of a house,

where the old flour is covered with filling material and a new floor is constructed on top of this material. This would result in a clearly observable sequence of strati, namely layers of filling material bounded with floors, representing a chronological sequence. Samples found in a higher stratum are consequently known to be younger than those from a lower stratum. Naturally, this rule can be broken by irregular cases, e.g. if a higher layer was filled with old material dug out elsewhere and so on. However, archaeologists are well used to deal with difficulties of that kind. A more general example for a stratigraphy is a sequence of strati that represent whole cultural phases. The latter are frequently identifiable by a characteristic style of pottery and may be separated by dramatic events as warlike destructions, possibly clearly observable by an ash layer of a fire destruction. Furthermore, even layers found on different locations of an excavation, although characterised by similar typical findings as e.g. pottery of equal style, can be assigned to the same temporal phase, and thus they are chronologically connected. In this way archaeologists establish a stratigraphy for the whole excavation, defining time relations of samples taken out of the strati. The association of cultural phases by characteristic pottery is not restricted to a single excavation, but is also done for different sites, if there was an exchange of pottery by trade. Thus, actually for samples found at distant places, e.g. on Cyprus and Crete, relative age relations can be established.
Stratigraphies as just described, result mainly in sample age relations stating that one group of samples has to be older or younger than another one. However, there can be other information about samples leading to various kinds of time relations. For example, if one finds wooden construction material with identifiable tree rings, it may be possible to take samples from single rings. In this case one knows the exact age difference of the samples, according to the number of rings between them. A comparable situation could be the knowledge that one sample is related to the beginning of the reign of an Egyptian Pharaoh and another one to the end, and the duration of the reign is known from temple inscriptions. But in this case one is on the way to change from archaeological to historical sources. A geological example could be the following: Samples are embedded in sediments that where formed with an approximately constant sedimentation rate. Consequently the local distances of the samples would be proportional to their age differences. On can imagine that there are many cases leading to different types of age relations, but in archaeological practice the younger-older relations resulting from stratigraphies are the most important ones.
It should be noted here, that this presentation of the archaeological situation is naturally only a very rough one, due to the limitations in understanding of the field by a physicist. Therefore, when using archaeological information in real dating applications, the description of the archaeological information available from the excavation has always to be performed by archaeologists.

1.3 BETTER RESULTS BY COMBING AVAILABLE INFORMATION

As described in section 1.1.3, the radiocarbon dating procedure has the serious disadvantage, that due to wiggles in the calibration curve the resulting probability density for the real sample age can be spread out over a wide time range. Fortunately, as pointed out in section 1.2, in many cases a series of samples with various known relations between their ages, deduced from the excavation site, is available. It will be

demonstrated now with a most simple but very illustrative example, that the combination of radiocarbon measurements with the available archaeological information can provide a significant shortening of the ranges for the possible real ages of the samples. In the 14th century AD the calibration curve shows a very significant wiggle, that fits very well for this consideration. Imagine one would have done two radiocarbon measurements falling in this time period as shown in Figure 1.3. The single sample calibrations of both measurements result in widely spread, double peaked probability densities. But now we assume that the stratigraphy gives the evidence, that sample 1 is older than sample 2. It is clear, that only the marked parts of the probability densities fulfil this constraint and the remaining parts can be excluded. Although, it is easy to see in which way the probability densities have to be modified in this simple example, one can imagine that this can not be done as easily, when many samples and more complex age relations are included. In the following chapter the mathematical framework is discussed that offers the possibility to include the archaeological information in a general way, leading to a gradual shaping of the sample age probability densities.

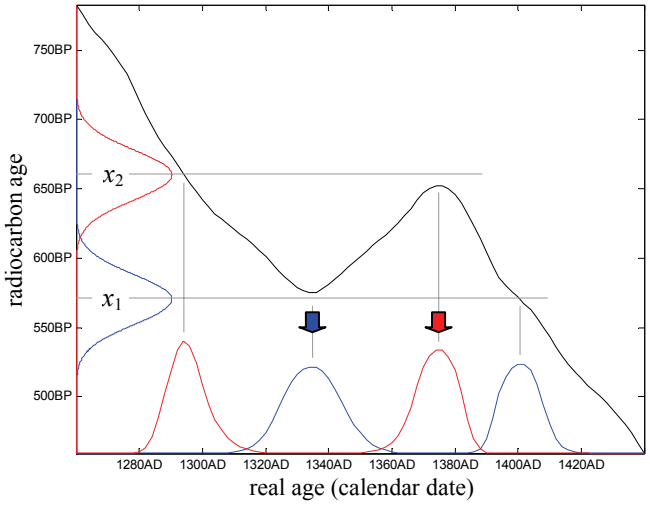

Figure 1.3: Illustration of the possible benefit due to additional information. The single sample calibrations of two radiocarbon measurements (x_1 blue and x_2 red) are shown on a significant wiggle of the calibration curve (black). Assuming a stratigraphic evidence, that sample 1 is older than sample 2, only the marked parts of the probability densities remain possible on logical reasons.

2 BAYESIAN SEQUENCING OF RADIOCARBON DATES: FORMALISM AND NUMERICAL REALISATION

This chapter gives a detailed introduction to the formalism of Bayesian sequencing, and also to the Bayes theorem, on which the method is based. The numerical realisation of the method via Gibbs sampling, a Markov chain Monte Carlo procedure, is further described precisely. The content of the chapter is common knowledge in principle, excepted the description of the developed specific program package (section 2.7), which realises the basic procedures as well as new investigations described in chapter 4 and 5.

2.1 DESCRIPTION OF THE BAYESIAN METHOD WITH A SIMPLE EXAMPLE

Here the mathematical method - named Bayesian sequencing or Bayesian multi-sample calibration - is introduced, which is able to change the probability densities of the real sample ages considering the additional stratigraphic information, as described in the example above. The theoretical base of the method is an application of the Bayes theorem, which was developed in the eighteenth century by Reverend Thomas Bayes (BAYES, 1763). However, the used notation is optimised to clarify the specific application and hides the theoretical background in some sense. Theoretical aspects are discussed more exactly in section 2.3.
The method will be demonstrated on the simple example from above, but with changed radiocarbon values producing more overlap of the single sample calibrations. This modification brings the example a step closer to real situations. One starts again from the single sample calibrations shown in Figure 2.1, called now single sample likelihoods or likelihood functions, expressions which are used in the Bayes theorem. It should be noted, that for this example real ages are also given in years BP (related to 1950 AD), because it is inconvenient to formulate mathematical relations on a BC/AD scale.
The mathematical expressions of the single-sample likelihood functions for sample 1 (l_1) and sample 2 (l_2) are the following, as already explained in section 1.1.3.

Equations 2.1:
$$l_1(t_1) \propto \exp\left(-\frac{(x_1 - c(t_1))^2}{2\sigma_1^2}\right), \qquad l_2(t_2) \propto \exp\left(-\frac{(x_2 - c(t_2))^2}{2\sigma_2^2}\right)$$

t_1 and t_2 are the unknown real ages of the corresponding samples, x_i are the measured radiocarbon ages, σ_i are their uncertainties and $c(t)$ is the calibration curve. Generally, all probability density distributions will be given without normalisation, remarked by using 'proportional' instead of 'equal'.

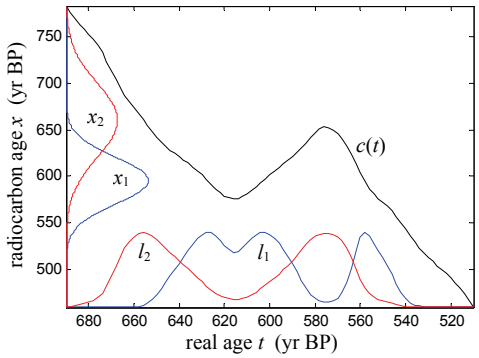

Figure 2.1: To demonstrate the principle of the Bayesian method, again two measurements are assumed to fall within the range of the significant wiggle of the calibration curve: x_1 = 595±25 yr BP and x_2 = 660±40 yr BP. The resulting single-sample likelihood functions l_1 and l_2 are the base for further calculations. (Normalisation as used in the plot is not required within the calculation.)
Note that now yr BP is used for the real ages too. (Some authors use the notation 'cal BP' in this case.)

The next step is to link both single-sample likelihood functions together within a two-dimensional mathematical space, spanned by the axes for the real ages of sample 1 and 2 (t_1 and t_2). The introduction of this two-dimensional space is needed to make a mathematical formulation of the available archaeological information possible, as shown below. Thus, a two-dimensional likelihood function ($l(t_1,t_2)$) is constructed, simply by multiplication of both single-sample likelihood functions ($l_1(t_1)$ and $l_2(t_2)$) point by point; see Figure 2.2 and Equation 2.2. In the following, the short notation 'age' stands always for a real age t, never for a radiocarbon age x.

Equation 2.2: $$l(t_1,t_2) \propto l_1(t_1) \cdot l_2(t_2)$$

The two-dimensional likelihood function can be interpreted analogous to the one-dimensional ones: It represents the probability density distribution for pairs of particular values for sample age 1 and 2, but again under the condition, that any pair is previously equally probable (and after normalisation). Each pair is represented by a particular point within the two-dimensional plane in the graph.
The available archaeological information is now mathematically formulated within the same two-dimensional space. Again - as in the qualitative discussion above - the assumption is, that from stratigraphical evidence, sample 1 has to be older than sample 2. The functional representation of this archaeological fact is shown in Figure 2.3 and given by Equation 2.3.

Equation 2.3: $$a(t_1,t_2) \propto \left\{ \begin{array}{l} 1 \text{ if } t_1 > t_2 \\ 0 \text{ if } t_1 \le t_2 \end{array} \right\}$$

Where $a(t_1,t_2)$ is the so-called 'a priori' probability density distribution or shortly prior function, because it reflects the status of knowledge previously to the measurements. The fact, that the given example represents - exactly spoken - no probability density, because it cannot be standardised due to its unrestricted extension, does not matter at present. However, seeing this function at the first time, it seems to be the only reasonable representation of the given information, namely that age 1 has to be older than age 2. Unfortunately, this impression turns out to be not correct when analysed exactly. The resulting consequences will be discussed in section 3.1 in detail. Disregarding the last remark for now, one has obtained two two-dimensional probability densities $l(t_1,t_2)$ and $a(t_1,t_2)$ so far, the first representing the

radiocarbon measurements including calibration, and the second the archaeological knowledge.

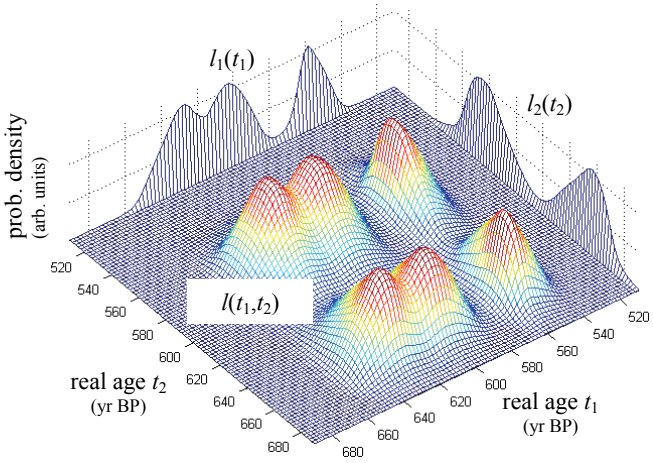

Figure 2.2: The two-dimensional likelihood function $l(t_1,t_2)$ is the pointwise product of the single sample calibrations $l_1(t_1)$ and $l_2(t_2)$. It represents the probability density distribution for pairs of sample age 1 and 2 based on the measurements, and under the condition, that any pair is previously equal probable. The available archaeological information is still ignored at this step.

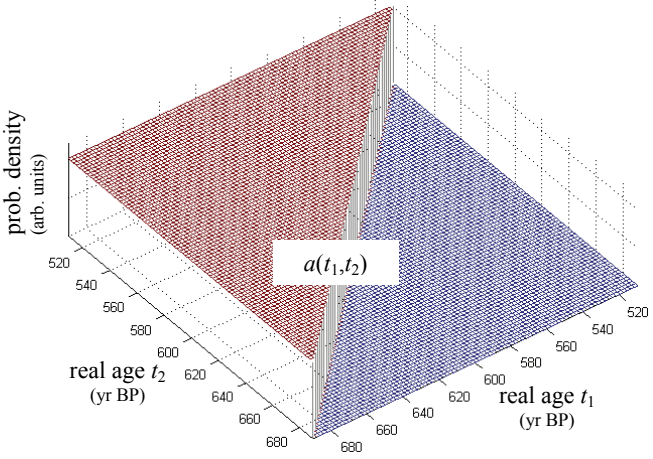

Figure 2.3: The prior function $a(t_1,t_2)$ carries the archaeological information that sample 1 is older than sample 2. Only age pairs achieving this condition are located within the region with non-zero probability density.

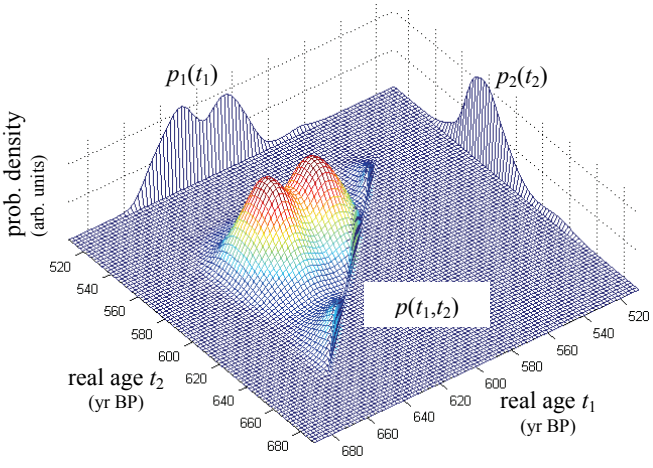

Figure 2.4: The posterior $p(t_1,t_2)$ is the resulting probability density for pairs of sample age 1 and 2, including both, the information from the radiocarbon measurement and the additional archaeological information. All parts of the likelihood function located in the region with zero prior probability have been suppressed. Projecting $p(t_1,t_2)$ onto the sample age axes results in the so-called marginal posterior probability densities of the individual sample ages $p_1(t_1)$ and $p_2(t_2)$.

The following step is the combination of both functions by multiplication, resulting in the so-called posterior probability density $p(t_1,t_2)$ or posterior function shown in Figure 2.4 and Equation 2.4.

Equation 2.4: $$p(t_1,t_2) \propto l(t_1,t_2) \cdot a(t_1,t_2)$$

This equation represents the Bayes theorem in simple notation; for the theoretically exact notation see section 2.3. The posterior probability density $p(t_1,t_2)$ is the resulting probability for any pair of sample age 1 and 2, including both the information from the radiocarbon measurement and the additional archaeological information. In this example, three of the four regions of sample age combinations with high probability in the likelihood function (representing the measurement) are mainly located in the region with zero prior probability (representing the stratigraphic information) and therefore they are strongly suppressed within the posterior probability.
$p(t_1,t_2)$ is already the result one has been looking for, but still represented in terms of sample age combinations. For this, the last step is to go back from the two-dimensional space of age combinations to common probability densities for the single samples ages. This is done by projecting the two-dimensional posterior probability density onto the individual sample age axes, resulting in the so-called marginal posterior probability densities $p_1(t_1)$ and $p_2(t_2)$, as shown in Figure 2.4 and given by Equations 2.5.

Equations 2.5: $$p_1(t_1) \propto \int_{-\infty}^{+\infty} p(t_1,t_2)\ \mathrm{d}t_2\ , \qquad p_2(t_2) \propto \int_{-\infty}^{+\infty} p(t_1,t_2)\ \mathrm{d}t_1$$

The marginal posterior probability densities for the individual sample ages give the common representation of the final result of the method, which is the mathematical combination of radiocarbon dating with the available archaeological information.

Figure 2.5 gives a concluding comparison of the single-sample likelihood functions (or common single sample calibrations in other words) with marginal posterior distributions for the individual samples resulting from the Bayesian method. One can see clearly, that these parts of the of the probability densities are suppressed, which are in disagreement with the assumed condition, that sample 1 is older than sample 2, just as expected from the method initially.

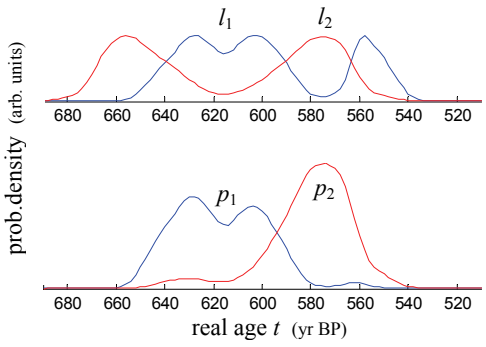

Figure 2.5: Concluding comparison of the single sample calibrations or likelihood functions l_1 and l_2 with the resulting marginal posterior probability densities p_1 and p_2. The former carry only the information due to the radiocarbon measurements, the latter include both the information from measurements and the archaeological information. Parts of the probability densities that are in disagreement with the assumed condition that sample 1 is older than sample 2, are suppressed within the posterior densities.

Although the given example explains the principle of Bayesian sequencing completely, the figures and equations are given within a two-dimensional space. In general, the dimensionality of the likelihood, prior, and posterior function is equal to the number of the samples, and can be quite high. While it is not possible to visualise

high-dimensional functions, the mathematical formulation of the method remains the same in principle. The generalised form of the equations is given in the next section.

Finally, a short remark to avoid possible misunderstandings concerning the prior function is given. It is not fundamental that the prior reflects only a yes or no decision as in the example above. For example, a different case would be a prior function deduced from the information that sample 1 is older than sample 2 by the particular time span of 50±15 years, shown in Figure 2.6. An information of this kind could e.g. be found for two samples taken from different parts of the same tree trunk, where the number of tree rings between the two samples could only be determined with the given accuracy, due to a bad state of preservation.

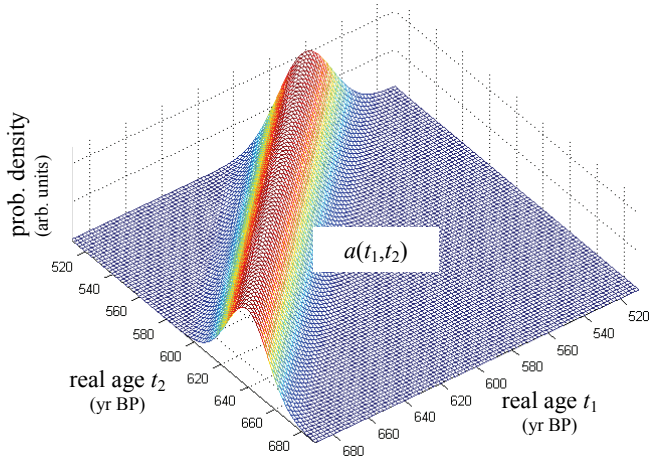

Figure 2.6: An alternative example for a prior function. This prior would represent the information that sample 1 is older than sample 2 by the particular time span of 50±15 years.

2.2 THE MATHEMATICAL FORMALISM IN GENERAL

2.2.1 The multi-dimensional formulation of the basic equations

Up to now the equations where given treating the calibration curve without uncertainty. Actually, the calibration curve is known only with a given accuracy due to the underlying measurements. This fact can be taken into account by the summation of the squares of the uncertainties of the measured radiocarbon age and of the age-dependent uncertainty of the calibration curve itself. But this is exactly true, only in the normalised formalism, as given in the next section; Equation 2.6 is - strictly speaking - an approximation. Details to this questions can be found in section 2.2.3.
The equations introduced in the simple example above will now be generalized to a notation for a free number of dimensions or a free number of samples in other words.

The single sample likelihoods l_i are given analogous to Equations 2.1 as:

Equation 2.6:
$$l_i(t_i) \quad \propto \quad \exp\left(-\frac{\left(x_i - c(t_i)\right)^2}{2\cdot\left(\sigma_i^2 + \sigma_c^2(t_i)\right)}\right)$$

t_i is the unknown real age for the i^{th} sample, x_i the determined radiocarbon age and σ_i its uncertainty. $\sigma_c(t)$ is the uncertainty of the radiocarbon-age value of the calibration curve $c(t)$ at the real-age position t.
The multi-dimensional likelihood function l follows similarly to Equation 2.2:

Equation 2.7: $$l(t_1, \ldots, t_n) \propto l_1(t_1) \cdot \ldots \cdot l_n(t_n)$$

The prior $a(t_1,...,t_n)$, which reflects the archaeological information on the samples, becomes a function on the multi-dimensional space too. A simple example, analogous to the two-dimensional case from above, could be constructed, assuming that there are older/younger relations between various samples known. Thus, the value of the prior function would be e.g. one at every point within the n-dimensional space where all relations are fulfilled, and zero elsewhere.
The multi-dimensional posterior function is, similar to Equation 2.4, the pointwise product of likelihood and prior for every point $(t_1,...,t_n)$ within the n-dimensional space of the real age axes:

Equation 2.8: $$p(t_1, \ldots, t_n) \propto l(t_1, \ldots, t_n) \cdot a(t_1, \ldots, t_n)$$

Finally, the marginal posterior probability distributions $p_i(t_i)$ are calculated by projecting $p(t_1,...,t_n)$ to the individual sample age axes, as described above. Projection to one specific sample-age axis within the multi-dimensional space means to integrate over all sample-age axes except the one examined:

Equation 2.9: $$p_i(t_i) \propto \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} p(t_1, \ldots, t_n) \, \mathrm{d}t_1 \ldots \mathrm{d}t_{i-1} \, \mathrm{d}t_{i+1} \ldots \mathrm{d}t_n$$

Only the four equations in this section are needed to carry out Bayesian multi-sample calibration in its basic form. Therefore the equations can be used directly with 'equal' signs, because normalisation constants do not affect the method. It is sufficient to normalise the resulting posterior marginals at the end.

2.2.2 The basic equations in detailed notation

In the previous section, the mathematical foundation of the method is given in a notation optimised for the application. Now, the equation will be shown in a completed form, considering normalisation and an exact notation of the probability densities.
Up to now, the single-sample likelihood function or single sample calibration was explained to be the probability density of the sample age, under the assumption that each age is originally equally probable. This was an adequate explanation to clarify the initial example, but it is not the complete definition of the likelihood function. From a more theoretical point of view, the likelihood function is the conditional probability density to determine a particular radiocarbon age x, under the condition of a given sample age t, denoted as $l(x|t)$. Fixing x at the actual determined value, it

becomes a function with the argument t that describes, how likely it is, to get the actual radiocarbon age for any assumed real sample age. Including normalisation the single-sample likelihood function of the i^{th} sample is expressed by Equation 2.10.

Equation 2.10:
$$l_i(x_i|t_i) = \frac{1}{\sqrt{2\pi\cdot\left(\sigma_i^2+\sigma_c^2(t_i)\right)}}\cdot\exp\left(-\frac{\left(x_i-c(t_i)\right)^2}{2\cdot\left(\sigma_i^2+\sigma_c^2(t_i)\right)}\right)$$

Note, that the real-age dependence of the uncertainty of the calibration curve does not destroy the correctness of the normalisation, because it is a normalisation on the radiocarbon axis, but see more to this point in section 2.3.
The complete notation of the multi-dimensional likelihood is:

Equation 2.11:
$$l(x_1,\ldots,x_n|t_1,\ldots,t_n) = l_1(x_n|t_n)\cdot\ldots\cdot l_n(x_n|t_n)$$

$l(x_1,...,x_n|t_1,...,t_n)$ characterizes the likelihood to determine the set of radiocarbon ages $(x_1,...,x_n)$ for any assumed set of sample ages $(t_1,...,t_n)$. It is a conditional probability density in the n-dimensional space of the radiocarbon ages, under a condition represented by a point in the n-dimensional space of the real ages.
The prior function carries information depending on the real ages only, thus it is a common non-conditional probability density in the real age space, denoted $a(t_1,...,t_n)$.
In complete notation, the equation to calculate the posterior $p(t_1,...,t_n|x_1,...,x_n)$ is exactly the Bayes theorem:

Equation 2.12:
$$p(t_1,\ldots,t_n|x_1,\ldots,x_n) = \frac{l(x_1,\ldots,x_n|t_1,\ldots,t_n)\cdot a(t_1,\ldots,t_n)}{\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty} l(x_1,\ldots,x_n|t_1,\ldots,t_n)\cdot a(t_1,\ldots,t_n)\ \mathrm{d}t_1\ldots\mathrm{d}t_n}$$

The volume integral in the denominator can be seen as normalisation factor, although it has a further meaning, shown later. The Bayes theorem will be discussed in detail in section 2.3. Equation 2.12 shows, that by the help of the prior function, from the likelihood function, which is a conditional probability density of possible radiocarbon ages at given real ages, the posterior function, a conditional probability density of the possible real ages at given radiocarbon ages, is derived. Naturally, the latter is the one, one is interested in.
At this point a short clarification to the initial explanation of the meaning of the likelihood function is given: Fixing the radiocarbon ages to the actually determined values and assuming a constant prior function, the likelihood and the posterior become equally shaped functions in the real-age space. This is the reason why the likelihood characterises the probability density of the real ages, assuming previously equal probable ages, which means a constant prior. So the common single sample calibration can be seen as the application of the Bayes theorem, using a constant prior function.
Finally, the complete notation for the posterior marginals is given by:

Equation 2.13: $$p_i(t_i \mid x_1, \ldots, x_n) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} p(t_1, \ldots, t_n \mid x_1, \ldots, x_n)\ \mathrm{d}t_1 \ldots \mathrm{d}t_{i-1}\ \mathrm{d}t_{i+1}\ \mathrm{d}t_n$$

In general, the posterior marginals depend on the whole set of radiocarbon measurements, because the corresponding real sample ages can be linked together by the prior function.
The equations given in this section are the exact foundation of basic Bayesian multi-sample calibration, although, for the numerical realisation the equations in section 2.2.1 are sufficient.

A compact explanation of the formalism of the Bayesian multi-sample calibration can also be found e.g. in BUCK *et al.* (1991) or more detailed in BUCK *et al.* (1996) or in CHRISTEN (1994).

2.2.3 Remarks to the treatment of the uncertainty of the calibration curve

In this section a short consideration is given, why the uncertainty of the calibration curve can be treated in the way as done in the sections 2.2.1 and 2.2.2.
The unbiased single-sample likelihood function for a particular determined radiocarbon age x with uncertainty σ, would be, in its normalised form: (For simplification the sample index i is omitted generally in the following equations.)

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\ \cdot\ \exp\left(-\frac{\left(x-c^*\right)^2}{2\sigma^2}\right)$$

Where $c^*(t)$ is the unknown true value of the calibration curve at the real age t, unlike the given mean value $c(t)$ of the real existing calibration curve. For further simplification of the notation, the dependences on the real age t are not explicitly noted within the equations, however, all equations are valid for any value of t. Although, the value c^* is not known, its probability density is given by:

$$\frac{1}{\sqrt{2\pi}\cdot\sigma_c}\ \cdot\ \exp\left(-\frac{\left(c^*-c\right)^2}{2\sigma_c^2}\right),$$

where $\sigma_c(t)$ is the accuracy of the calibration curve. Both expressions are probability densities on the radiocarbon-age axis (given for any value of the real age t). The first is the probability density for the radiocarbon age x at a given possible true value of the calibration curve c^*. The second is the probability density for a possible true value of the calibration curve at a given mean value c with uncertainty σ_c. Thus, multiplying these two densities and integrating over all possible values of c^*, results in a probability density for the radiocarbon age, not depending on the unknown true value of the calibration curve. This is the likelihood function including the uncertainty of the calibration curve, one was looking for:

$$l = \frac{1}{2\pi\cdot\sigma\cdot\sigma_c}\cdot\int_{-\infty}^{+\infty}\exp\left(-\frac{\left(x-c^*\right)^2}{2\sigma^2}\right)\cdot\exp\left(-\frac{\left(c^*-c\right)^2}{2\sigma_c^2}\right)\mathrm{d}c^*$$

The product of the two Gaussians can be transformed elementarily as follows:

$$l = \frac{1}{2\pi \cdot \sigma \cdot \sigma_c} \cdot \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-c)^2}{2 \cdot (\sigma^2 + \sigma_c^2)} \right) \cdot \exp\left(-\frac{(c^* - x_{prod})^2}{2 \cdot \sigma_{prod}^2} \right) \mathrm{d}c^*$$

with $x_{prod} = \frac{c \cdot \sigma^2 + x \cdot \sigma_c^2}{\sigma^2 + \sigma_c^2}$ and $\sigma_{prod} = \frac{\sigma \cdot \sigma_c}{\sqrt{\sigma^2 + \sigma_c^2}}$

After this transformation the integration can be done easily, resulting in:

$$l = \frac{1}{\sqrt{2\pi \cdot (\sigma^2 + \sigma_c^2)}} \cdot \exp\left(-\frac{(x-c)^2}{2 \cdot (\sigma^2 + \sigma_c^2)} \right)$$

This equation shows, that the correct way to consider the uncertainty of the calibration curve, is to use the quadratic sum of the latter with the uncertainty of the radiocarbon age, as introduced in Equation 2.6 and Equation 2.10.
As already mentioned earlier, the non-normalised form of the single sample likelihood as given in Equation 2.6, where only the exponential-function term of the equation above is used (the normalisation term $1/\sqrt{\ldots}$ is skipped), has not exactly the same shape as the normalised form (when treated as function of the real age t at fixed radiocarbon age x). This is caused by the age dependence of σ_c in the skipped part. But this difference is usually negligible, because the uncertainty of the calibration curve does not change so much within the short region, relevant for a single sample likelihood, and furthermore σ^2 is frequently much lager than σ_c^2. Both facts make the skipped part approximately constant, at least for archaeological applications, which use the younger part of the calibration curve having only small uncertainties.

2.3 SOME REMARKS ON THE BAYES THEOREM

The Bayes theorem is given by Equation 2.12 above. The used notation reflects the common application, where an existing 'a priori' information on unknown quantities, here the archaeological knowledge about the real sample ages $a(t_1,...,t_n)$, is upgraded by new findings or measurements, here the determined radiocarbon ages $(x_1,...,x_n)$, carried by the likelihood function $l(x_1,...,x_n|t_1,...,t_n)$. The result is a new improved knowledge on the searched quantities, the real sample ages $(t_1,...,t_n)$, given by the posterior $p(t_1,...,t_n|x_1,...,x_n)$. This process can be repeated, if once again new findings are available. In this case the posterior becomes the new prior, and the calculation is repeated with a new likelihood due to the new measurements. Thereby, the new prior is strictly spoken already a conditional probability density depending on the former measurements, but this has no meaning for the formalism focusing on the new measurements.
Although, the given form of the Bayes theorem is most useful for the application, the notation shown below is better for a clear visibility of all relations of the involved probability densities. For a more compact writing the vector notation $\mathbf{t} = (t_1,...,t_n)$ and $\mathbf{x} = (x_1,...,x_n)$ will be used and multi-dimensional volume integrals are noted e.g. as $\int_{vol} ... \mathrm{d}\mathbf{t}$.

The notation of the Bayes theorem (Equation 2.12) is changed now by labelling the 'normalisation integral' explicitly with $v(\mathbf{x})$ (the letter v is taken from the word 'volume', with regard to the volume of $l \cdot a$ in the **t**-space), resulting in

$$p(\mathbf{t} \mid \mathbf{x}) \cdot v(\mathbf{x}) = l(\mathbf{x} \mid \mathbf{t}) \cdot a(\mathbf{t}) \qquad \text{with} \qquad v(\mathbf{x}) = \int_{\text{vol}} l(\mathbf{x} \mid \mathbf{t}) \cdot a(\mathbf{t}) \; d\mathbf{t}$$

Using this form, it is easy to see, that integrating the Bayes theorem in the radiocarbon-age space (**x**) leads to

$$\int_{\text{vol}} p(\mathbf{t} \mid \mathbf{x}) \cdot v(\mathbf{x}) \; d\mathbf{x} = a(\mathbf{t}) \; ,$$

because the integral of the right side of the theorem gives $a(\mathbf{t})$ due to the fact that $l(\mathbf{x} \mid \mathbf{t})$ is a probability density normalized in **x**, thus

$$\int_{\text{vol}} l(\mathbf{x} \mid \mathbf{t}) \; d\mathbf{x} = 1$$

for any point in the **t**-space.
The following complete summary of the Bayes theorem, including all relations and normalisations, shows the symmetric structure of the theorem:

Equation 2.14:

$$p(\mathbf{t} \mid \mathbf{x}) \cdot v(\mathbf{x}) = l(\mathbf{x} \mid \mathbf{t}) \cdot a(\mathbf{t})$$

Equations 2.15:

(a): $v(\mathbf{x}) = \int_{\text{vol}} l(\mathbf{x} \mid \mathbf{t}) \cdot a(\mathbf{t}) \; d\mathbf{t}$ (b): $a(\mathbf{t}) = \int_{\text{vol}} p(\mathbf{t} \mid \mathbf{x}) \cdot v(\mathbf{x}) \; d\mathbf{x}$

Equations 2.16:

(a): $1 = \int_{\text{vol}} l(\mathbf{x} \mid \mathbf{t}) \; d\mathbf{x}$ (b): $1 = \int_{\text{vol}} p(\mathbf{t} \mid \mathbf{x}) \; d\mathbf{t}$

(c): $1 = \int_{\text{vol}} a(\mathbf{t}) \; d\mathbf{t}$ (d): $1 = \int_{\text{vol}} v(\mathbf{x}) \; d\mathbf{x}$

Usually the likelihood l and the prior a are given and they are normalised as specified by part (a) and (c) of Equations 2.16 and v is defined by part (a) of Equations 2.15. In this case, the posterior p results from Equation 2.14, and the three remaining equations (the equations on the right-hand side) are deducible.
The detailed meaning of the probability densities can be understood clearest by assuming for the present, that the process of taking a fixed number of samples with a set of real ages **t** and determining the corresponding set of radiocarbon ages **x** could be repeated independently infinite times, which is the usual statistical view. In this case the densities have the following meanings:
$a(\mathbf{t})$ ... distribution of the different sets of real ages **t** of the collected sets of samples
$v(\mathbf{x})$ ... distribution of the determined sets of radiocarbon ages **x** of the collected sets of samples
$l(\mathbf{x} \mid \mathbf{t})$ ... conditional probability density to get a particular set of radiocarbon ages **x** at any given set of real ages **t**
$p(\mathbf{t} \mid \mathbf{x})$ ... conditional probability density for a particular set of real ages **t** at any determined set of radiocarbon ages **x**
It is important to remark, that the definitions above are idealisations, assuming an ideal match between the quantities within the Bayes theorem and the real facts. In the real world, the quantities can only be the best possible approximations of the reality. For example, the exact distribution of possible sets of real ages is not known,

however, one knows some conditions from the stratigraphy, determining the distribution partially.
Although the picture of infinite repetition of the whole process is powerful for fundamental understanding, it is not applicable in reality. A real excavation is unique, with a finite number of definite samples, taken from a finite number of existing samples too. So there is no distribution, but only one fixed set of real ages of the samples. Thus the prior function $a(\mathbf{t})$ has to be seen as the probability density for the occurrence of a particular set of real sample ages, before using the information from the measurements. The actual used prior function tries to approximate this unknown density, based on the 'a priori' information, e.g. from stratigraphic evidence. In a similar way, $v(\mathbf{x})$ is the probability density for the possible sets of radiocarbon ages, without knowing the real ages of the collected samples, but already their probability densities. Knowing the principle relation between real ages and radiocarbon ages, which is expressed by the likelihood function $l(\mathbf{x}|\mathbf{t})$, $a(\mathbf{t})$ and $v(\mathbf{x})$ are directly related via part (a) of Equations 2.15. So more realistically, the explanation of $a(\mathbf{t})$ and $v(\mathbf{x})$ should be summarised as:
$a(\mathbf{t})$ ... probability density for sets of real ages $\mathbf{t}$ only reflecting the 'a priori' information
$v(\mathbf{x})$ ... probability density for sets of radiocarbon ages $\mathbf{x}$ corresponding to the 'a priori' information only; therefore $v(\mathbf{x})$ is usually denoted as 'prior predictive distribution'
The explanations for $l(\mathbf{x}|\mathbf{t})$ and $p(\mathbf{t}|\mathbf{x})$ remain the same as given above.
It should be noted, that $l(\mathbf{x}|\mathbf{t})$ is normalised in the radiocarbon space and $p(\mathbf{t}|\mathbf{x})$ is normalised in the real age space; see Equations 2.16. In the application of this work, the likelihood function is treated as function in the $\mathbf{t}$-space at fixed $\mathbf{x}$ values as illustrated in Figure 2.2 for the two-dimensional example. It has to be kept in mind, that exactly spoken, this function $l(\mathbf{t})$ is not a normalised probability density. The likelihood function has to be normalised only in the radiocarbon space, if normalisation is required. This is as well the answer to the question that occurred earlier in section 2.2.2, why the normalisation of Equation 2.10 is correct, although the normalisation constant depends on the real ages. To get a probability density function normalised in the real age space as desired, realised by the posterior function $p(\mathbf{t}|\mathbf{x})$, it is necessary to apply the Bayes theorem, including the need of a prior function.
Comparing the exact formulation of the Bayes theorem, as given in this section, with the example given in section 2.1, a difficulty becomes evident: Equations 2.16 shows, that the prior has to be normalised, but this is not possible for a prior as used in the example, which is not restricted within the sample-age space. Normally this does not matter, because normalisation is not essential for the result of the method. A way to deal with unlimited prior functions, when normalisation is needed, is shown in section 4.3.1.
Finally it should be mentioned, that the notation of the four probability densities within the Bayes theorem with different letters (a, l, p, v) for clarification, is not common. Frequently all terms are equally denoted with p (for probability), and can be distinguished by the function arguments only. In this case the Bayes theorem would look like this:

$$p(\boldsymbol{\theta}\,|\,\mathbf{x})\cdot p(\mathbf{x}) \;=\; p(\mathbf{x}\,|\,\boldsymbol{\theta})\cdot p(\boldsymbol{\theta}) \qquad \text{with} \qquad p(\mathbf{x}) \;=\; \int_{\text{vol}} p(\mathbf{x}\,|\,\boldsymbol{\theta})\cdot p(\boldsymbol{\theta})\; d\boldsymbol{\theta}$$

In general, **x** is a set of stochastic variables, whose distribution depends on a set of parameters $\boldsymbol{\theta}$, according to the determined radiocarbon ages and the real sample ages in the application of this work.

Theoretical exact considerations to the Bayes theorem are given e.g. in the first chapter of JEFFREYS (1961). A very compact introduction of the theorem can be found e.g. in the first chapter of GILKS *et al.* (1996).

2.4 BASIC NUMERICAL REALISATION

The four equations given in section 2.2.1 are sufficient to perform Bayesian sequencing in its basic form. To perform the integration as expressed by Equation 2.9, the multi-dimensional posterior function can be numerically realised by an *n*-dimensional array, where *n* is the number of samples and each coordinate represents an individual sample age. This array can be calculated by an element-by-element multiplication of the likelihood function and the prior function, which can also be expressed as *n*-dimensional arrays. For clarification, remember the two-dimensional example from section 2.1, whose posterior function is once again plotted in Figure 2.7. Each crossing within the grid of the graph can be numerical realised by an element of a two-dimensional array, carrying the corresponding function value. Thus, the integration can be performed simply by summing up the array elements of the posterior function along all age coordinates that occur as integrands. Dealing with only very few samples, the calculations can actually be done in this direct way, as it was done for the example of section 2.1. However, usually many samples are included within Bayesian sequencing, causing a serious problem for this straightforward evaluation: Imagine one had to treat a set of e.g. 20 samples and would use a division of e.g. 100 points on each sample-age axes. This would lead to a 20-dimensional arrays for the likelihood, prior and posterior functions, containing now $100^{20} = 10^{40}$ entries, which cannot be handled any more. Fortunately, the Gibbs-sampling method, which belongs to the family of Markov chain Monte Carlo methods (MCMC), can solve this problem in a very convenient way.

2.4.1 Performing the calculations by Gibbs sampling

The basic principle when using Monte Carlo methods, is to perform the evaluations only at randomly chosen points in the multi-dimensional space, instead of evaluating all elements of an array as described above. The straightforward approach would be the use of uniformly distributed points. Unfortunately, this has the drawback, that a lot of sampled points could lie at positions with nearly zero function values, where they would hardly contribute to the result. This is a very serious problem in spaces with higher dimension. In contrast, Gibbs sampling is an efficient algorithm to find points in a way, that the density of their pattern is proportional to the value of an investigated multi-dimensional function, which has to have the properties of a probability distribution. Or in other words, Gibbs sampling draws randomly points out of a given probability distribution function. Therefore, only a few points will occur in regions with negligible function values, which makes the evaluation

efficient. The method requires only evaluations of one-dimensional cross sections through a multi-dimensional probability distribution, so that a high dimension of the distribution is no problem any more in principle.

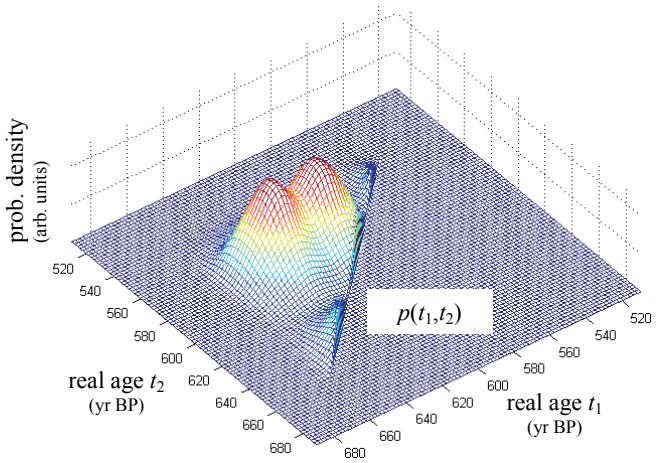

Figure 2.7: Two-dimensional posterior probability density to demonstrate the Gibbs sampling method as shown below.

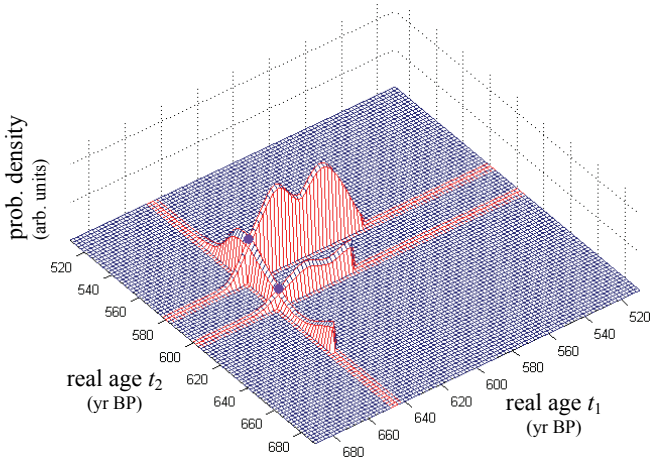

Figure 2.8: Gibbs-sampling of the distribution from above: Imagine one starts with a cross section along the first dimension at the arbitrary fixed position of t_2 (at 580 BP; the largest one in the picture). Out of this conditional probability distribution for t_1, a point is drawn randomly. At the drawn t_1-value (645 BP) a cross section along the second dimension is calculated, which is a conditional probability density for t_2, from which the next point is drawn, and so on.

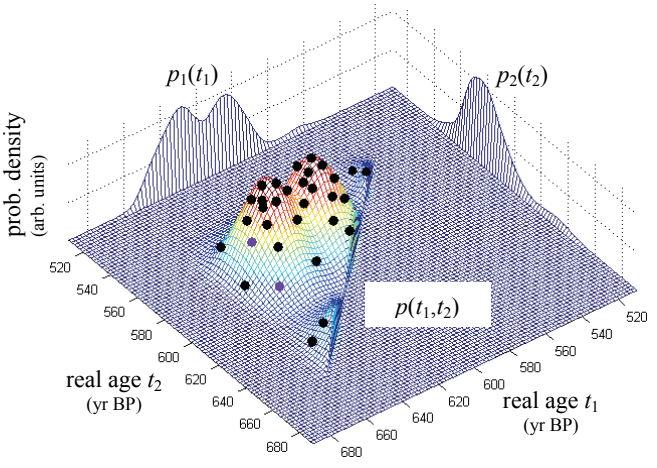

Figure 2.9: The repetition of the procedure from above would generate a point pattern as approximately indicated in the figure. For an increasing number of points, the density of the pattern becomes proportional to the function value. Histograms of the projections of the point positions to the axes would generate the posterior marginals.

The Gibbs-sampling algorithm is the following: One starts with calculating a cross section along an arbitrary dimension of the probability distribution, setting all other coordinates to starting values that have to be chosen in a way, that the initial cross section is not completely zero. Such a cross section is, aside from normalisation, the conditional probability for the investigated coordinate at the specific fixed values for

the other coordinates. Looking right into the posterior probability, a cross section along the i^{th} dimension can be mathematically expressed as $p(t_i|t_1,...,t_{i-1}, t_{i+1},...,t_n)$. Exactly spoken, the radiocarbon ages are conditions too, but it is not useful to indicate them in this context. Out of this conditional probability distribution of t_i, a position on the real-age axis of the i^{th} sample t_i^* is randomly drawn. After this, a further cross section is calculated along the next dimension $i' = i+1$, located at the position of the first draw. That means, $t_i=t_{i'-1}$ is set to the value t_i^*, i' is the new investigated age coordinate, all other ages remain the same. The new conditional probability distribution is $p(t_{i'}|t_1,...,t_{i'-1}, t_{i'+1},...,t_n)$. Again, a position is randomly drawn out of that distribution. This procedure is repeatedly performed, continuing at the first coordinate, when reaching the last. With each change of any coordinate a new point is found. It can be shown theoretically, that in general, the density of their pattern converges to the processed probability distribution. The mathematical foundation for this is roughly given in section 2.5.1.
Figure 2.8 illustrates the procedure of repeatedly calculating cross sections at the positions of the previously drawn point, using the posterior function of the two-dimensional example from section 2.1, given in Figure 2.7. The procedure results in a point pattern representing the probability density, approximately symbolised in Figure 2.9.
To find the marginal posterior densities of the individual sample ages by evaluating Equation 2.9, is very simple, when using Gibbs sampling. The integration is easily done by projecting every drawn point onto every sample age axis, and counting the frequency of their occurrence within a histogram. The frequency of projected points is actually proportional to the integral, caused by the fact, that the density of the points represents already the posterior probability distribution. Thus, the marginal posterior density can be calculated by using only one-dimensional cross sections trough the posterior function. Also for the likelihood and prior functions only these one-dimensional cross sections are needed, because the posterior cross sections can be calculated as product of the corresponding likelihood and prior cross sections.
Further information about Gibbs-sampling is given e.g. by GILKS *et al.* (1996), KRAUSE (1994) or LEVINE *et al.* (2005), and also in section 2.5 below.
It should be mentioned that the basics explained up to now, are also described in WENINGER *et al.* (2006), however in a shortened form and in slightly different notation.

2.4.2 Essential program structure to run Bayesian sequencing

The essential program structure for running the most basic form of Bayesian multi-sample calibration by the means of Gibbs-sampling can be formulated in a very compact way. A possible realisation is given below, using Matlab language. In the actual program package, which was developed for this thesis and will be described in section 2.7, the basic features are realised very similarly in principle.

First, a short explanation of the needed Matlab functions and syntax information is given:
The used variables can be either scalars or matrices, defined by their computation. The special case of a matrix with only one column and `m` rows, an `m x 1` matrix, is a possible representation for a vector, a 'column-vector'. An equivalent vector

representation is a matrix with only one row and `n` columns, a `1 x n` matrix or 'row-vector'. Simple operators as `*` or `/` or `^2` act in the matrix sense. For example if `A` is a row-vector and `B` a column-vector with the same number `z` of entries, `A*B` results in a scalar, the inner product, and `B*A` results in a `z x z` matrix, both according to the rules of matrix multiplication. To get element-by-element operations, the notations `.*` or `./` or `.^2` have to be used. Addition and subtraction is always notated with + and - only.
`M(i,j)` gives the element in row `i` and column `j` of the matrix `M`. `M(:,j)` extracts the j^{th} column of `M`, leading to an `m x 1` matrix or column-vector. `A(i)` is the i^{th} element of the row- or column-vector `A`. `M'` gives the transposed of matrix `M`; the transposed of a row-vector gives a column-vector.
The expression `(x>y)` results in one if `x` is larger than `y` and gives zero otherwise; for vectors or matrices this is done element by element.
Only the following few, built-in Matlab functions are needed:
`ones(m,n)` or `zeros(m,n)` create `m x n` matrices with all elements one or zero.
`size(M,1)` or `size(M,2)` give the number of rows or columns of the matrix `M`.
`exp(M)` gives the exponential function for each element of `M`.
`cumsum(V)` gives the cumulative sum of the vector V (a vector with same size as V)
`rand` creates a random number between zero and one
`find(V>=x)` lists (e.g.) all indices of elements of V that are larger or equal than x (given as vector with varying size)
`for i=1:n` defines a loop with `i` running from one to `n`.

Previous to executing the Gibbs-sampling procedure, in a first step a matrix `Li`, carrying within each column a single-sample likelihood function, is evaluated following Equation 2.6 in section 2.2.1; see the listing 'Essential program structure - part 1'. Hereby `Val` is a matrix that provides the radiocarbon ages of the samples in the first, and their one-sigma errors in the second column. `Cal` carries a relevant region of the calibration curve. The first column gives the real age scale, the second the corresponding radiocarbon age, and the third the one-sigma error of the calibration. It is reasonable to use generally BP-ages within the calculations. In this basic implementation, the range and resolution of the first column of `Cal` defines directly range and resolution of the calculation. `m` is the number of used points on the real age scale and `n` is the number of samples. For compact calculation of the `m x n` matrix `Li`, the auxiliary `m x n` matrices `X`, `x`, `C` and `c` are evaluated. `X` and `x` consist of `m` equal rows, each carrying the radiocarbon ages of the `n` samples or their errors respectively. `C` and `c` contain `n` equal columns, each carrying the radiocarbon ages or errors of the `m` used positions on the calibration curve.

Essential program structure - part 1: Calculation of the single-sample likelihood functions:

```
m = size (Cal, 1);
n = size (Val, 1);
X = ones (m,1) * Val (:,1)';
x = ones (m,1) * Val (:,2)';
C = Cal (:,2) * ones (1,n);
c = Cal (:,3) * ones (1,n);
Li = exp ( - (X-C).^2 ./ (2 * (x.^2+c.^2)) );
```

Having obtained the single-sample likelihood functions, the Gibbs sampling (see section 2.4.1) can now be implemented. 'Essential program structure - part 2' gives the corresponding program code, executing Equation 2.7 to Equation 2.9 of section 2.2.1 implicitly. The evaluated column-vectors `Lcut`, `Acut` and `Pcut` are the one-dimensional cross sections through the likelihood, prior, and posterior functions, at a position, indicated by the index-vector `ind`. The latter contains one element for each sample, carrying an index between `1` and `m`, corresponding to a specific real-age value of the associated sample. The relation between index and real age value is defined by the first column of `Cal`. In case of the likelihood, a cross section through the multi-dimensional function is, aside from normalisation, identical with the single-sample likelihood function of that sample, that corresponds to the dimension, along which the cross section is executed. The function `priorsection` provides a cross section through the application-specific prior and is described later. It has to be ensured, that the starting index `ind0` does not point to a position where the posterior probability is zero, or at least, that not the whole first posterior cross section is zero. The sampling runs a number of `cyc` full cycles or iterations, each stepping through all `n` dimensions. The adequate value of `cyc` is related to the question of convergence that will be discussed in section 2.5.2. `Pi` is the finally resulting matrix, each column representing an individual marginal posterior probability density of the corresponding sample. It results from the collected drawn points, projected onto the sample axes. For efficiency, each drawn point is projected only to the current dimension, which is equivalent to a projection to all dimensions, aside of a meaningless constant factor.

Essential program structure - part 2: Processing the Gibbs sampling:

```
ind = ind0;
Pi = zeros (m, n);

% Loops over iterations j and samples i:
for j = 1 : cyc
   for i = 1 : n

        % Evaluating the one-dimensional sections:
        Lcut = Li (:, i);
        Acut = priorsection (i, ind, Cal(:,1));
        Pcut = Lcut .* Acut;

        % Drawing a point from the distribution Pcut
        % and updating the index-vector:
        Int = cumsum (Pcut);
        I = find (Int >= rand * Int(m));
        im = I (1);
        ind (i) = im;

        % Evaluating the marginal posterior distributions
        % by collecting the point projections:
        Pi (im, i) = Pi (im, i) + 1;
   end
end
```

Finally, a simple implementation of the function `priorsection` used above is given in 'Essential program structure - part 3'. This function performs the transformation of the prior from its simple form used to define the prior in a convinent way (as shown in the last line of the code below), to a specific cross section through the multi-dimensional prior function, which is needed within the Gibbs sampling. The used prior example `(T(1)<T(2))*((T(2)<T(3))` would define that sample 1 has to be younger than 2, and sample 2 has to be younger than 3. `poin` defines the real age scale and is set equal to the real age column of the calibration curve by the code above. The vector `T` contains the real age values along the calculated cross section. All values but this of the investigated dimension `i` are fixed according to the index `ind`. Stepping the i^{th} element of `T` along the real age axis, the function values of the prior cross section `Acut` are calculated.
It should be mentioned, that the given implementation is the most simplest form. It is not optimised regarding run time, because the function `userprior` has to be evaluated for each single calculated point. In the actually used program, it is possible to evaluate the user-specified prior in a fast vector-related way, keeping the simple form of user input.

Essential program structure - part 3: Calculating a cross section through the prior function:

```
function Acut = priorsection (i, ind, poin);
m = size (poin, 1);
n = size (ind, 2);

% Setting the age values for the fixed dimensions:
T = zeros (1, n);
for j = 1 : n
   T (j) = poin (ind(j));
end

% Evaluating the prior function values along the current dimension:
Acut = zeros (m, 1);
for j = 1 : m
   T (i) = poin (j);
   Acut (j) = userprior (T);
end

% Sub-function with the specific prior function (example):
function A = userprior (T);
A = (T(1)<T(2)) * (T(2)<T(3));
```

Notwithstanding of its very compact form, the shown code gives a fully executable Matlab program for Bayesian sequencing in its basic form, just by putting part 1 to part 3 together in their given order. However, the complete developed program actually used (see section 2.7), which provides a package of features beyond the basic case illustrated above, and additionally elaborate input and output interfaces, is not of comparable size to shown code any more.

2.5 SOME GENERAL REMARKS ON THE GIBBS SAMPLING METHOD

2.5.1 Mathematical justification of the procedure

Here, the mathematical foundation is given that shows, that the Gibbs sampling procedure, as described in detail in section 2.4.1, is actually able to reproduce the underlying probability density.
We assume, that after J full iteration cycles (after $J \cdot n$ individual steps, where n is the number of samples) the method finds a point, based on a probability distribution $\pi^{(J)}$, which may differ from the actual probability density p. However, with given $\pi^{(J)}$ the distribution after the next iteration cycle $\pi^{(J+1)}$ can be calculated. For simplification the equation is given for the two-dimensional case first:

Equation 2.17:
$$\pi^{(J+1)}(t_1^*, t_2^*) = \int_{-\infty}^{+\infty} p(t_2^* \,|\, t_1^*) \cdot p(t_1^* \,|\, t_2) \cdot \pi^{(J)}(t_2)\ \mathrm{d}t_2$$

Where (t_1,t_2) is the point after J iteration cycles. (t_1^*,t_2^*) is the point after the $(J+1)^{\text{th}}$ cycle, which is drawn out of a cross section along the second dimension, after drawing a previous point out of a cross section along the first dimension. $\pi^{(J)}(t_2)$ is the probability density for the initial point, to lie within the latter cross section, whose position is denoted with t_2. $p(t_1^* \,|\, t_2)$ is the probability density to draw the next point at t_1^* out of this cross section, and $p(t_2^* \,|\, t_1^*)$ is the density to draw the final point at t_2^* out of the cross section along the second dimension located at t_1^*. See Figure 2.10 for clarification. By integrating over the t_2 coordinate of the initial point, or over all possible positions of the first cross section respectively, on gets the probability density $\pi^{(J+1)}(t_1^*,t_2^*)$ after the $(J+1)^{\text{th}}$ iteration cycle. Now, the three factors within the integral in Equation 2.17 can be expressed by the non-conditional two-dimensional representation of the probability densities:

Equation 2.18:
$$\pi^{(J+1)}(t_1^*, t_2^*) = \int_{-\infty}^{+\infty} \frac{p(t_1^*, t_2^*)}{\int_{-\infty}^{+\infty} p(t_1^*, t_2')\,\mathrm{d}t_2'} \cdot \frac{p(t_1^*, t_2)}{\int_{-\infty}^{+\infty} p(t_1', t_2)\,\mathrm{d}t_1'} \cdot \left(\int_{-\infty}^{+\infty} \pi^{(J)}(t_1, t_2)\,\mathrm{d}t_1\right)\ \mathrm{d}t_2$$

By setting $\pi^{(J)} = p$ the Equation 2.18 leads to

$$\pi^{(J+1)}(t_1^*, t_2^*) = \pi^{(J)}(t_1^*, t_2^*) = p(t_1^*, t_2^*)$$

and this means, that $\pi = p$ is actually the stationary distribution of the Gibbs sampling procedure.

In the general n-dimensional case the initial equation would look like this:

Equation 2.19:
$$\pi^{(J+1)}(t_1^*, \ldots, t_n^*) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p(t_n^* \,|\, t_1^*, \ldots, t_{n-1}^*) \cdot p(t_{n-1}^* \,|\, t_1^*, \ldots, t_{n-2}^*, t_n) \cdot \quad \ldots \quad \cdot$$
$$\cdot\, p(t_2^* \,|\, t_1^*, t_3 \ldots, t_n) \cdot p(t_1^* \,|\, t_2, \ldots, t_n) \cdot \pi^{(J)}(t_2, \ldots, t_n)\,\mathrm{d}t_2 \ldots \mathrm{d}t_n$$

Transforming the Equation 2.19 similar to Equation 2.17 above, leads again to the result, that $\pi = p$ is the stationary distribution.

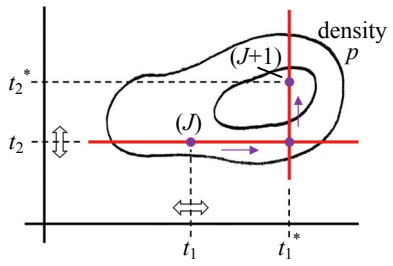

Figure 2.10: Illustration to Equation 2.17 and Equation 2.18. In the two-dimensional case the point (t_1^*, t_2^*) at the end of the $(J+1)^{th}$ iteration cycle is constructed by two sampling steps, based on the point (t_1,t_2) at the end of the previous cycle. One has to integrate over all possible positions for t_1 and t_2.

2.5.2 Convergence and 'burn in'

As shown above, when Gibbs sampling is executed on a probability density p, this density is in fact the stationary distribution for the sampled points. Unfortunately, this statement is not exactly true, because it was not shown, that there are no other stationary distributions possible. Actually, one can find easily an example for a distribution p that allows a stationary distribution π for the sampling process, which is not equal to p. Figure 2.11 shows this example for the two-dimensional case. If the original probability density is separated totally in two parts that lie along the diagonal within the coordinate system, each single part alone defines already a stationary distribution π. This can easily be seen by analysing Equation 2.18 with p and π as shown in Figure 2.11, which results in $\pi^{(J+1)} = \pi^{(J)}$ (π is input as $\pi^{(J)}$ into the equation). This result can also be seen aside from the equation, when reflecting the sampling procedure: Assuming the sampling starts in the upper right part of p, there is no way to reach the lower left part with cross sections along the axis, as used in the procedure.

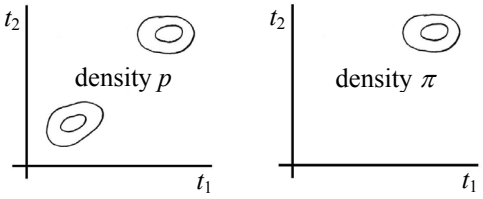

Figure 2.11: Example for a stationary distribution π that differs from the actual sampled distribution p. Distributions p that are split into two parts, which are totally unconnected may show this behaviour.

However, for the archaeological application in this work, probability densities with totally separated parts are unusual. For connected densities the Gibbs sampling will always produce a distribution equal to the original distribution p (there is a theoretical proof; see e.g. GILKS *et al.*, 1996). Although, it is obvious, that the number of points needed to realise the stationary distribution, or in other words the speed of convergence, will depend on the shape of the investigated distribution. For example, if the distribution is similar to p within Figure 2.11, except a tiny connection with low density between the two peaks, the convergence will be slow. This is because the sampling will stay for very many steps within one peak. Only with a small probability the sampling will be able to get to the other peak. Therefore,

the small number of changes from one peak to the other defines the quality of the statistics, rather than the much larger number of sampled points.
In fact, for this simple example the convergence problem could be solved easily by a 45°-rotation of the coordinate system. Furthermore, the basic Gibbs sampling procedure is just one very special method of the Markov chain Monte Carlo family (see section 2.5.3), offering methods with improved convergence. However, for the investigations of this work it was sufficient to use only basic Gibbs sampling within the developed computer program (described in sections 2.4.2 and 2.7). Nevertheless, it is necessary to check for sufficient convergence. What actually can be done, is to test, whether the distribution of the sampled points has already reached a stationary state. In the developed program this is performed by recording ten intermediate results during the sampling procedure. Two criteria are used, first the integrals over the absolute values of the differences between the marginal posterior functions, and second the shift of the centroid of the marginal posterior functions.

Closely related to the question of convergence is an often used procedure, called 'burn in'. The idea is to discard a number of initially sampled points, to give the sampler the chance, to find the stationary distribution first. This procedure can be useful, if the total number of sampled points is too small to represent the stationary distribution. So it does not matter if the starting point of the sampling is chosen at a position with very low probability that would hardly be reached within the total number of sampled points, because this initial points would be dropped. From a theoretical point of view, assuming a very large number of sampled points, there is no need of a burn-in procedure, because even regions with low probability will be reached within the sampling, and so there is no bias of the result, if the procedure starts at such a position. In the program developed for this work, the initial points are not automatically discarded.

2.5.3 A few words to Markov chain Monte Carlo methods in general

As mentioned above, the used Gibbs sampling is a special form of a Markov chain Monte Carlo (MCMC) method. The general principle to find points representing a given distribution $p(\boldsymbol{\theta})$ by a Markov chain is the following: A possible new point in the chain $\boldsymbol{\theta}^*=(\theta_1^*,...,\theta_n^*)$ is found first with the help of a conditional selection distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, where the condition $\boldsymbol{\theta}$ is the preceding point. The selection distribution could e.g. be an n-dimensional Gaussian around $\boldsymbol{\theta}$. A drawn point is accepted as an actual new point with a certain probability α, as described below. If the point is not accepted, the procedure continuous with $\boldsymbol{\theta}$ once more. The widely used 'Metropolis-Hastings algorithm' defines α in the following way:

$$\alpha = \min\left(1, \frac{p(\theta^*)}{p(\theta)} \cdot \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$$

For selection distributions with $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ equal to $q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ (e.g. the Gaussian mentioned above) α is reduced to

$$\alpha \quad = \quad \min\left(1,\ \frac{p(\theta^*)}{p(\theta)}\right),$$

and the method is then called 'Metropolis algorithm'.

A special kind of the Metropolis-Hastings algorithm is the single component Metropolis-Hastings algorithm. In this method, instead of generating the complete new point, component by component is generated by the algorithm. For each coordinate an individual selection function q_i can be defined, and an individual probability α_i to accept the recent component is calculated:

$$\alpha_i \quad = \quad \min\left(1,\ \frac{p(\theta_i^* \,|\, \boldsymbol{\theta}_{\text{except } i})}{p(\theta_i \,|\, \boldsymbol{\theta}_{\text{except } i})} \cdot \frac{q_i(\theta_i \,|\, \theta_i^*, \boldsymbol{\theta}_{\text{except } i})}{q_i(\theta_i^* \,|\, \theta_i, \boldsymbol{\theta}_{\text{except } i})}\right)$$

$\boldsymbol{\theta}_{\text{except } i}$ is a shortcut for $(\theta_1,...,\theta_{i-1},\theta_{i+1},...,\theta_n)$, where θ_{i-1} is the previous renewed component, and θ_{i+1} will be renewed next. Gibbs sampling is a further specialisation of the this algorithm: As the selection distribution can be defined widely free, one can define it as well as the cross section along the i^{th} coordinate through the original distribution (i.e. the original conditional distribution for θ_i at the position $\boldsymbol{\theta}_{\text{except } i}$):

$$q_i(\theta_i^* \,|\, \theta_i, \boldsymbol{\theta}_{\text{except } i}) \quad = \quad p(\theta_i^* \,|\, \boldsymbol{\theta}_{\text{except } i})$$

and respectively

$$q_i(\theta_i \,|\, \theta_i^*, \boldsymbol{\theta}_{\text{except } i}) \quad = \quad p(\theta_i \,|\, \boldsymbol{\theta}_{\text{except } i})$$

This leads directly to

$$\alpha_i \quad = \quad \min\left(1,\ \frac{p(\theta_i^* \,|\, \boldsymbol{\theta}_{\text{except } i})}{p(\theta_i \,|\, \boldsymbol{\theta}_{\text{except } i})} \cdot \frac{p(\theta_i \,|\, \boldsymbol{\theta}_{\text{except } i})}{p(\theta_i^* \,|\, \boldsymbol{\theta}_{\text{except } i})}\right) \quad = \quad 1\ ,$$

which means, that the drawn new point is always accepted. Thus, the algorithm draws points out of the one-dimensional cross sections along the individual coordinates and accepts this point always, which is actually the Gibbs sampling algorithm as described previously.
Further informatio about Markov chain Monte Carlo methods can be found e.g. in GILKS *et al.* (1996).

2.6 INTRODUCING STATISTICAL OBJECTS BEYOND SAMPLE AGES

2.6.1 The mathematical framework

In all explanations given up to here, only the basic application of Bayesian statistics has been described. A set of n statistical parameters $(t_1,...,t_n)$ representing the real ages of n samples was used exclusively. The results were probability distributions for the real sample ages only. However, in most applications probability distributions of additional parameters are of interest. For example, in archaeological sequences one commonly wants to know the probability distributions of phase boundaries, defined to be younger than the samples of a particular phase and older than the samples of

the following phase. Or in some situations depth-age models can be used, and one wants to get the probability distributions of the parameters defining the depth age relation (e.g. deposition rate). Situations of this kind can be handled in a general form by introducing an application dependent model function f_{model} with a free number p of model parameters (s_1,...,s_p), previous to the Bayesian framework, as described in section 2.2.1.:

Equation 2.20: $$(t_1, \ldots, t_n) = f_{model}(s_1, \ldots, s_p)$$

In the numerical realisation, f_{model} is a user-definable function. The particular relation of the real ages (t_1,...,t_n) and the model parameters will be clarified by examples in the following two sections. The equations for the single-sample likelihood function l_i (Equation 2.21), the multi-dimensional likelihood function l (Equation 2.22) and posterior probability distribution p (Equation 2.23) remain in principle the same as shown in section 2.2.1, but now, the real sample ages are themselves deduced from the model parameters.

Equation 2.21: $$l_i(t_i(s_1, \ldots, s_p)) \propto \exp\left(-\frac{\left(x_i - c(t_i(s_1, \ldots, s_p))\right)^2}{2 \cdot \left(\sigma_i^2 + \sigma_c^2(t_i(s_1, \ldots, s_p))\right)}\right)$$

Where t_i is the unknown real age for the i^{th} sample, x_i the determined radiocarbon age and σ_i its uncertainty. $\sigma_c(t)$ is the uncertainty of the radiocarbon-age value of the calibration curve $c(t)$ on the real-age position t.

Equation 2.22: $$l(s_1, \ldots, s_p) \propto l_1(t_1(s_1, \ldots, s_p)) \cdot \ldots \cdot l_n(t_n(s_1, \ldots, s_p))$$

Equation 2.23: $$p(s_1, \ldots, s_p) \propto l(s_1, \ldots, s_p) \cdot a(s_1, \ldots, s_p)$$

Prior function, likelihood function and posterior function are now defined within the p-dimensional model-parameter space, instead of the n-dimensional real-age space, as they have been before. Consequently, also the Gibbs sampling is performed in this p-dimensional space. Of course, when using this parameter algorithm, probabilities of simple sample ages can still be calculated, just by setting this ages directly equal to parameters when creating the model function.
At this point one can evaluate the marginal posterior probabilities of the parameters analogous to Equation 2.9 in section 2.2.1., leading to:

Equation 2.24: $$p_i(s_i) \propto \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} p(s_1, \ldots, s_p)\ \mathrm{d}s_1 \ldots \mathrm{d}s_{i-1}\ \mathrm{d}s_{i+1} \ldots \mathrm{d}s_p$$

However, it is useful to make a second generalisation here. The reason therefore is, that frequently one is interested in probability distributions of functions based on the parameters or sample ages, as e.g. a probability distribution for the time difference between two particular ages. Therefore a free number q of user-definable result quantities (r_1,...,r_q), described by the function g_{result}, is introduced:

Equation 2.25: $$(r_1, \ldots, r_q) = g_{result}(s_1, \ldots, s_p,\ t_1(s_1, \ldots, s_p), \ldots, t_n(s_1, \ldots, s_p))$$

The notation reflects the fact, that for convenience, g_{result} can be defined within the developed program on the model parameters and on the sample ages as well, although the sample ages are themselves functions on the parameters. Finally, one is interested in the probability distributions of the individual result quantities, expressed by Equation 2.26. These distribution have now the meaning of the marginal posterior distributions in the basic formalism, and will therefore be denoted analogous with $p^{(r)}{}_j(r_j)$.

Equation 2.26: $$p_j^{(r)}(r_j) \propto \int \ldots \int_{\substack{\text{on the domain where} \\ r_j - \delta r_j < r_j'(s_1', \ldots, s_p') < r_j + \delta r_j \\ \text{with constant } \delta r_j \to 0}} p(s_1', \ldots, s_p')\ \mathrm{d}s_1' \ldots \mathrm{d}s_p'$$

Unfortunately, in contrast to the calculation of the posterior marginals, the analytic expression for $p^{(r)}{}_j(r_j)$ has got a little cumbersome. However, the numerical evaluation performed by Gibbs sampling remains simple: As the sampling generates points with a density reflecting $p(s_1,...,s_p)$, just the result quantities $r_j(s_1,...,s_p)$ have to be calculated for each point and collected in individual histograms. These results (aside of normalisation) directly in the required probability densities for the r_j, because the total number of points collected for a particular value of the result function is proportional its total probability.

2.6.2 Important application: realisation of phase boundaries

In archaeological applications phase boundaries are the most prominent example for the need of parameters beyond sample ages. A phase defined within a model could, for example, be a continuous time of settlement in a village between a preceding and a subsequent rebuilding of the village. In many cases the available samples can be assigned to the different phases, but there is no information about time relations between the samples of the same phase. So, the model defines a sequence of phase boundaries with the sample ages lying between them.

The following artificial example assumes a phase containing three samples (number 2, 3 and 4) that is separated from a preceding phase containing sample 1 and a subsequent phase containing sample 5. This results in a model with seven parameters, where (e.g.) the first five parameters represent the sample ages, the parameters s_6 represents the ages of the younger and s_7 the age of the older boundary:

model (f_{model}):

$t_1 = s_1$
$t_2 = s_2$
$t_3 = s_3$
$t_4 = s_4$
$t_5 = s_5$

prior (a):

$$a = \begin{cases} 1 & \text{if } s_1 > \underline{s_6} > \{s_2, s_3, s_4\} > \underline{s_7} > s_5 \\ 0 & \text{else} \end{cases}$$

The prior function connects the boundaries s_6 and s_7 with the sample ages s_1 to s_5. The Matlab expression for this prior function, that has to be input into the program, is (see chapter 2.4.2 for syntax explanations and used notation):

```
prior = (S(1)>S(6)) * (S(2)<S(6)) * (S(3)<S(6)) * (S(4)<S(6)) * ...
                      (S(2)>S(7)) * (S(3)>S(7)) * (S(4)>S(7)) * (S(5)<S(7));
```

As mentioned already in section 2.1, it has to be noted here again, that this kind of realisation of the prior function is not the only possible form. The same is true for the next example in section 2.6.3. Although this point is very essential within this work, it will be discussed later (chapter 3).
For the present example the calculation was performed assuming the following radiocarbon ages:

sample	radiocarbon age (yr BP)
1	2155 ± 40
2	2140 ± 30
3	2115 ± 35
4	2060 ± 40
5	2020 ± 45

Figure 2.12 shows the single sample calibrations (likelihood functions) together with the resulting marginal posteriors for both, the sample ages and boundaries, given by Equation 2.24.

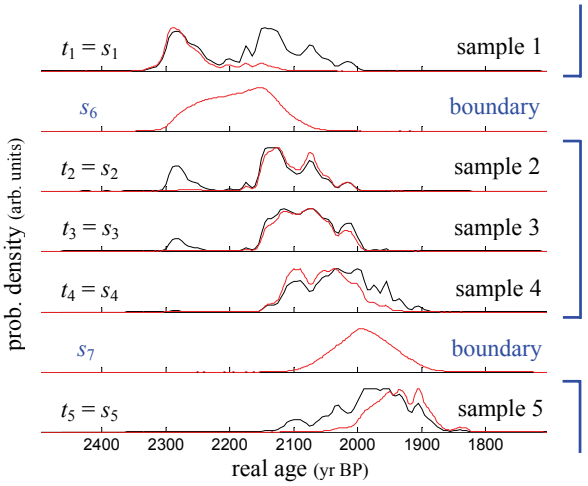

Figure 2.12: Realisation of phase boundaries: The black lines show the single-sample likelihood functions or single sample calibrations for the five samples. The red lines represent the calculated marginal posterior probabilities for the model parameters. Parameters 1 to 5 represent the sample ages; parameter 6 and 7 the two boundaries. The blue brackets indicate the phases separated by the boundaries.

It should be shortly noted here that in the developed program code (see section 2.7) all resulting densities have to be defined as result quantities. To get the results shown in Figure 2.12, one defines the result quantities $r_1,...,r_7$ equal to $s_1,...s_7$. For result quantities that are equivalent to model parameters themselves, there is no difference whether using Equation 2.26 or Equation 2.24.
Now, one could additionally ask for the probability density function of the duration of the phase containing sample 2, 3 and 4, which is the age difference of the two boundaries. For this, one easily defines an additional result quantity $r_8=s_6-s_7$. The

resulting probability density $p^{(r)}{}_8(r_8)$ according to Equation 2.26 is already the density for the phase length, which is shown in Figure 2.13.

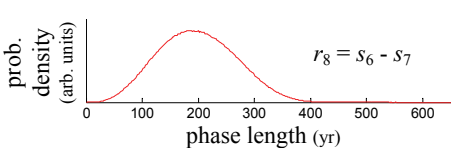

Figure 2.13: Probability density distribution $p^{(r)}{}_8(r_8)$ for the age difference $r_8=s_6-s_7$ of the two boundaries. This is the probability density for the length of the phase that contains the samples 2, 3 and 4.

A further way to use result quantities, is to calculate discrete probability values, as e.g. the probability, that sample 4 is the youngest within the phase. Therefore, one defines a result quantity r_9 that is one if sample 4 is actually younger than sample 2 and 3, and zero otherwise. The corresponding mathematical realisation according to Equation 2.25 can be expressed in Matlab language as:

```
R(9) = (T(2)>T(4)) * (T(3)>S(4))
```

Hence the number of sampled counts for $r_9=1$, compared with the total number of counts, gives the probability asked for. For this example, the probability that sample 4 is the youngest within the phase, results in 59.5%.

2.6.3 A more general usage of parameters: an accumulation rate model

Although the realisation of phase boundaries is essential in archaeological applications, it illustrates only a restricted way of the usage of parameters, because the parameters are still objects on the age scale, as the sample ages are. In this - also artificial - example, the accumulation rate of a deposit, containing dateable samples, shall be modelled. The use of accumulation rates is not very frequent in archaeology, because the accumulation of archaeological sites is commonly very inhomogeneous. On the other hand, one can think of geological sedimentation processes that can be very homogeneous.
For this example we assume, three samples that have been found at defined heights within a sedimentation layer:

sample	height (m)	radiocarbon age (yr BP)
1	6.1	15820 ± 40
2	3.7	16250 ± 50
3	1.9	16450 ± 45

Assuming a constant accumulation rate leads to the following model:

model (f_{model}):

$t_1 = s_2 - 6.1\text{m} \cdot s_1$
$t_2 = s_2 - 3.7\text{m} \cdot s_1$
$t_3 = s_2 - 1.9\text{m} \cdot s_1$

prior (a):

$a = 1$

The parameter s_1 is the duration per height growth or the reciprocal accumulation rate with dimension yr/m, and parameter s_2 is the real age at the bottom of the

sediment layer in kyr BP. As the model assumption of a constant growth rate is already realised by the specific definition of the model parameters, there is no available information left for the prior function, which is set to constant one. (Again, the choice of the particular used prior function is not the only possible one; see the remark at the former example.)

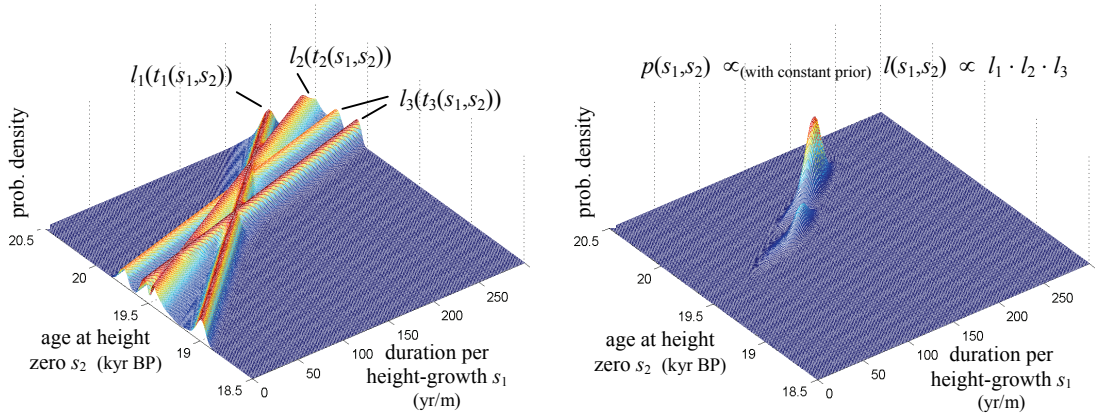

Figure 2.14: *Left side:* The single-sample likelihood functions of the three samples; each a two-dimensional function within the parameter space. *Right side*: The combined likelihood function, which is the product of the three functions shown in the right plot. The likelihood function is the posterior function as well, because the prior was chosen constant.

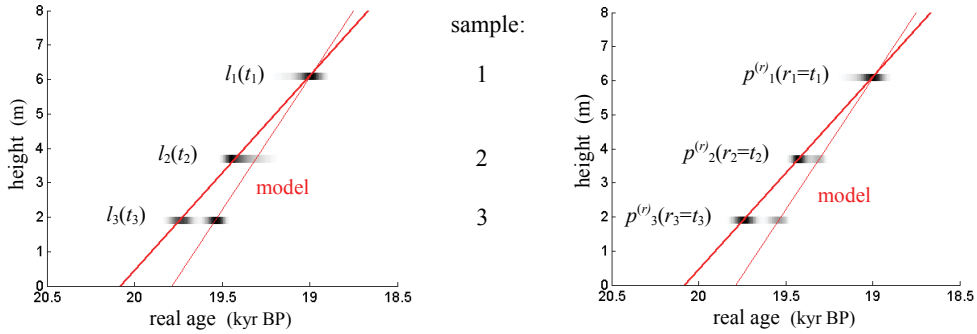

Figure 2.15: Comparison of the single sample calibrations (grey scaled in the left plot) and also the posterior probabilities of the sample ages (grey scaled in the right plot) with the two most likely height age relations (red lines), according to the two peak maxima of the posterior function. The thick line represents the major, the thin line the minor peak of the posterior function.

According to Equation 2.21, the single sample likelihoods are functions on the two-dimensional parameter space. They are shown together within a single plot; the left one in Figure 2.14. The shape of the single-sample likelihood functions reflect the fact, that a particular measured radiocarbon age for a sample at a fixed height in the layer, can be explained by various combinations of the age at the bottom of the layer and the value for the duration per height growth. There is a linear relation with a slope that depends on the sample height. The cross sections along s_2 at s_1=0 shows the common single sample calibrations. The double peaked structure for sample three originates from a wiggle in the calibration curve. The plot at the right side within Figure 2.14 gives the combined likelihood function, which is the product of the three single sample likelihoods. As the prior function is constant this function is already

the posterior function. The two resulting peaks reflect two different combinations of age at bottom and growth rate that are in agreement with the data; one more likely than the other. The height-age relations defined by the maxima of these two peaks are given by the red lines in Figure 2.15, together with the single sample calibrations in the left plot, and together with the posterior probabilities of the sample ages in the right plot. As the Gibbs sampling runs on the two-dimensional parameter space, the latter are calculated by defining result quantities that are set equal to the sample ages t_1,t_2,t_3. The height-age relation characterised by the thick red line is more likely than that characterised by the thin one. This is the reason for the significant difference between the single sample calibration and the posterior density, especially for sample three, which can be seen more detailed in Figure 2.16.

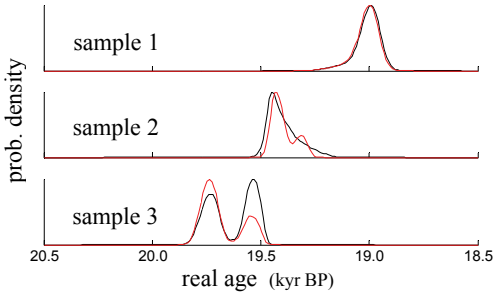

Figure 2.16: Single sample calibrations of the three samples (black), compared with their modelled posterior probability densities (red).

Conclusively it should be mentioned that there are descriptions and applications of deposition models published; see for example the papers of BLAAUW *et al.* (2007) and BRONK RAMSEY (2008).

2.7 A BRIEF DESCRIPTION OF THE DEVELOPED PROGRAM CODE

The mathematical and numerical foundations to perform Bayesian multi-sample calibration have been already described above. In this section a view words to the most important features and to the basic architecture of the developed program shall be made. There were two main reasons to do a self-made program, instead of using available packages: First to understand the method accurately from the basics, and second to have every freedom, to test modifications to improve the procedure.
The used programming language is Matlab (by The MathWorks, Inc., Natick, Massachusetts, USA), which is very comfortable for mathematical topics, because the syntax is compact and very near to mathematical notations. As there is no explicit compilation necessary, new program parts can be tested easily during the development process, which simplifies the program development significantly. The drawbacks of using this specific mathematical language are a higher runtime, and the fact, that the code cannot be executed independently from the Matlab program package.

All developed program features are controlled by a common user interface, based on a set of input and output files. First the user runs an initial program that creates an initial input file for specifying the different program modes, see Example 1:

Example 1: Initial input file. The actual input expressions are marked in blue, the user defined values in red:

```
% Input form  B01_FORM__CONTROL :   PROGRAM FLOW SETTINGS

% Here you choose the program flow by selecting different alternatives. Start the program B01_PREPARE
% afterwards and, depending on your selection, you will get the needed input forms for the main program
% B01_RUN.

% BUILT-IN EXAMPLES:
% To become familiar with the input syntax you can use built in examples. By calling an example this and
% all other needed input forms will be filled in automatically when running B01_PREPARE. Of course you
% can change the forms manually again.

% Call an example by number (set F_CON_stand):
%    0 :   Do not call an example, fill in or change the form manually
%   51 :   Realistic archaeological example
%   52 :   Artificial example demonstrating principles of the method (2-dim)
%   53 :   Artificial example demonstrating principles of the method (3-dim)
%   54 :   Artificial example working with a parameter model
%   55 :   Artificial example for robust Bayesian analysis
     F_CON_stand = 0;
%   Attention: If you call an example, already done entries in this and all other input forms will be
%   deleted !

% CALCULATION:
% To run calculations set F_CON_calculate to 'yes', otherwise to 'no':  (If the calculation is already
% done and you want to change the additional evaluations or the graphics only, take 'no')
     F_CON_calculate = 'yes';

%    Choose calculation mode and corresponding calculation parameter input:
%         1 :   common Gibbs-sampling mode
%         8 :   non-Gibbs direct calculation (usage of models or robust analysis is not possible in
%               this mode)
     F_CON_param = 1;

%    Choose the mode of prior information input:
%         1 :   direct input of the prior as a mathematical function
%         2 :   simplified input for common time sequencing situations
%        11 :   functional prior input for robust Bayesian analysis
     F_CON_prior = 1;

%    To use a parameter model or user defined result functions
%         set F_CON_model to 'yes', otherwise to 'no':
     F_CON_model = 'yes';

% ADDITIONAL EVALUATIONS (essential for robust analysis):
% To run additional evaluations (calculation of generalised probability ranges;
% resulting unified ranges for robust analysis) set
% F_CON_evaluate to 'yes', otherwise to 'no':
     F_CON_evaluate = 'no';

% CREATION OF GRAPHICS:
% To create graphics set F_CON_graphics to 'yes', otherwise to 'no':
     F_CON_graphics = 'yes';
```

Dependent on the chosen procedures, a further short setup program creates a set of four to eight additional input files, which are adapted to the users choice, to request the specifically needed data input. To make the use of the program more convenient, many input files provide the choice of different standard inputs. For example the calibration curve can be input manually into the corresponding file for sure, however, one can also call the IntCal-curve to be input automatically. Further, all needed inputs can be filled with data of a couple of different examples. This was very useful to test the program during development, and is still convenient to test program upgrades. Additionally, these examples can be used to get quickly consistent input data that do not generate runtime errors, by starting from an appropriate example, and changing the data subsequently. This is useful, since the program was designed to develop and analyse new methods; there was not paid to much attention on testing against inconsistent or wrong formatted inputs.
The resulting data and additional information (e.g. convergence indicators) are provided within a set of output files, and various plots are created, some of them in different definable modes.

In the following the essential features oft the program will be listed roughly. First, it should be mentioned, that the program offers aside from the common Gibbs-

sampling mode, a non-Gibbs direct calculation (see Example 1, third input). In this mode the multi-dimensional functions are calculated completely point by point on a grid of chosen resolution. For sure, this mode is only useable for low-dimensional problems, however, it is very useful to visualise fundamental properties of the method. The plots shown in section 2.1 are generated with this mode.
To define the prior function, there are two different ways available: Aside from the functional input, as used within the example in section 2.6.2, there is also the possibility to define the prior in matrix form, if younger/older relations between the sample ages or model parameters are used exclusively.
If a parameter model is used, two additional input forms are offered to define the relations of the real sample ages or parameters on the one hand, and the set of the desired result functions on the other hand; see the explanations in section 2.6.1. It should be additionally noted, that parameter models can be executed also completely without radiocarbon measurements, which can e.g. be used to analyse a prior function. The program further permits the opportunity to define multivalued (vectorial) result functions, which allow to associate a complete probability density distribution (rather than an individual value) with each sampled point, and add these densities up. This feature can be used to define for example a constant probability for all age values lying between the sampled ages of two different samples, and add this densities up for all sampled points. The result is a probability density for an event that is know to lie between these two sample ages. This procedure could be an alternative to using a boundary between the two ages, because it does not influence the marginal posterior functions for the ages themselves, as boundaries do.
The essential part of the parameter-definition form is shown as Example 2:

Example 2: Essential part of the file for model definition. The values are taken from the accumulation-rate example of section 2.6.3.:

```
% Range limits for the particular parameters:
%    1st column: lower limits
%    2nd column: upper limits
     F_MODPAR_grenzen = [18500  20500
                            0    300];


% Define the model function in the following way:
%    ( T(1),T(2),T(3),...) = free function of ( S(1),S(2),...);
%    S(1), S(2), ... are the model parameters;
%    T(1), T(2), ... are the true sample ages (yr BP);
%    intermediate steps with additional variables are permitted; the normal Matlab syntax can be used,
%    but EACH LINE HAS TO START WITH %$ SIGNS; do not change the special signs for the first and the
%    last line;

%$ $BEG:F_MODPAR_modfunc$
%$ T(1) = S(1) - 6.1*S(2);
%$ T(2) = S(1) - 3.7*S(2);
%$ T(3) = S(1) - 1.9*S(2);
%$ $END:F_MODPAR_modfunc$
```

On default, the results of the program are the marginal posterior functions or the densities of the result functions. An additional evaluation is provided that calculates highest-posterior-density ranges (defined in section 5.2.1) and single sample based agreement indices (defined in section 4.1). The ranges are simultaneously calculated for any confidence level on a continuous scale from zero to hundred percent.
There are features implemented to build a sum function of the single-sample likelihood functions for a definable group of samples, and alternatively, functions that are the weighted sums of likelihood functions, weighted by the samples probability to be the oldest/youngest of the defined group. These features where used for tests concerning the problem of the 'statistical pressure', which is discussed in

section 3.3. Further there is an 'overlap method' implemented that extends the multi-dimensional posterior function, depending on specific relations between the likelihood function and the prior constraints. This method will be explained in section 5.5.
The program offers the option to calculate the multi-dimensional volumes of the posterior function, the prior function and the likelihood function. This is needed to deduce a fundamental characterisation of the agreement of a used model (i.e. prior) with the measured radiocarbon ages, as explained in section 4.2. A Gibbs-sampling based multi-dimensional integration method was developed for this purpose, which is discussed in section 4.4.
An important feature is to perform 'robust Bayesian analysis', which can be done by varying the shape of the prior function, and unify the resulting highest-posterior-density ranges, as explained in chapter 5. For this purpose the program provides a prior-input form that enables the definition of parametric prior functions. There are two fundamentally different modes to unify the resulting highest-posterior-density ranges possible: A mode that allows to define a weighting function based on different available agreement parameters, and a second mode that discards results below a threshold based on the agreement parameters.

The program is structured as a hierarchically ordered set of specific program modules, that are all controlled by a common main program. However, essential parts of the program are designed in a way, that they can be executed on their own too. Aside of the modules with the actual program code, there are also sets of files carrying the patterns for the input and output files and the available sets of standard input values. During the run, the program creates also additional files that contain functions in the needed internal form, carrying the user defined prior and model information. This internal functions are saved, together with essential variables, in a reserved sub-directory within the user defined application directory. This enables the run of more than one instance of the program simultaneously, and further, to repeat the additional evaluations and the graphics creation, without repeating the sampling process.
Presently the program consist of about 150 individual files, containing all in all about 10 000 text lines. However, as the code contains a lot of comments and additionally parts with a high level of redundance, because some slightly different features are realised by code copies with adequate changes, the actual informative core of the code is represented by a few thousand lines.
The total runtime of the program depends highly on the processed application. The more of the following characteristics are present, the longer is the needed runtime: high sample number or parameter number, using a parameter model instead of calculating with samples only, bad convergence, using special methods as robust analysis with a high number of individual prior shapes. Thus, although a simple low-dimensional case can be done in a few minutes, robust analysis of complex applications, with e.g. fifty dimensions, can need a few days to give sufficiently convergent results, when running the program on a standard personal computer.

The complete program code is available on the web site

http://homepage.univie.ac.at/franz.weninger/radiocarbon-sequencing.html

to be viewed, tested or used. However, it should be remembered that the code is not optimised for save and convenient use.

Finally in this chapter, where the principles of Bayesian sequencing had been explained, it should be mentioned that since Caitlin Buck established this Bayesian statistical method in the field of archaeology (BUCK *et al.*, 1991), a large number of applications had been published using this powerful new method. Just a few articles shall be cited representatively: MANNING *et al.* (2006) is an example for a sequence based on archaeological evidences indicating the temporal order of the samples. Applications of this kind are mainly focused on within this thesis. GALIMBERTI *et al.* (2004) and BRONK RAMSEY *et al.* (2001*b*) show a frequent application too, which is called wiggle matching and models the well known age differences, as those between different rings of an individual piece of wood. Finally BRONK RAMSEY *et al.* (2010) is an example for a model based on information from historical records, specifically the reign-length records of Egyptian kings.

3 THE PRIOR FUNCTION: PROPERTIES AND PROBLEMS

This chapter is exclusively dedicated to the prior function, since it is the most crucial part in the Bayesian sequencing method. The reason therefore is the ambiguity when defining the prior function (discussed in section 3.1), what is the central issue that motivated this thesis. The analysis of the 'prior marginals' (section 3.2) is a powerful way to uncover an unwanted information content of the prior function, which has been not obvious when defining the prior function. By the help of these analyses some commonly used specific types of prior functions can be justified mathematically (section 3.3). Although, this thesis turns away from using only one single prior function (see section 5), a very general procedure to find the 'best possible' prior function shape, the 'maximum entropy method' is described finally (section 3.4).
Chapter 3 describes still common knowledge, however, the given analyses were important to get the necessary conception about the fundamental properties of priors.

3.1 THE SUBJECTIVITY OF THE USED PRIOR FUNCTION SHAPE

Subjectivity of the choice of the prior function is one of the most discussed problems in Bayesian statistics (see for continuative information e.g. KASS and WASSERMAN, 1996), and it is the main topic in this work too. Looking closer at this problem, it can be split into two different aspects. First, it is never possible to deduce the archaeological constraints from a site with absolute certainty. There will always remain a risk, that the historical facts deviate from the evidence obtained when analysing the site. So, one always deals with an archaeological model, that is only an approximation of the historical reality. However, some sites can be undisturbed and well preserved, so that the experienced archaeologist is able to deduce facts with high reliability. Unfortunately, even assuming that the archaeologists are able to define a final set of given constrains, it is not possible to transform this fixed prior information to a mathematical function - the prior function - in an unambiguous way. This part of the problem is possibly more serious than the first one, because it remains even when the archaeological information is well defined, as demonstrated at the following basic example. (Remember, that the simple notation 'age' is used always for real ages and not for radiocarbon ages.)
The knowledge, that sample 1 is older than sample 2 can be described correctly by both of the following prior functions, shown in Figure 3.1. One knows, that age 1 must not be younger than age 2, so the prior function has to be zero on the right side of the diagonal. But any function shape one chooses on the positive left side, defines a particular probability density for each combination of the two ages. Notwithstanding the fact, that this is not at all founded on available information, because one knows only that sample 1 is older than 2, and nothing more. Although, the procedure of Bayesian sequencing, as described in chapter 2, requires the use of a particular shaped prior function. That means, that Bayesian modelling has to assume

information that is actually without foundation. Of course, this assumption affects the resulting posterior density as well, and thus, brings in subjectivity into the method. How to overcome this problem will be the topic of chapter 5.

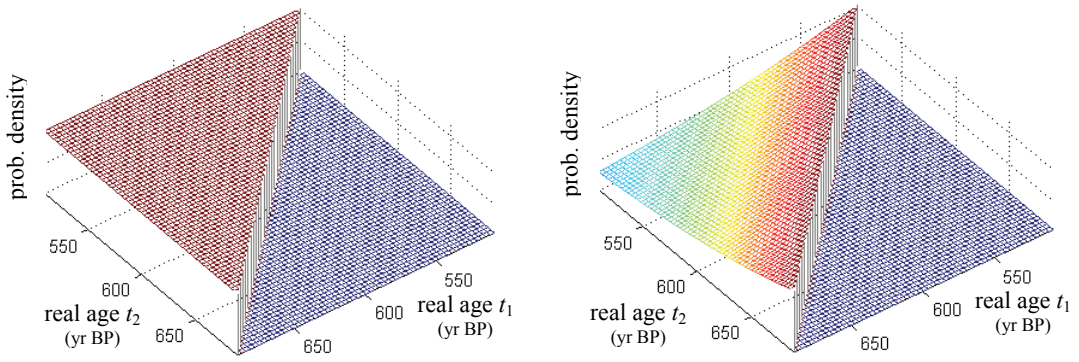

Figure 3.1: Two different functional representations of the prior information 'age 1 is older than age 2'. The shape of the part of the prior function with correctly ordered ages is ambiguous.

However, as shortly mentioned earlier, when looking at the two functions of Figure 3.1, one could get the impression, that the left function with constant probability in the region of possible age pairs - which is usually called a 'uniform prior' - is the only reasonable representation of the given fact, that age 1 has to be older than age 2. Actually it is not, and this fact has two different reasons: First, the definition that an age difference of e.g. fifty years is as likely as one of hundred years (left graph) is not less artificial than, that an age difference of fifty years is more probably than hundred years (right graph). Both definitions are actually not based on available information. And even if one can not accept this argument, one should be aware, that the uniform prior does not remain uniform, if one changes the scaling of the axes. For sure, it is not easy to argument why one should use something different than a linear age scale. However, thinking more generally, there is not always a 'natural' scale for the model parameters, as easily to see, when remembering the accumulation rate example from section 2.6.3. There, the reciprocal accumulation rate and the bottom age of the layer were chosen as model parameters, and a uniform prior was used. It was not discussed on purpose, that one could take directly the accumulation rate instead of the reciprocal as well, which defines, together with the unchanged bottom-age parameter, a new parameter set. The uniform prior for the original set can not remain uniform for the new set, without changing the result. (The correct transformation of the prior probability density would have to be in accordance with $a(\mathbf{s})\cdot d\mathbf{s} = a'(\mathbf{s}')\cdot d\mathbf{s}'$; where a is the original and a' the transformed prior, which are functions of the two different sets of parameters $\mathbf{s}$ and $\mathbf{s}'$.) As both parameterisations are reliable, this example shows clearly the inevitable ambiguity within the choice of the prior function, because not even the uniform prior has an exceptional position.

3.2 THE MEANING OF PRIOR MARGINALS

The former section shows, that the choice of the prior function is ambiguous and brings in subjectivity in Bayesian statistics. Therefore, the choice of the prior function is a critical step, and the finally used prior function should be analysed very

carefully. The prior can become a highly complex function, assuming a large number of archaeological constraints within a high-dimensional model. This function can hardly be characterised as a whole. Fortunately, prior marginals can be built in the same way as described for the posterior function in chapter 2. According to Equation 2.9, the prior marginal for a particular age is:

Equation 3.1:
$$a_i(t_i) \propto \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} a(t_1, \dots, t_n)\ \mathrm{d}t_1 \dots \mathrm{d}t_{i-1}\, \mathrm{d}t_{i+1} \dots \mathrm{d}t_n$$

Where $a_i(t_i)$ is the prior marginal for the real age t_i of the i^{th} sample, and $a(t_1,...,t_n)$ is the n-dimensional prior function. For the present, the framework without non-age parameters is used for simplification.
The prior marginal can be calculated numerically with the Gibbs-sampling procedure in the same way, as done for the posterior marginals, explained in section 2.4.1. However, this is only true for a regular standardisable prior, i.e. a prior function with finite integral. In the practical use of Bayesian sequencing, there are often priors with infinite integral used. Even so, the posterior function has usually a finite integral, caused by the finite integral of the likelihood function, and therefore, there is no difficulty to calculate the marginal posterior functions. However, prior marginals can be meaningful for not standardised priors too, as shown in the next section.
The marginal prior function for a given sample age expresses the probability density distribution that arises from the prior information only, although in the specific representation of the chosen prior function. These densities for the individual samples can show very clearly, whether the prior function contains information that was not intended to be used.
A basic and classical example to illustrate the behaviour of prior marginals, is the following. We assume three samples with ages of known chronological order (t_1 younger than t_2, and t_2 younger than t_3), all lying within a fixed time range between t_a and t_b. Choosing the uniform prior, the corresponding prior function has the form:

$$a(t_1, t_2) \propto \left\{ \begin{array}{ll} 1 & \text{if} \quad t_a < t_1 < t_2 < t_3 < t_b \\ 0 & \text{else} \end{array} \right\}$$

In this simple case the prior marginals can be calculated easily based on Equation 3.1, resulting in the three marginals, given by the following equations and shown in Figure 3.2. The reason why the integrands are one is, that the prior function is expressed by the limits of the integrals completely.

$$a_1(t_1) \propto \int_{t_1}^{t_b} \mathrm{d}t_2 \int_{t_2}^{t_b} \mathrm{d}t_3 \quad 1 = \frac{1}{2} \cdot (t_b - t_1)^2$$

$$a_2(t_2) \propto \int_{t_a}^{t_2} \mathrm{d}t_1 \int_{t_2}^{t_b} \mathrm{d}t_3 \quad 1 = (t_2 - t_a) \cdot (t_b - t_2)$$

$$a_3(t_3) \propto \int_{t_a}^{t_3} \mathrm{d}t_1 \int_{t_1}^{t_3} \mathrm{d}t_2 \quad 1 = \frac{1}{2} \cdot (t_3 - t_a)^2$$

Looking on these prior marginals, cursorily it seems feasible that sample 1 has a high probability at young ages, sample 3 at old ages and sample 2 in the middle of the

range. On the other hand, we must not forget, that we have only information on the order of the samples, but not on their age difference. So if we assume, that the total span of the three ages is much less than the full range, they could lie close together anywhere within the limits, which is not consistent with the extremely low probabilities of sample age 1 at the old side or sample age 3 at the young side. Thus, even this very simple example shows already a fundamental problem of Bayesian sequencing. To reduce this difficulty is a main topic within this thesis.

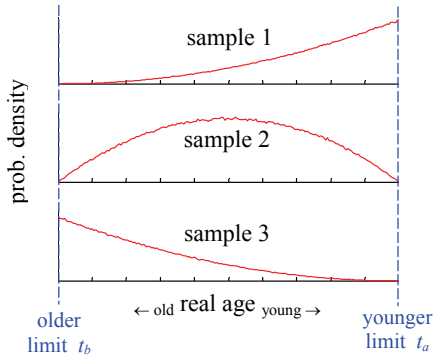

Figure 3.2: Marginals of the uniform prior for a sequence of three samples within fixed limits: t_1 younger than t_2 younger than t_3. (The plots are calculated with the normal Gibbs-sampling procedure, although direct integration would also be possible in this simple case.)

Although the example with a prior between two limits is illustrative, priors do not need limits in practical use, because the likelihood function limits the result anyway. Expanding the limits for the example above to infinity, gives undefined values as e.g. $a_1(t_1) \propto (\infty - t_1)^2$. This is not surprising, since the prior contains only relative time relations, and thus it can not deliver useful information on the absolute age scale. However, it is possible to get useful information on relative values, as e.g. on the total span covering all ages, or on the age difference between the oldest and the youngest sample in other words. This can be done analytically by transforming the coordinate system in a way, that there is a coordinate for this age difference available, and subsequent calculating the corresponding prior marginal. For a sequence of n samples, assuming t_1 youngest and t_n oldest, the transformation can be written in the following way:

$$\begin{aligned} \tau_1 &= t_1 \\ \tau_2 &= t_2 \\ &\vdots \quad\quad \vdots \\ \tau_{n-1} &= t_{n-1} \\ \vartheta &= t_n - t_1 \end{aligned}$$

Where $\tau_1,\ldots,\tau_{n-1}$ remain the normal ages of all but the last sample, and ϑ is the total span of the sequence. The uniform prior for this sequence, on an unrestricted age scale, is:

Equation 3.2: $$a(t_1,\ldots,t_n) \propto \left\{ \begin{array}{ll} 1 & \text{if} \quad t_1 < t_2 < \ldots < t_n \\ 0 & \text{else} \end{array} \right\}$$

The Jacobi determinant for the given transformation is constant and equal one. Therefore, the prior marginal of the total span τ can be evaluated, according to Equation 3.1, as below. The prior function is again defined by the integration limits.

$$a_\vartheta(\vartheta) \propto \int_{-\infty}^{+\infty} \mathrm{d}\tau_1 \int_{\tau_1}^{\tau_1+\vartheta} \mathrm{d}\tau_2 \int_{\tau_2}^{\tau_1+\vartheta} \mathrm{d}\tau_3 \ \ldots \int_{\tau_{n-3}}^{\tau_1+\vartheta} \mathrm{d}\tau_{n-2} \int_{\tau_{n-2}}^{\tau_1+\vartheta} \mathrm{d}\tau_{n-1}\ 1$$

The integral can be evaluated straightforward (by using repeatedly the substitution $(\tau_1+\vartheta-\tau_i)\ \mathrm{d}\tau_i = -\omega_i\ \mathrm{d}\omega_i$ for $i \le n$-2) resulting in:

$$a_\vartheta(\vartheta) \propto \vartheta^{n-2} \cdot \frac{1}{(n-2)!} \cdot \int_{-\infty}^{+\infty} \mathrm{d}\tau_1$$

Thus, the prior marginal of the total span is actually proportional to ϑ^{n-2}, because the two other factors can be summarised as a standardisation constant, no matter that this constant is infinite in the present case, using the unrestricted uniform prior. This means, that the uniform prior strongly favours longer total spans, which is a serious problem, because this was clearly not intended. The prior function should only carry the information of the temporal order of the samples.
To evaluate the prior marginal of the total span numerically, there is no coordinate transformation necessary. One simply defines the age difference t_n-t_1 as a result function in the way described in section 2.6.1. Naturally, the unrestricted uniform prior can not be evaluated on an infinite range, and range limits would alter the result. To overcome this problem, one can fade out the prior function by superposing a multi-dimensional Gaussian, centred on any point in the coordinate system with equal values for all sample ages. This will not corrupt the resulting prior marginal for values of the total span that are small compared with the (arbitrary) width of the Gaussian, and thus reflect the behaviour of the prior marginal. This procedure can simplest be performed with the developed program, by defining a linear calibration curve and setting artificially samples, all with an equal but arbitrary radiocarbon age and also an equal very large uncertainty. The upper graph of Figure 3.3 shows the result of a calculation performed in this way for a sequence of three samples, using the uniform prior. The prior marginal shows the expected behaviour of an increase with ϑ^{n-2}, which is a linear increase for n=3.
Further discussions to the problem of unwanted trends generated by the use of the uniform prior can e.g. be found by STEIER and ROM (2000) and BRONK RAMSEY (2000).

3.3 ENHANCED PRIOR SHAPES FOR SEQUENCES

To overcome the problem of the enhancement of longer total spans for a sequence with known temporal order when calculated with the uniform prior, it has become common, to use the so called 'uniform span prior', given by Equation 3.3, again for a sequence of n samples, where sample 1 is the youngest and sample n the oldest one.

Equation 3.3:
$$a(t_1, \ldots, t_n) \propto \left\{ \begin{array}{cc} 1 \,/\, (t_n - t_1)^{n-2} & \text{if} \quad t_1 < t_2 < \ldots < t_n \\ 0 & \text{else} \end{array} \right\}$$

Similar as done for the uniform prior in the previous section, the prior marginal for the total span can be calculated, using the same transformation as above, resulting in:

$$a_\vartheta(\vartheta) \propto \int_{-\infty}^{+\infty} \mathrm{d}\tau_1 \int_{\tau_1}^{\tau_1+\vartheta} \mathrm{d}\tau_2 \int_{\tau_2}^{\tau_1+\vartheta} \mathrm{d}\tau_3 \ \ldots \int_{\tau_{n-3}}^{\tau_1+\vartheta} \mathrm{d}\tau_{n-2} \int_{\tau_{n-2}}^{\tau_1+\vartheta} \mathrm{d}\tau_{n-1} \frac{1}{\vartheta^{n-2}} = \frac{1}{(n-2)!} \cdot \int_{-\infty}^{+\infty} \mathrm{d}\tau_1$$

The equation shows, that the prior marginal for the span of the sequence is constant, aside from the fact, that the prior has an infinite integral again, resulting in an infinite standardisation constant for the marginal.
As explained at the end of the previous section, the prior marginal can also be calculated numerically by fading out the unrestricted prior by an *n*-dimensional Gaussian. Figure 3.3 compares the behaviour of the marginals for the total span for the uniform prior at one hand, and the uniform span prior at the other hand, calculated for a sequence of three samples. The red lines extrapolate the results for small spans, the blue dots are the actually calculated values that decrease, caused by the restriction with the Gauss function. As mentioned above, the marginal of the uniform prior increases linearly with the span. By contrast the marginal of the uniform span prior (second graph) is constant, as expected.

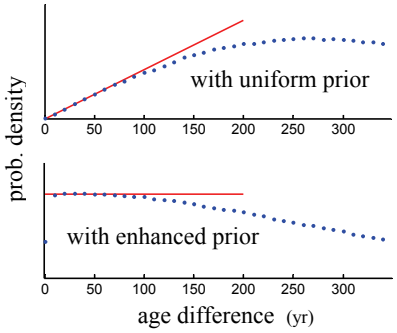

Figure 3.3: Comparison of the marginals of the total span (age difference between the outer sample ages) of the simple uniform prior (first plot), and the enhanced uniform span prior (second plot). To perform the calculations, the prior function is faded out by superposing a multi-dimensional Gaussian with arbitrary width (σ = 200 yr in this case). The red lines extrapolate the result for small spans, the blue dots are the actually calculated values that decrease, caused by the restriction with the Gauss function.

A very frequent situation in archaeology - at least in principle - is a temporal sequence of phases separated with boundaries (introduced in section 2.6.2). Each phase includes an arbitrary number of samples without known further time relation. In this case the uniform prior generates a problem that could be termed as 'statistical pressure' of samples. Fundamentally this problem is closely related to the widening of the span of a sequence, discussed just before. Again the problem is best illustrated by a simple example: One assumes an older phase including four samples, divided by a boundary b_2 from a younger phase including two samples, and fixed outer limits b_3 and b_1:

b_3	... fixed older limit
$t_{O,1}, t_{O,2}, t_{O,3}, t_{O,4}$	... four samples in older phase *O*
b_2	... boundary with unknown age
$t_{Y,1}, t_{Y,2}$	... two samples in younger phase *Y*
b_1	... fixed younger limit

The corresponding uniform prior is of the following form:

$$a = \begin{cases} 1 & \text{if} \quad b_3 > \{t_{O,1}, t_{O,2}, t_{O,3}, t_{O,4}\} > b_2 > \{t_{Y,1}, t_{Y,2}\} > b_1 \\ 0 & \text{else} \end{cases}$$

The first plot in Figure 3.4 shows the numerical calculation of the prior marginal for the boundary b_2 between the two phases, which is the probability for this boundary derived from the information carried by the uniform prior only, without considering radiocarbon measurements of the samples. It can be seen clearly, that the most likely duration of the older phase is higher as that of the younger phase, caused only by the different number of samples within the phases. This behaviour is clearly artificial and unwanted, because there can be arbitrary numbers of samples collected for the different phases, and this numbers need not to be related to the duration of the phases. So again, the mathematical description of the prior information has brought in artificial information that was not intended. Therefore, the following factor is usually added to the uniform prior:

$$a = \begin{cases} 1 / (b_3\text{-}b_2)^4 \cdot (b_2\text{-}b_1)^2 & \text{if} \quad b_3 > \{t_{O,1}, t_{O,2}, t_{O,3}, t_{O,4}\} > b_2 > \{t_{Y,1}, t_{Y,2}\} > b_1 \\ 0 & \text{else} \end{cases}$$

Where the powers of the phase-length factors reflect the number of sample within the phase. Calculating the marginal of boundary b_2 with this prior, results in a constant probability between the limits b_3 and b_1, independent of the number of samples within the phase. A constant probability is a reliable choice, as nothing is know about the actual durations of the phases. (For sure, even this prior is one particular subjective choice, as discussed in section 3.1.) The corresponding numerical calculation is given by the second plot of Figure 3.4. The deviation from the ideal constant function is caused by the finite step size and by the finite number of sampled points within the Gibbs-sampling process.

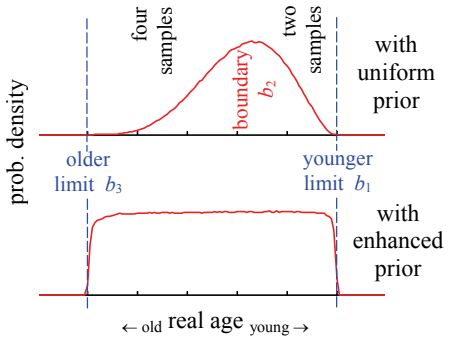

Figure 3.4: Prior marginal for a boundary between two phases with different numbers of samples within. The first plot shows the marginal of the uniform prior and the second the marginal of an enhanced prior, including factors that eliminate the 'statistical pressure' of the samples within the phases.

In case of this simple example the prior marginal can also be calculated analytically. For the enhanced prior the marginal results actually in a constant, which can be seen by straightforward integration as follows:

$$a_{b_2}(b_2) \propto \int_{b_1}^{b_2} dt_{O,1} \ldots \int_{b_1}^{b_2} dt_{O,z_O} \int_{b_2}^{b_3} dt_{Y,1} \ldots \int_{b_2}^{b_3} dt_{Y,z_Y} \frac{1}{(b_3 - b_2)^{z_O} \cdot (b_2 - b_1)^{z_Y}} = 1$$

Where z_O and z_Y are arbitrary numbers of samples in the older and in the younger phase. Again the prior constraints are realised by the integration limits, as repeatedly done earlier.

Finally, one can generalise this prior factors for an arbitrary number of phases, and additional add the span-correcting term - discussed at the beginning of this section -

for the whole sequence of boundaries, resulting in a commonly used basic prior form, which will be denoted as 'uniform overall-span prior' for a sequence of phases furthermore. The formal description of this prior type is:

Equation 3.4:

$$a(t_1,\dots,t_n,b_1,\dots,b_m) \propto \left\{ \begin{array}{cc} (b_m - b_{m-1})^{-z_{m,m-1}} \cdot \dots \cdot (b_2 - b_1)^{-z_{2,1}} \cdot (b_m - b_1)^{-(m-2)} & \text{for correct order} \\ 0 & \text{for incorrect order} \end{array} \right\}$$

The equation describes the prior of a sequence of m-1 phases, limited by m boundaries of unknown ages $b_1,\dots,b_m$. The sequence includes totally n samples with unknown real ages $t_1,\dots,t_n$. Each sample is allocated to a particular phase, where $z_{j,j\text{-}1}$ is the number of samples within the phase limited by the boundaries b_j and $b_{j\text{-}1}$. The notation 'correct order' means, that all boundaries are in the correct temporal order, corresponding to the known order of the phases, and additionally, all real sample ages fall between the boundaries of the phase they are allocated too.
Just for completeness, the uniform prior for a sequence of phases is given by Equation 3.5, using the same explanations:

Equation 3.5:

$$a(t_1,\dots,t_n,b_1,\dots,b_m) \propto \left\{ \begin{array}{cc} 1 & \text{for correct order} \\ 0 & \text{for incorrect order} \end{array} \right\}$$

Similarly to a simple sequence of sample ages only, the uniform overall-span prior for a sequence of phases shows a constant marginal for the total span (i.e. the age difference of the outer boundaries b_m-b_1) too. This can be seen analytically by using the following transformation:

$$\begin{array}{rcl} \tau_1 & = & t_1 \\ \vdots & & \vdots \\ \tau_n & = & t_n \\ \beta_1 & = & b_1 \\ \vdots & & \vdots \\ \beta_{m\text{-}1} & = & b_{m\text{-}1} \\ \vartheta & = & b_m - b_1 \end{array}$$

Where t and b are the original sample and boundary age coordinates. τ, β and ϑ define the transformed coordinate system, where ϑ is the total span of the sequence. As by the analogous transformation, used for a simple sequence of samples earlier, the Jacobi-determinant is constant for the current transformation too. Accordingly, the prior marginal of the total span can be written as:

$$a_\vartheta(\vartheta) \propto \int_{-\infty}^{+\infty} \mathrm{d}\beta_1 \int_{\beta_1}^{\beta_1+\vartheta} \mathrm{d}\beta_2 \int_{\beta_1}^{\beta_2} \mathrm{d}\tau_{..} \dots \int_{\beta_1}^{\beta_2} \mathrm{d}\tau_{..} \int_{\beta_2}^{\beta_1+\vartheta} \mathrm{d}\beta_3 \int_{\beta_2}^{\beta_3} \mathrm{d}\tau_{..} \dots \int_{\beta_2}^{\beta_3} \mathrm{d}\tau_{..} \dots$$
$$\dots \int_{\beta_{m-3}}^{\beta_1+\vartheta} \mathrm{d}\beta_{m-2} \int_{\beta_{m-3}}^{\beta_{m-2}} \mathrm{d}\tau_{..} \dots \int_{\beta_{m-3}}^{\beta_{m-2}} \mathrm{d}\tau_{..} \int_{\beta_{m-2}}^{\beta_1+\vartheta} \mathrm{d}\beta_{m-1} \int_{\beta_{m-2}}^{\beta_{m-1}} \mathrm{d}\tau_{..} \dots \int_{\beta_{m-2}}^{\beta_{m-1}} \mathrm{d}\tau_{..} \int_{\beta_{m-1}}^{\vartheta+\beta_1} \mathrm{d}\tau_{..} \dots \int_{\beta_{m-1}}^{\vartheta+\beta_1} \mathrm{d}\tau_{..} \frac{1}{\vartheta^{m-2}} \cdot$$
$$\cdot \frac{1}{(\beta_2-\beta_1)^{z_{2,1}} \cdot \dots \cdot (\beta_{m-1}-\beta_{m-2})^{z_{m-1,m-2}} \cdot (\vartheta+\beta_1-\beta_{m-1})^{z_{m,m-1}}}$$

Where each block of integrals over the sample age coordinates $\tau_{..}$, includes one integral for each sample, which is allocated to the phase defined by the integration limits. The fact that the last two sample age blocks are not separated by an integral over a boundary, is caused by the missing integral over the total span coordinate ϑ, which does not occur, because ϑ is the argument of the marginal. Evaluating the whole integral, the sample-age blocks cancel out with the factors in the denominator of the last fraction, leading to:

$$a_\vartheta(\vartheta) \propto \int_{-\infty}^{+\infty} \mathrm{d}\beta_1 \int_{\beta_1}^{\beta_1+\vartheta} \mathrm{d}\beta_2 \int_{\beta_2}^{\beta_1+\vartheta} \mathrm{d}\beta_3 \ \ldots \int_{\beta_{m-3}}^{\beta_1+\vartheta} \mathrm{d}\beta_{m-2} \int_{\beta_{m-2}}^{\beta_1+\vartheta} \mathrm{d}\beta_{m-1} \ \frac{1}{\vartheta^{m-2}}$$

This integral is analogous to that of a simple sequence of sample ages already discussed, and results in

$$\frac{1}{(m-2)!} \cdot \int_{-\infty}^{+\infty} \mathrm{d}\beta_1 \quad ,$$

which has to be seen as constant, apart from the fact, that the used uniform overall-span prior for sequences of phases has also infinite integral on an unrestricted age scale.

Concluding, the use of the uniform span prior for a simple sequence of samples (Equation 3.3) or the uniform overall-span prior for a sequences of phases (Equation 3.4) will deliver more reliable results, because both avoid the trend of the uniform prior (Equation 3.2 for the first and Equation 3.5 for the second case) to overrate large total spans, and the later avoids unwanted effects of the 'statistical pressure' within the phases too. Although this enhanced prior functions are beneficial, one must not forget, that their choice is still subjective, as discussed in section 3.1.
The basic priors types for specific types of age sequences introduced in this section are commonly uses in present archaeological sequencing and are available e.g. in the widely used program package 'OxCal' (BRONK RAMSEY 1995 and 2001*a*). The mathematical foundations can further be found by BUCK *et al.* (1992), STEIER and ROM (2000) and NICHOLLS and JONES (2001).

3.4 THE MAXIMUM ENTROPY METHOD

As discusses in section 3.1, it is not possible to define the prior function in an unambiguous way, even when the prior information can be described unambiguously in a non-functional way. For example, there are many different possible prior functions to realise the information 'sample 1 is older than sample 2'.
However, there is a theoretical concept to find the 'best' of all possible prior functions realising a given prior information, by choosing that function with the lowest content of unwanted artificial information. Therefore, one uses the entropy measure to define the information content of a probability density function, which can be briefly justified in the following way: For the present, we deal with a discrete set of probabilities P_i, defining the probability of the occurrence of an event i, instead of a continuous density function. Thus, the value

$$I_i = - \ln P_i$$

is a well defined measure for the information obtained by the occurrence of the event *i*, because it shows the following reliable characteristics: The smaller the probability of the event, the higher the information obtained. If the probability of the event is one, the obtained information by its occurrence is zero. The total information obtained by the independent occurrence of several events i,j,k,... is the sum of the individual information values $I = I_i+I_j+I_k+...$, because the total probability for the occurrence of the events is $P_i \cdot P_j \cdot P_k \cdot ...$. In the next step one can calculate the mean value (expectation) of the obtained information by the occurrence of an event, which is:

$$\sum_i P_i \cdot I_i \quad = \quad -\sum_i P_i \cdot \ln P_i \quad = \quad S$$

A high mean value of the obtained information means a high degree of indeterminacy of the considered system, or a high entropy *S* in other words. So, to get a set of probabilities (or a probability density function in the continuous equivalent) that contains the lowest amount of unwanted artificial information, one has to maximize its entropy *S*, under the given constraints. Thinking specifically on the prior function, the constraints carry the known prior information.
For specific cases, the maximisation of the entropy can define a particular prior function, best illustrated by the following example: What is the 'best' prior function for a value y that has to be positive, and of which one knows further its expectation value μ. One could think e.g. on the age difference of two samples found in two neighbouring layers, and one knows estimates of the most likely time spans associate with the layers from other evidence. Let us denote the continuous density function for the prior with $p(y)$, and asign a discrete set of probabilities P_i, that are the probabilities for the rounded age differences $y_i=i$ years. (We stay in this content exceptionally by the general notation '*p*' for probability density, instead of using the specific notation '*a*' for the prior density function.) The prior function should be normalised, which leads to the first condition

$$1 = \int_0^\infty p(y)\,\mathrm{d}y \qquad \text{or} \qquad 1 = \sum_i P_i \quad .$$

The knowledge of the expectation value leads to a further condition, which is:

$$\mu = \int_0^\infty y \cdot p(y)\,\mathrm{d}y \qquad \text{or} \qquad \mu = \sum_i y_i \cdot P_i$$

Now the entropy

$$S = -\sum_i P_i \cdot \ln P_i$$

has to be maximized. This can be done by using the method of Lagrange multipliers, which means to solve the following equation for each *i*:

$$\frac{\partial F}{\partial P_i} = 0 \qquad \text{with}$$

$$F = -\sum_i P_i \cdot \ln P_i + \lambda_1 \cdot \left(1 - \sum_i P_i\right) + \lambda_2 \cdot \left(\mu - \sum_i y_i \cdot P_i\right)$$

This results in:

$$P_i = e^{-\lambda_1 - 1} \cdot e^{-\lambda_2 \cdot y_i}$$

Changing now to the continuous representation and using this result within the conditions from above, leads to

$$\mu = \int_0^\infty y \cdot e^{-\lambda_1 - 1} \cdot e^{-\lambda_2 \cdot y} \, \mathrm{d}x \qquad \text{and} \qquad 1 = \int_0^\infty e^{-\lambda_1 - 1} \cdot e^{-\lambda_2 \cdot y} \, \mathrm{d}y \quad ,$$

resulting further in

$$\frac{1}{\mu} = \lambda_2 \qquad \text{and} \qquad \frac{1}{\mu} = e^{-\lambda_1 - 1} \quad .$$

Putting this into the equation got for the probabilities P_i from above, leads to the final result for the 'ideal' prior probability density:

$$p(y) = \frac{1}{\mu} \cdot e^{-\frac{y}{\mu}}$$

Thus, knowing the expectation value for a non-negative value, an exponential decreasing function with this expectation value is the probability density that expresses this knowledge with a minimum of unwanted additional information.
A detailed explanation of the principle of maximum entropy can be found e.g. by SIVIA (1996), section 5.2. A fundamental explanation of the entropy measure and additionally of the method of Lagrange multipliers can be found e.g. by ADAM and HITTMAIR (1999).

Concluding, the method of maximising the entropy could be an approach to counter the problem of ambiguity in defining the prior function (section 3.1). However, even though the maximum entropy criterion is a reliable, one must not forget that the prior found by this method is still a particular choice out of an infinite number of possible different shaped functions, and therefore, can not fully solve the ambiguity problem. In this work a more general approach to deal with this problem (introduced in chapter 5) will be analysed in detail.

4 MEASURES FOR THE AGREEMENT OF MODEL AND DATA

The model-data agreement has turned out to be an essential indicator when trying to establish a procedure to perform robust Bayesian analysis, which is described in chapter 5. In principle, agreement of model and data is given if the model does not result in a prediction of real sample ages that can not produce the measured radiocarbon ages. Simplified, this is realised if the resulting modelled probability density (represented by the posterior function) lies in regions that are covered by the un-modelled density (represented - if accepting a slight theoretical imprecision - by the likelihood function).
Below, two basically different ways to classify the model-data agreement are discussed. The single sample based agreement index (section 4.1) is commonly used, although the shown way to define a reliable threshold level (described in section 4.1.2) is new. An alternative system for defining an agreement measure (described in section 4.3) is based on the prior predictive distribution (or more exact speaking, on the particular point on it, which is relevant for the actual set of measurements), which delivers an integral agreement measure for the model on the whole. The detailed meaning of the prior predictive distribution is discussed in section 4.2. Since the latter agreement system provides a theoretically strictly defined absolute measure, there arise fundamental problems when using unrestricted prior functions. Thus, in section 4.3.2 a procedure is developed to solve this problem by restricting the prior function with the help of a 'domain function' based on the whole set of used radiocarbon ages. As for the single sample based index above, a reliable method to define a threshold level is developed again (also described in section 4.3.2).
The use of the second agreement measure results mathematically in the need of integrating multi-dimensional volumes. Since there were no simple numerical procedures available, a new method was developed (described in section 4.4), which is also based on the Gibbs sampling procedure, as the basic sequencing procedures are. (If the reader is not interested in the description of this integration method, section 4.4 can be skipped without a loss in understanding of the remaining parts.)

4.1 A MEASURE BASED ON SINGLE SAMPLE AGREEMENTS

4.1.1 Definition

Technically, the simplest way to test the agreement of model and data is the comparison of the marginal posterior functions for each sample with the according single-sample likelihood functions. A measure that represents the model-data agreement, as qualitatively characterised just above, is given by Equation 4.1:

Equation 4.1:
$$I_i = \frac{\int_{-\infty}^{+\infty} l_i(t_i) \cdot p_i(t_i)\ \mathrm{d}t_i}{\int_{-\infty}^{+\infty} l_i^2(t_i)\ \mathrm{d}t_i}$$

Where the single-sample likelihood function $l_i(t_i)$ and the marginal posterior function $p_i(t_i)$ for i^{th} sample are defined according to Equation 2.6 and Equation 2.9, however, they have to be standardised (with respect to t_i) previously to their use in Equation 4.1. The 'single sample agreement indices' I_i defined in this way, is widely used to analyse sequencing results (e.g. in OxCal program; BRONK RAMSEY 1995 and 2001*a*). The few cases given in Figure 4.1, where artificial rectangular function shapes are used to see the resulting agreement index immediately, illustrates the behaviour of the index very clearly, which shows the required characteristics: It is high (one) if the posterior lies completely inside the likelihood, and it decreases the more the posterior lies outside.
It should be noted, that in general, the single sample agreement index can even exceed the value one, what is the case, if the marginal posterior falls in a region with a particular high likelihood value.
It is plausible, that the product of all single sample agreements (Equation 4.2) is a possible agreement index for the sequence as a whole.

Equation 4.2:
$$I_\Pi = \prod_i I_i$$

To make agreement indices really usefull, one has to understand the meaning of the specific values found. Or in other words, one has to be able to define a reliable threshold that indicates good agreement. A possible approach is given in the following section.

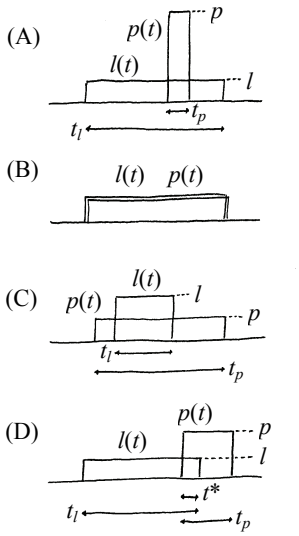

Figure 4.1: Assuming rectangular function shapes for single sample likelihood and marginal posterior, the agreement index I can be calculated simply for characteristic cases and shows the required behaviour. (Thereby one considers, that standardisation implies the relation $l \cdot t_l = p \cdot t_p = 1$; the index i for the specific sample is skipped in this illustration.)

Case A: The marginal posterior lies fully within the likelihood:
$I = (p \cdot l \cdot t_p) / (l^2 \cdot t_l) = (p \cdot t_p) / (l \cdot t_l) = 1$

Case B: Marginal posterior and likelihood are identically:
$I = 1$

Case C: The marginal posterior covers fully the likelihood:
$I = (p \cdot l \cdot t_l) / (l^2 \cdot t_l) = p/l = t_l/t_p$

Case D: The marginal posterior overlaps partially with the likelihood:
$I = (p \cdot l \cdot t^*) / (l^2 \cdot t_l) = (p/l) \cdot (t^*/t_l) = (t_l/t_p) \cdot (t^*/t_l) = t^*/t_p$

4.1.2 Quantitative meaning

To understand quantitatively the value of the product of all single sample agreements (I_Π), which shall be used as agreement index for the total sequence, one can analyse the following simple but meaningful situation. Imagine there is a set of n samples measured with accuracies σ_i. Assuming a linear calibration curve $c(t)$ with slope one, each sample generates a corresponding Gaussian likelihood function $l_i(t_i)$ with width σ_i. (The slope of the calibration curve is irrelevant for the following considerations, but a slope of one simplifies the notation, because there is no scaling factor between the radiocarbon and the real age axes.) Now, one can calculate the agreement index for the best possible model or prior function, which would be for sure that, where the actual real ages t_i^* are already known. The realisation of this prior function would be an n-dimensional delta-function (or a narrow Gaussian) at the position of the actual real-age set. The prior would force the marginal posterior functions $p_i(t_i)$ to be delta-functions (or narrow Gaussians) at the positions of the actual individual sample ages. Due to the uncertainties of the radiocarbon measurements, the centres of the single-sample likelihood functions would deviate from these positions by the measurement errors ε_i, illustrated in Figure 4.2.

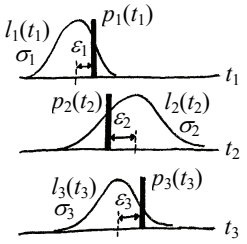

Figure 4.2: Illustration of the 'ideal' model, where the actual sample ages would be known. This leads to marginal posterior $p_i(t_i)$ that are delta-functions at the actual real ages. The single-sample likelihood functions $l_i(t_i)$ deviate from the marginal posteriors by the actual appeared measurement errors ε_i.

Now the single sample agreement indices can be calculated as:

$$I_i = \frac{\int_{-\infty}^{+\infty} l_i(t_i)\cdot p_i(t_i)\ \mathrm{d}t_i}{\int_{-\infty}^{+\infty} l_i^2(t_i)\ \mathrm{d}t_i} = \frac{l_i(t_i^*)}{\int_{-\infty}^{+\infty} l_i^2(t_i)\ \mathrm{d}t_i} = \frac{\frac{1}{\sqrt{2\pi}\cdot\sigma_i}\cdot e^{-\frac{\varepsilon_i^2}{2\cdot\sigma_i^2}}}{\int_{-\infty}^{+\infty} \frac{1}{2\pi\cdot\sigma_i^2}\cdot e^{-\frac{t_i^2}{2\cdot\sigma_i^2}}\ \mathrm{d}t_i} = \sqrt{2}\cdot e^{-\frac{\varepsilon_i^2}{2\cdot\sigma_i^2}}$$

Thus, for the total agreement I_Π for the whole set of samples follows:

$$I_\Pi = \prod_i I_i = \prod_i \sqrt{2}\cdot e^{-\frac{\varepsilon_i^2}{2\cdot\sigma_i^2}} = 2^{\frac{n}{2}}\cdot e^{-\frac{1}{2}\cdot\sum_i\left(\frac{\varepsilon_i}{\sigma_i}\right)^2}$$

If we assume independent measurement errors, the sum $\Sigma_i(\varepsilon_i/\sigma_i)^2$ (denoted as χ^2) follows the well known χ^2-distribution, which means that

$$\sum_i\left(\frac{\varepsilon_i}{\sigma_i}\right)^2 \qquad \text{is less than} \qquad \chi^2_{th}(P^*, n) \qquad \text{with a probability of } P^*,$$

where the value χ_{th}^2 (threshold level) for a particular probability P^* and a particular number of samples n (i.e. degree of freedom) is given via its inverse function in the following way:

$$P^*(\chi_{th}^2, n) = P_\Gamma\left(\frac{\chi_{th}^2}{2}, \frac{n}{2}\right) \quad \text{with} \quad P_\Gamma(x,y) = \frac{\int_0^x e^{-z} \cdot z^{y-1}\,\mathrm{d}z}{\int_0^\infty e^{-z} \cdot z^{y-1}\,\mathrm{d}z}$$

Where P_Γ is the so called incomplete gamma-function, which can be evaluated numerically. The mathematical framework to the χ^2-distribution can be found e.g. by PRESS *et al.* (1992).
Accordingly, with a probability of P^*, for the 'ideal' model, the total agreement index I_Π is larger than

Equation 4.3: $$I_\Pi^* = 2^{\frac{n}{2}} \cdot e^{-\frac{1}{2}\cdot\chi_{th}^2(P^*, n)}$$

Hence, if a threshold is needed to decide whether a model (prior function) is in agreement with the measurements sufficiently or not, it could be based on this value. Thus, one has to set P^* to an adequate high level (e.g. 0.9545, 2σ), which guarantees, that the 'ideal' model is always accepted, except for the unavoidable number of cases with very large measurement errors, which occur with a probability of 1-P^*. Real models, which probably do not agree with the data as well, will fail the threshold with corresponding higher probability.
For sure, the values for I^*_Π are highly dependent on the number of samples. This can be avoided by using $I_\Pi^{1/\sqrt{n}}$ as total agreement index, which is e.g. done in the OxCal program (BRONK RAMSEY 1995 and 2001*a*). The recommended threshold there, which is based on slightly different considerations, results in roughly similar values as given by Equation 4.3 for P^*=0.9545. (To use $1/\sqrt{n}$ as exponent is reasonable, because when assuming the single sample agreement indices scattering Gaussian around the value one with given equal deviations, the total agreement would scatter around one with the same deviation, independently from the number of samples; $\sigma((\Pi I_i)^{1/\sqrt{n}})=\sigma_i$. That means, that $I_\Pi^{1/\sqrt{n}}$ deviates from one with equal sensitivity, for any number of samples.)
The values for $I^{*\,1/\sqrt{n}}_\Pi$, when I^*_Π is calculated by Equation 4.3 for different probabilities P^*, are shown in Figure 4.3.

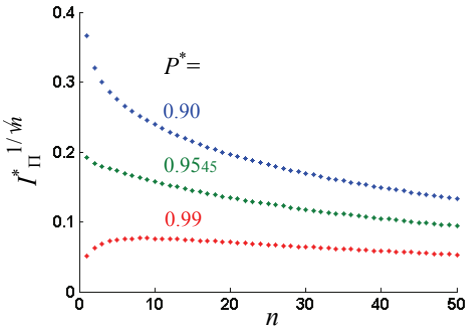

Figure 4.3: Shows the threshold level of $I_\Pi^{1/\sqrt{n}}$, which is statistically passed with a probability of P^*, in case of an ideal model. Where I_Π is the total agreement index defined as product of all single sample agreement indices and n is the sample number. The threshold level can be calculated by Equation 4.3 numerically.

For the special case of n=1 and P^*=0.9545 (2σ), $\chi_{th}^2(P^*,n)$ has, according to its definition, the exact value of 2^2, and I^*_Π results in $\sqrt{2}/e^2$ (0.191). When analysing the samples of a sequence individually, this value could be used as a threshold for the single sample agreement indices.
It should be noted here, that sufficient agreement criteria and thresholds will become essential for robust Bayesian analysis, which will be described in chapter 5.

4.2 THE MEANING OF THE 'NORMALISATION' TERM WITHIN THE BAYES THEOREM

The Bayes theorem was introduced in its exact standardised form by Equation 2.12 in section 2.2.2 as follows:

$$p(t_1,\dots,t_n \mid x_1,\dots,x_n) \;=\; \frac{l(x_1,\dots,x_n \mid t_1,\dots,t_n)\cdot a(t_1,\dots,t_n)}{\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} l(x_1,\dots,x_n \mid t_1,\dots,t_n)\cdot a(t_1,\dots,t_n)\;\mathrm{d}t_1\dots\mathrm{d}t_n}$$

Where a, l, p are prior function, likelihood function and posterior function; t_i the real age coordinates and x_i the radiocarbon ages. In section 2.3, the 'normalisation term' in the denominator was denoted as 'prior predictive distribution' labelled with v. Using additionally the vector notation $\mathbf{t}$ = (t_1,...,t_n) and $\mathbf{x}$ = (x_1,...,x_n) led to the following representation of the theorem:

$$p(\mathbf{t} \mid \mathbf{x}) \;=\; \frac{l(\mathbf{x} \mid \mathbf{t})\cdot a(\mathbf{t})}{v(\mathbf{x})} \qquad \text{with} \qquad v(\mathbf{x}) \;=\; \int_{\mathrm{vol}} l(\mathbf{x} \mid \mathbf{t})\cdot a(\mathbf{t})\;\,d\mathbf{t}$$

The exact mathematical relations and the fundamental meaning of the terms within the equations were already discussed in section 2.3.

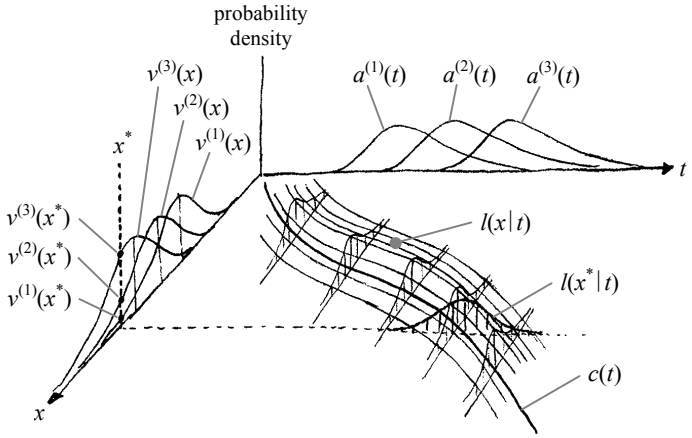

Figure 4.4: Illustration of the prior predictive distributions $v^{(i)}(x)$ and the 'prior predictions' $v^{(i)}(x^*)$ for three different one-dimensional models $a^{(i)}(t)$. See details in the text. The high value of $v^{(3)}(x^*)$ indicates, that model 3 shows the best agreement with the measured radiocarbon age x^*.

To calculate the marginal posterior distributions, $v(\mathbf{x})$ is not needed and can be skipped, as done in basic Bayesian sequencing, described in section 2.2.1. Nevertheless, $v(\mathbf{x})$ has a fundamental meaning in characterising the agreement of the

prior function and measured data set. This can be understood clearly by a fundamental consideration illustrated by Figure 4.4. The figure shows the correlation between a set of different chosen prior functions and the corresponding set of prior predictive distributions.

Note, that the graph shows a one-dimensional model (only one single real age) at a two-dimensional coordinate system, built up by the real-age axis and the radiocarbon-age axis. This is a completely different visualisation compared to these used many times before, where the probability densities of two-dimensional models where plotted on the two-dimensional space, given by the two real-age coordinates.

The two dimensional Gaussian ridge along the calibration function $c(t)$ represents the complete likelihood function $l(x|t)$, when not restricted to a particular measured radiocarbon age. With its help, a particular prior function $a^{(i)}(t)$ can be 'transformed' to a corresponding prior predictive distribution $v^{(i)}(x)$, by evaluating

$$v^{(i)}(x) = \int_{-\infty}^{\infty} l(x \mid t) \cdot a^{(i)}(t)\ dt$$

for any value of x. The prior predictive distribution is the probability density with which a particular radiocarbon age is expected, before a radiocarbon measurement has been performed. If a measured radiocarbon age (x^*) is available, one can focus on the value of the prior prediction distributions at this particular age, which will be denoted shortly as 'prior prediction' ($v^{(i)}(x^*)$) in the following. It is obvious, that this value should be a fundamental measure for the model-data agreement. The problem is, that $v^{(i)}(x^*)$ has no absolute meaning, because it is not the probability that the radiocarbon age x^* occurs, it is just a probability density that allows to determine the probability, that the radiocarbon age occurs in a particular interval. Fortunately, this problem diminishes when one wants a comparison between two models only, because the ratio between the prior predictions of two different models ($v^{(i)}(x^*)/v^{(j)}(x^*)$) is as well the ratio between the probabilities to get the actual measured radiocarbon age, according to one or to the other model. Therefore, the ratio of the prior predictions is obviously a reliable measure to compare the level of model-data agreement for different models, and it is a well known value in model comparison in general, denoted as 'Bayes factor' (see e.g. GARCIA-DONATO, 2005 or BERGER and PERICCHI, 1996):

Equations 4.4:

$$B_{ij} = \frac{v^{(i)}(\mathbf{x})}{v^{(j)}(\mathbf{x})} \qquad \text{with} \qquad v^{(k)}(\mathbf{x}) = \int_{\mathrm{vol}} l(\mathbf{x} \mid \mathbf{t}) \cdot a^{(k)}(\mathbf{t})\ d\mathbf{t}$$

B_{ij} is the Bayes factor that compares model $a^{(i)}$ with model $a^{(j)}$; $v^{(i)}$ and $v^{(j)}$ are the according prior predictions. In this notation, the actual measured radiocarbon value is not indicated by '*' any more, to be consistent with the notation used mainly in the text, where x means already the actual measured value. Further, the Bayes factor is given in its general form for the n-dimensional case, where $\mathbf{x}$ is the whole set of measured radiocarbon ages, and $\mathbf{t}$ indicates the n-dimensional real age space.

4.3 AN AGREEMENT MEASURE BASED ON THE PRIOR PREDICTIVE DISTRIBUTION

4.3.1 Definition

As described in the previous section, the ratio of the prior predictions, i.e. the Bayes factor, is already a reliable measure to compare the levels of sample-data agreements of different models. However, the prior prediction is only well defined in case of a standardised prior function. Actually, it is common in Bayesian sequencing to use unrestricted priors with infinite integral too. For example, all priors discussed in sections 3.2 and 3.3 that were defined on an unrestricted time range have infinite integrals. Thus, for calculating the prior-prediction, one has to standardise that kind of priors too. The problems resulting from the use of non-standardised priors (or fequently denoted as 'improper' priors) are well known in the literature; see e.g. BERGER and PERICCHI, 1996 or WASSERMAN, 1996.
Standardisation of all kinds of prior functions can be achieved by confining every prior function to a restricted domain generally. A possible approach is to characterise this domain by a kind of overall probability density for possible positions of real ages. This density can then be used as a weighting function that defines the domain gradually. The idea introduced in this thesis to construct such a density is the following: One assumes, that the real age of each sample could show any value that can result for any of the measured radiocarbon ages. Common for all sample ages, the likelihood to achieve this requirement is proportional to the sum of all single-sample likelihood functions (when assuming a constant a-priori possibility on the real-age axis). This single sample densities are combined to a multi-dimensional probability density, which will be called 'domain function' $\lambda(t_1,...,t_n)$; see Equation 4.5. (The simplified notation for a fixed set of determined radiocarbon ages, not indicating the radiocarbon ages explicitly, is used.)

Equation 4.5:
$$\lambda(t_1,\ldots,t_n) \propto \left(\sum_{i=1}^{n}\left(\frac{l_i(t_1)}{\int_{-\infty}^{+\infty} l_i(t)\,\mathrm{d}t}\right)\right)\cdot\ \ldots\ \cdot\left(\sum_{i=1}^{n}\left(\frac{l_i(t_n)}{\int_{-\infty}^{+\infty} l_i(t)\,\mathrm{d}t}\right)\right)$$

Where l_i are the single-sample likelihood functions as defined in Equation 2.6
Equation 4.5 shows the domain function for a simple example and compares it with the likelihood function for clarification.

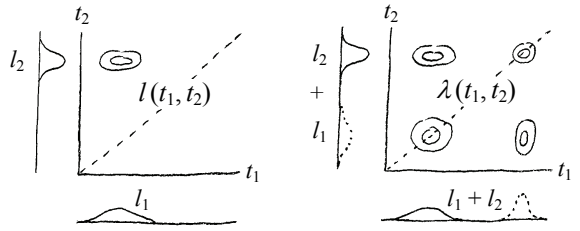

Figure 4.5: To build the domain function, all single-sample likelihood functions l_i are added in the same way for all sample age axes. Based on that, the multi-dimensional domain function λ (right) is built similarly to the multi-dimensional likelihood function (left).

The domain function can also be extended to cases where additional parameters are used, insofar the parameters are ages, as e.g. phase boundaries. If there is for example a model with $n+m$ parameters in total ($s_1,...,s_{n+m}$), containing n sample ages ($t_1,...,t_n$) and m boundaries ($b_1,...,b_m$), one can use the domain function:

Equation 4.6:
$$\lambda(s_1,\dots,s_{n+m}) \;=\; \lambda(t_1,\dots,t_n,b_1,\dots,b_m) \;\propto\; \prod_{j=1}^{n+m}\left(\sum_{i=1}^{n}\left(\frac{l_i(s_j)}{\int_{-\infty}^{+\infty} l_i(t)\,\mathrm{d}t}\right)\right)$$

Of course, this seems to be not really natural, because if there are gaps between the single sample likelihoods, they are also excluded for the boundaries. Therefore one could think about using a domain function that includes these gaps. However, this would make the function less simple and less directly related to the measurements, and it is not really necessary, because the concept, as explained below, is not violated in principle, although the domain function emphasises the possible older and younger limits for the boundaries.

Based on the domain function one can calculate the prior-prediction (see Equations 4.4 right part) with the restricted prior function, in the way given by Equation 4.7. All used functions within the equation (l ... likelihood, a ... prior function, λ ... domain function) may have arbitrary absolute values; the needed standardisations are explicitly included in the equation):

Equation 4.7:
$$J \;=\; \frac{v_{a\lambda}}{v_{\lambda}} \;=\; \frac{\int_{vol} l\cdot\dfrac{a\cdot\lambda}{\int_{vol} a\cdot\lambda\,\mathrm{d}\mathbf{t}'}\,\mathrm{d}\mathbf{t}}{\int_{vol} l\cdot\dfrac{\lambda}{\int_{vol}\lambda\,\mathrm{d}\mathbf{t}'}\,\mathrm{d}\mathbf{t}} \;=\; \frac{\int_{vol} l\cdot a\cdot\lambda\;\mathrm{d}\mathbf{t}\;\Big/\;\int_{vol} a\cdot\lambda\;\mathrm{d}\mathbf{t}}{\int_{vol} l\cdot\lambda\;\mathrm{d}\mathbf{t}\;\Big/\;\int_{vol}\lambda\;\mathrm{d}\mathbf{t}}$$

One can see, that the prior function a needs not to be standardisable any more, it is sufficient if $a\cdot\lambda$ has finite volume, what is usually the case due to the strong decrease of the single-sample likelihood functions towards ages that do not agree with the measurements.
$v_{a\lambda}$ is the prior-prediction, calculated with a prior that combines the original prior function with the domain probability, and v_λ is the prior-prediction using a prior that is the domain function only. The value J is the Bayes factor that compares these two, and delivers the increase ($J>1$) or decrease ($J<1$) of the probability to get the determined set of radiocarbon ages, that occurs when the using the prior information. Thus, J is a reliable and quantitative measure of the agreement of the prior function with the radiocarbon ages; it will be called 'agreement factor' furthermore. Note that the fact, that the prior-prediction cannot deliver absolute values, because it is a density (see section 4.2), is no difficulty any more, because only a ratio (the Bayes factor) is used.

To get an impression, that the agreement factor J is a very meaningful measure, one can think on a sequence of n samples, where the single-sample likelihood functions are assumed to be clearly separated. If J would be calculated without prior information (prior is constant at any point), it gives the neutral value of one. (This can be seen trivially, because $v_{a\lambda}$ and v_λ become identical.) If one uses the uniform

prior for a sequence of samples, that is one for age combinations that are in the order as supposed by the prior information, and zero elsewhere, then there are two results possible: If the assumed prior order is in disagreement with the order of the radiocarbon ages, the agreement factor would be zero (For nearly totally separated single sample likelihoods, the multi-dimensional likelihood lies nearly completely within the region where the prior is zero). But if the assumed prior order is in agreement with the order of the radiocarbon ages, the resulting value is the faculty of n.
This can be deduced by the following consideration: As the multi-dimensional likelihood lies completely in the constant non-zero part of the prior, the values for $v_{a\lambda}$ and v_{λ} differ only by a factor arising from the standardisation of $a \cdot \lambda$ and λ. Caused by the symmetry of λ this factor is equivalent to the factor arising just from the normalisation of the uniform prior compared with a constant prior. See Figure 4.5 for clarification, which represents exactly the discussed situation in case of two samples, when assuming a with a constant non-zero part (e.g.) left above from the diagonal, and a zero part right below. Thus, the needed factor is the ratio between the volume of the constant prior and the volume of the uniform prior, calculated on equal but arbitrary age range ($t_{lim,Y}$ to $t_{lim,O}$):

$$\frac{vol\,(a_{constant})}{vol\,(a_{uniform})} = \frac{\int_{t_{lim,Y}}^{t_{lim,O}} \mathrm{d}t_1 \int_{t_{lim,Y}}^{t_{lim,O}} \mathrm{d}t_2 \int_{t_{lim,Y}}^{t_{lim,O}} \mathrm{d}t_3 \ \ldots \int_{t_{lim,Y}}^{t_{lim,O}} \mathrm{d}t_n \ 1}{\int_{t_{lim,Y}}^{t_{lim,O}} \mathrm{d}t_1 \int_{t_1}^{t_{lim,O}} \mathrm{d}t_2 \int_{t_2}^{t_{lim,O}} \mathrm{d}t_3 \ \ldots \int_{t_{n-1}}^{t_{lim,O}} \mathrm{d}t_n \ 1} = \frac{(t_{lim,O} - t_{lim,Y})^n}{\frac{(t_{lim,O} - t_{lim,Y})^n}{n!}} = n!$$

(The integration in the denominator can easily be evaluated using $\tau_i = t_{lim,O} - t_i$.)
Thus we get $J=n!$ when the prior is consistent with the actual order of the samples. This characterises the model-data agreement adequately, because $1/n!$ is the probability to get the right order of n ages by chance. Thus, it is reliable, that the agreement factor of the uniform prior, which defines one particular order, is $n!$ times higher (when consistent with the data) than that of the constant prior, which does not fix any order. Or in order words, if the real sample ages are actually in the order as defined by the prior, the probability to get the measured radiocarbon ages is $n!$ times higher than in a case, where they can have any order.

For sure, the characterisation of the domain function λ, as given by Equation 4.5, is not the only possible way. Different definitions (e.g. if one includes the regions between determined radiocarbon ages by an appropriate definition) can lead to different results for the agreement factor. However, this reflects just the fundamental problem, that unrestricted priors, which can not be standardised, are no probability densities with absolute meaning. Therefore, arbitrariness is unavoidable when applying an absolute measure as the prior-prediction.
It should be remembered at this point, that the domain function is exclusively used for the calculation of an agreement measure. The usual calculation of the resulting posterior marginals is performed in an unchanged way, without standardisation of the prior function.

Aside from the problems occurring from prior standardisation, the agreement factor J remains an absolute measure for the model or prior quality, because it delivers directly the increase of the probability to get the determined radiocarbon age set, caused by the prior information. However, in robust Bayesian analysis, which will be the application of the agreement factor in this work, this absolute character can be disadvantageous, due to reasons discussed the following section and in section 5.3.1. Therefore an alterative version of the agreement factor is defined too, which is related to a reference prior $a^{(ref)}$:

Equation 4.8:

$$J_{REL} = \frac{J}{J^{(ref)}} = \frac{v_{a\lambda}/v_{\lambda}}{v_{a^{(ref)}\lambda}/v_{\lambda}} = \frac{v_{a\lambda}}{v_{a^{(ref)}\lambda}} = \frac{\int_{vol} l \cdot a \cdot \lambda \ \mathrm{d}\mathbf{t} \ / \ \int_{vol} a \cdot \lambda \ \mathrm{d}\mathbf{t}}{\int_{vol} l \cdot a^{(ref)} \cdot \lambda \ \mathrm{d}\mathbf{t} \ / \ \int_{vol} a^{(ref)} \cdot \lambda \ \mathrm{d}\mathbf{t}}$$

4.3.2 Quantitative meaning

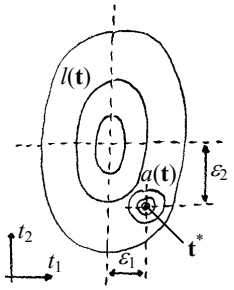

Figure 4.6: Illustration of the used 'ideal model' (two-dimensional case). The likelihood function $l(\mathbf{t})$ and the prior function $a(\mathbf{t})$ are shown. The latter is a delta-function or a narrow Gaussian around the real age set $\mathbf{t}^*$. The ε_i are the actual deviations of the measured radiocarbon ages from the real ages. (A linear calibration curve with slope one is assumed, so that time differences are equal on the real age and on the radiocarbon age scale.)

To get a quantitative rating of the values of the agreement factor J, a similar estimate is performed, as done in section 4.1.2 for the single sample based total agreement factor I_{Π}. Again we assume a linear calibration curve $c(t)$ with slope one, which generates a likelihood function $l(\mathbf{t})$, which is a multi-dimensional Gaussian. Thus, the widths of the Gaussian in the individual dimensions reflect just the measurement accuracies σ_i of the determined radiocarbon ages x_i.

$$l = \frac{1}{\sqrt{2\pi}\cdot\sigma_1} \cdot e^{-\frac{(x_1 - c(t_1))^2}{2\cdot\sigma_1^2}} \cdot \ \ldots \ \cdot \frac{1}{\sqrt{2\pi}\cdot\sigma_n} \cdot e^{-\frac{(x_n - c(t_n))^2}{2\cdot\sigma_n^2}}$$

According to Equation 4.5, the domain function based on this likelihood is:

$$\lambda = \left(\frac{1}{\sqrt{2\pi}\cdot\sigma_1} \cdot e^{-\frac{(x_1 - c(t_1))^2}{2\cdot\sigma_1^2}} + \ \ldots \ + \frac{1}{\sqrt{2\pi}\cdot\sigma_n} \cdot e^{-\frac{(x_n - c(t_1))^2}{2\cdot\sigma_n^2}} \right) \cdot \ \ldots \ \cdot$$

$$\cdot \ \ldots \ \cdot \left(\frac{1}{\sqrt{2\pi}\cdot\sigma_1} \cdot e^{-\frac{(x_1 - c(t_n))^2}{2\cdot\sigma_1^2}} + \ \ldots \ + \frac{1}{\sqrt{2\pi}\cdot\sigma_n} \cdot e^{-\frac{(x_n - c(t_n))^2}{2\cdot\sigma_n^2}} \right)$$

Again, an 'ideal' model (prior function) $a(\mathbf{t})$ is analysed, which determines the actual set of real ages $\mathbf{t}^*$ precisely. Particularly we assume a multi-dimensional δ-function (or Gaussian with very small sigmas) at $\mathbf{t}^*$:

$$a = \delta(\mathbf{t}-\mathbf{t}^*) \approx \frac{1}{\sqrt{2\pi}\cdot\varsigma}\cdot e^{-\frac{(t_1-t_1^*)^2}{2\cdot\varsigma^2}} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\cdot\varsigma}\cdot e^{-\frac{(t_n-t_n^*)^2}{2\cdot\varsigma^2}} \quad \text{with} \quad \varsigma << \{\sigma_1,\ldots,\sigma_n\}$$

Based on this three equations (which are already notated in their standardised form), the prior-prediction $v_{a\lambda}$ that includes the actual prior information, as defined in Equation 4.7, results in:

$$v_{a\lambda} = \frac{1}{\sqrt{2\pi}\cdot\sigma_1}\cdot e^{-\frac{(x_1-c(t_1^*))^2}{2\cdot\sigma_1^2}} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\cdot\sigma_n}\cdot e^{-\frac{(x_n-c(t_n^*))^2}{2\cdot\sigma_n^2}} = \frac{1}{\left(\sqrt{2\pi}\right)^n\cdot\prod_i\sigma_i}\cdot e^{-\frac{1}{2}\cdot\sum_i\left(\frac{\varepsilon_i}{\sigma_i}\right)^2}$$

$$\text{with} \quad \varepsilon_i = x_i - c(t_i^*)$$

Where the ε_i are the actual deviations of the radiocarbon ages from the real ages (see Figure 4.6). $v_{a\lambda}$ is independent of the shape of λ, because the latter acts on the narrow prior function just as a multiplication with a constant factor, which is eliminated by the standardisation term (see Equation 4.7) again. In contrast, the prior-prediction v_λ, that uses just the domain function as prior, is dependent on the domain function. The domain function λ itself is determined by the likelihood function. If the single sample likelihoods are clearly separated (as e.g. in Figure 4.5), the domain is largest, if they all overlap fully (all radiocarbon ages are nearly equal), the domain is smallest. Assuming equal accuracies for all measurements, the equations from above show directly the following results for these two extreme cases:

$$v_\lambda \approx \begin{cases} \frac{1}{\left(2\sqrt{\pi}\right)^n\cdot\sigma^n}\cdot\frac{1}{n^n} & \text{for clearly separated single sample likelihoods} \\ \frac{1}{\left(2\sqrt{\pi}\right)^n\cdot\sigma^n} & \text{for fully overlapping single sample likelihoods} \end{cases}$$

The value for v_λ in case of fully overlapping likelihoods with additional equal measurement accuracies, is actual the highest possible one, because only in this case, the domain function and the likelihood function are identical. Thus, one can find a lower limit for the agreement factor J (according to Equation 4.7):

$$J = \frac{v_{a\lambda}}{v_\lambda} \geq \frac{\frac{1}{\left(\sqrt{2\pi}\right)^n\cdot\sigma^n}}{\frac{1}{\left(2\cdot\sqrt{\pi}\right)^n\cdot\sigma^n}}\cdot e^{-\frac{1}{2}\cdot\sum_i\left(\frac{\varepsilon_i}{\sigma}\right)^2} = 2^{\frac{n}{2}}\cdot e^{-\frac{1}{2}\cdot\sum_i\left(\frac{\varepsilon_i}{\sigma}\right)^2}$$

This lower limit for the agreement factor J for the used 'ideal model' shows the same value as the single sample based total agreement index I_Π does, in case of the comparable 'ideal model' used; see section 4.1.2. The same considerations as in the mentioned section lead to a similar possible definition of a threshold level J^* here:

Equation 4.9: $$J^* = 2^{\frac{n}{2}}\cdot e^{-\frac{1}{2}\cdot\chi^2_{th}(P^*,n)}$$

Where χ_{th}^2 denotes that level for $\Sigma_i(\varepsilon_i/\sigma_i)^2$, which is statistically undershoot with a probability of P^*. Thus, in case of the 'ideal model', the agreement factor J is with a

probability of at least P^* higher than J^*. The reason, that the probability can be higher then P^* is the fact, that the used calculation for J is the just minimal possible.
As J^* is equal to I^*_Π (the threshold for the single sample based total agreement index I_Π; see section 4.1.2), numerical values for $J^{*\,1/\sqrt{n}}$ can be seen in Figure 4.3, where $I^{*\,1/\sqrt{n}}_\Pi$ is plotted.

It is meaningful to discuss the ratio of n^n that occurs between the two extreme cases within the calculation of v_λ, as described above, and is accordingly given for the agreement factor J itself too. The value n^n would also be the resulting agreement factor J for a sequence of clearly separated measurements, when a prior is applied, that models correctly the order of the ages and defines additionally, that the age differences of neighbouring ages have to be so high, that the ages cannot lie within a single peak of the domain function. It was already demonstrated in section 4.3.1, that a prior that defines the order alone results in a value $n!$, which is the reciprocal of the probability to get the right order by chance. The reciprocal of the probability, that each of the n real age lies exclusively within one of the n different possible age ranges (given by the domain function), is $n^n/n!$. The product of both is n^n, which is in accordance with the result for J.
The fact, that the factor n^n occurs between the cases with separated and overlapping measurements, is caused by restriction of the domain for the real ages to the small common range, defined by the equal measurements in the latter case. Actually, it is much less likely for a model to be consistent with the data, if the possible ranges for the real ages are primarily much wider (given for the separated case), and thus, if there is an agreement in this case, the agreement factor is accordingly higher.
It should be remembered at this point, that (different to section 4.1) the term 'agreement' is used here in a richer sense, than just to describe the consistence of model and data. It is rather an absolute measure for the model quality, based on the model's predicted probability for the actual determined radiocarbon age set.

In the application of the agreement factor in this work (see section 5.3.1), usually different priors are analysed that all base on common basic constraints, based on archaeological facts. One wants to discard priors that are in bad agreement with the measurements, but without considering the quality of the common constraints, or the total model quality in other words. Therefore, one relates the agreement factors to that of a reference prior that defines the used constraints in a basic form (e.g. the uniform prior for a sequence), and thus characterises the total model quality.
The uniform prior for a sequence results in an agreement factor $J=n!$ for clearly separated likelihoods and in $J=1$ for fully overlapping likelihoods. (The first was demonstrated in section 4.3.1 and the second results from the fact, that a $n!^{\text{th}}$ part is cut out of the likelihood by the prior, which compensates the $n!$-factor from the $a\cdot\lambda$ standardisation.) So, assuming this prior as reference prior and analysing the ideal model in relation to it, one can see that J_{REL} (defined by Equation 4.8) is smallest in the second case, similar then J itself, and it has moreover the same lower limit:

$$J_{REL,\,unif} \;=\; \frac{J}{J_{unif}} \;\geq\; 2^{\frac{n}{2}} \cdot e^{-\frac{1}{2}\cdot\sum_i\left(\frac{\varepsilon_i}{\sigma}\right)^2}$$

Thus, it is reasonable to use the same threshold as defined by Equation 4.9 also for the relative agreement factor:

Equation 4.10: $$J^*_{REL} = 2^{\frac{n}{2}} \cdot e^{-\frac{1}{2}\cdot \chi^2_{th}(P^*, n)}$$

Conclusively it should be recognised, that the definition of the artificial domain function λ (Equation 4.5), that is needed for all calculations above when dealing with unrestricted priors, remains somehow arbitrary, which is disadvantageous. So one could ask, why not take the likelihood function itself as domain function ($\lambda=l$), which is for sure the most objective definition. Of course, this is possible and results in an index already known; see next section. Although, this would simplify the considerations from above, there would be a significant loss of information about the model quality within the index too. This can be seen best, considering the example of a sequence with fully separated measurements again. The agreement factor distinguishes clearly between the uniform prior ($J=n!$) and the constant prior ($J=1$). Contrary, if λ is set equal to l, both priors would result in an agreement factor of one. Thus, the index is reduced to test the consistency of model and data and neglects the fact, that the uniform prior, which carries the whole information about the order of the ages, accords at a higher level with the data than the constant prior, that allows the ages to lie in any order. Generally spoken, the more an agreement index is reduced to test the consistency only, the more really informative models are discriminated, because their information content is ignored, and they have for sure a higher risk to show inconsistencies than models with very poor information content.

4.3.3 Relations to other indices

If we just formally generalise the definition of the single sample agreement index I (Equation 4.1) by using the multi-dimensional functions (similarly standardised on the real age space) instead of the single sample likelihoods and posterior marginals, and switch to multi-dimensional integration accordingly, one gets

$$\frac{\int_{vol} l \cdot p \ \mathrm{d}\mathbf{t}}{\int_{vol} l \cdot l \ \mathrm{d}\mathbf{t}} \ ,$$

which is an established index (noted as F_{model} by BRONK RAMSEY, 2009*a*). The agreement factor J defined above, would become equal to this index, by simplifying the definition of the domain function λ (Equation 4.5) to $\lambda=l$, as shown bellow. $\lambda=l$ changes J (defined by Equation 4.7) to:

$$J = \frac{\int_{vol} l \cdot \frac{a \cdot l}{\int_{vol} a \cdot l \ \mathrm{d}\mathbf{t}'} \ \mathrm{d}\mathbf{t}}{\int_{vol} l \cdot \frac{l}{\int_{vol} l \ \mathrm{d}\mathbf{t}'} \ \mathrm{d}\mathbf{t}}$$

The Bayesian theorem (see Equation 2.12) shows, that the term $a \cdot l \ / \int_{vol} a \cdot l \ \mathrm{d}\mathbf{t}$ is equal to the posterior function p, where p is already standardised on the real age space (see

section 2.3). Thus, J results in (for clarification '$^{\int=1}$' indicates densities standardised on the real age space)

$$J = \frac{\int_{vol} l \cdot p^{\int=1} \, \mathrm{d}\mathbf{t}}{\int_{vol} l \cdot l^{\int=1} \, \mathrm{d}\mathbf{t}} = \frac{\int_{vol} l^{\int=1} \cdot p^{\int=1} \, \mathrm{d}\mathbf{t}}{\int_{vol} l^{\int=1} \cdot l^{\int=1} \, \mathrm{d}\mathbf{t}} \quad ,$$

which is 'F_{model}' as defined above.

It should be finally noted, that if the model is reduced just to the one-dimensional case, $\lambda=l$ is valid without changing the definition of the domain function. And consequently, J results in

$$J = \frac{\int_{-\infty}^{+\infty} l_1^{\int=1} \cdot p_1^{\int=1} \, \mathrm{d}t}{\int_{-\infty}^{+\infty} l_1^{\int=1} \cdot p_1^{\int=1} \, \mathrm{d}t} \quad ,$$

which is equal to the single sample agreement index I, described in section 4.1.1.

4.4 DEVELOPMENT OF A GIBBS SAMPLING METHOD TO EVALUATE VOLUME INTEGRALS TO DETERMINE THE PRIOR-PREDICTION

To calculate a prior-prediction (Equations 4.4, second one), which is the base of the agreement factor discussed in the previous section, one has to evaluate a volume integral of dimension n, where n is the number of samples (or model parameters in general). For the same reason as for the basic calculations of the posterior marginals, the integration has to be done by a Monte Carlo method. (The problem is discussed at the beginning of section 2.4.) In the following a possible integration method is introduced that is based on the simple Gibbs sampling procedure, which was already chosen for the calculation of the posterior marginals too. However, there is a fundamental difference between the simple procedure to calculate the prior marginals (see section 2.4.1) and the calculation now: The counting of Gibbs sampled points can deliver only relative results (e.g. the shape of a posterior marginal is deduced from the relative numbers of sampled points related to various positions on the age axis). Absolute values as the volume integral of the whole function do not result by the original method. The developed method described below can overcome this problem.

4.4.1 The fundamental principle

The basic idea to get an absolute value for a function volume is, to compare the function with another function, which is simple enough to be integrated analytically. In principle the method is the following: A reference function - e.g. a constant function value on a hyper-spherical domain - is defined, and the sum function of this reference function and the original function is built (symbolised in Figure 4.7). Now, a Gibbs-sampling run is performed on this sum function. Thereby, a special procedure is executed: For each sampled point the value one is divided into two

portions, representing the ratio between the original function and the reference function at this point. Both values are added up separately over the whole sampling run. This leads - aside of statistical deviations and under the condition of convergence of the Gibbs sampling - to two sums that reflect the ratio of the volumes of original and reference function. Since the volume of the reference function can be calculated analytically, the volume for the original function follows directly.

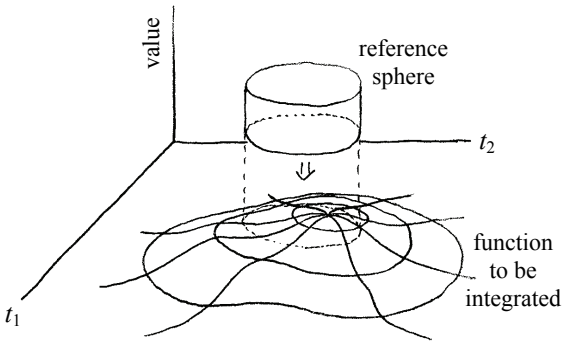

Figure 4.7: Adding a reference function to the function of interest. The reference function can e.g. be defined to show constant function value within a hyper-sphere, and zero outside. (In the shown two-dimensional case the hyper sphere is a circle.)

The statement, that Gibbs sampling on a sum function, with splitting each point into two parts proportional to the ratio of the to functions, leads to two sums, whose ratio is that of the two function volumes, has to be justified:
One denotes the two function as $f(\mathbf{t})$ and $g(\mathbf{t})$ with $\mathbf{t} = (t_1,...,t_n)$, and assumes a partition of the coordinate space defined in a way, that all parts are small enough to treat the function values as roughly constant within each part. Thus, $f(\mathbf{t}\in k)$ and $g(\mathbf{t}\in k)$ shall denote the function values corresponding to the k^{th} part of the partition. (One always speaks of functions with non-negative values and finite integral, which are at least partially overlapping.) $U_{(f)}$, $U_{(g)}$ and U shall be the function volumes (integrals) of f, g and f+g. Similarly $u_{k,(f)}$, $u_{k,(g)}$ and u_k, are the corresponding volumes of the k^{th} part of the partition, and $w_{k,(f)}$, $w_{k,(g)}$, and w_k, the proportionate and the total counts for this part. Thus, one obtains the following relations:

$$U_{(f)} = \sum_k u_{k,(f)} \qquad \text{(Ia)}$$
$$U_{(g)} = \sum_k u_{k,(g)} \qquad \text{(Ib)}$$
$$U = \sum_k u_k \qquad \text{(Ic)}$$
$$u_{k,(f)} / u_k = f(\mathbf{t}\in k) / (f(\mathbf{t}\in k)+g(\mathbf{t}\in k)) \qquad \text{(IIa)}$$
$$u_{k,(g)} / u_k = g(\mathbf{t}\in k) / (f(\mathbf{t}\in k)+g(\mathbf{t}\in k)) \qquad \text{(IIb)}$$
$$w_{k,(f)} / w_k = f(\mathbf{t}\in k) / (f(\mathbf{t}\in k)+g(\mathbf{t}\in k)) \qquad \text{(IIIa)}$$
$$w_{k,(g)} / w_k = g(\mathbf{t}\in k) / (f(\mathbf{t}\in k)+g(\mathbf{t}\in k)) \qquad \text{(IIIb)}$$
$$w_k = C \cdot u_k \qquad \text{(IV)}$$

The equations of the first group are clear; the second group follows trivially under the assumption of constant function values within each element of the partition. The third group of equations reflect the recipe to split the value one into two proportionate fractions at each point. Finally Equation IV describes the nature of Gibbs sampling to generate a point density proportional to the value of the sampled function, which is equivalent to generating a number of points proportional to the function volume in each partition element (the sampled function is the sum function f+g; C is the constant of proportionality). It has to be mentioned, that

Equation IV is influenced by statistical deviations and would be exactly true for an infinite number of counts only.
Combining IIa and IIIa gives

$$w_{k,(f)} / w_k = u_{k,(f)} / u_k \; ;$$

and combining this further with Equation IV leads to

$$w_{k,(f)} = C \cdot u_{k,(f)} \; .$$

Summing up all $w_{k,(f)}$, and using additionally Equation Ia, results in

$$\sum_k w_{k,(f)} = C \cdot U_{(f)} \; .$$

Similarly from IIb, IIIb, IV and Ib one derives

$$\sum_k w_{k,(g)} = C \cdot U_{(g)} \; .$$

Finally, combining the two latter equations delivers the relation that had to be verified:

$$\sum_k w_{k,(f)} / \sum_k w_{k,(g)} = U_{(f)} / U_{(g)} \; .$$

4.4.2 The actual procedure

In this section the particular realisation of the integration procedure, as it has been realised in the developed Matlab sequencing program, is described.

First step: Finding the centre, the function value and a first estimate of the radius for a reference hyper-sphere

As mentioned above, a simple reference function has to be built. Actually, it is realised within the program by defining a hyper-sphere with adequate centre and radius, and setting the reference function value to a particular constant level inside the sphere, and to zero outside. The reference function will be denoted g and the original function f in the following.
A practical choice for the centre of the hyper-sphere is the centroid of the original function. It is evaluated by performing a short Gibbs sampling run and summing up the components of all sampled points separately. (A point in the argument space will be denoted with $\mathbf{t} = (t_1,...,t_n)$ here, notwithstanding the fact, that there could be others than sample-age coordinates too.) Due to the fact, that the density of the sampled points is proportional to the local function value $f(\mathbf{t})$, the centroid $\boldsymbol{\beta} = (\beta_1,...,\beta_n)$ is simply:

$$\boldsymbol{\beta} = \frac{1}{w} \cdot \left(\sum_{j=1}^{w} t_1^{(j)}, \; \ldots \; , \sum_{j=1}^{w} t_n^{(j)} \right)$$

Where $t_i^{(j)}$ is the i^{th} component of the j^{th} of w sampled points.
It should be noted, that there are functions where the centroid is not an adequate choice for the centre of the reference sphere. Thus, the program includes a procedure to cope with this cases.

To achieve good convergence (see details in section 4.4.3), both the function volume and the base volume of the reference function should be of comparable extent with respect to those of the function to be integrated. (The term 'function volume' will always be used for the value of the volume integral, and the term 'base volume' denotes the volume of this part within the function-argument space, that generates the bulk of the function volume.) Thus the constant value ϕ for that part of the reference function g within the hyper-sphere, has to be chosen in a range, comparable to the characteristic values of the original function f. Actually, ϕ is set equal to the median of the function values, which occur within a sampling of the original function f. It can easily be calculated from the same points used already for the centroid above:

$$\phi \quad = \quad \text{median} \left(f(\mathbf{t}^{(j)}) \right)$$

Hereby, $f(\mathbf{t}^{(j)})$ is the function value of the j^{th} sampled point $\mathbf{t}^{(j)}$. The so found median is an adequate choice for the purpose of volume integration, because half of the function volume origins from values above ϕ and the other half from values below ϕ. The radius of the reference hyper-sphere will be adapted within an iteration process to find a reference function volume close to the original function volume (see next step below). A suitable starting value ρ_0 for this iteration is the radius of a sphere that divides the function volume into an outer and an inner half. To get this starting radius, all distances $\delta^{(j)}$ from the sampled points $\mathbf{t}^{(j)}$ to the centroid $\boldsymbol{\beta}$ are calculated:

$$\delta^{(j)} \quad = \quad \sqrt{\left(t_1^{(j)} - \beta_1\right)^2 + \ \ldots \ + \left(t_m^{(j)} - \beta_m\right)^2}$$

Finally, the starting radius results as median of the distances (because the same number of points inside and outside ρ_0 means same function volume inside and outside too):

$$\rho_0 \quad = \quad \text{median} \ (\delta^{(j)})$$

Second step: Finding a proper radius for the reference hyper-sphere

Now, the iteration to bring the reference function volume roughly to the value of the original function volume is performed. Therefore, the Gibbs-sampling is done on the sum of original function f and reference function g, counting the points in separate portions for the two functions, as described above. The iteration is started with a first short sampling run on the sum-function f+g, using the starting value ρ_0 for the reference sphere. This results in a particular ratio $w_{(f)}/w_{(g)}$ of the point-sum portions, which reflects the current rate of the volumes of the two functions. $w_{(f)}$ and $w_{(g)}$ are obtained as the following sums over all w sampled points on the sum-function:

Equations 4.11:

$$w_{(f)} = \sum_{j=1}^{w} \frac{f(\mathbf{t}^{(j)})}{f(\mathbf{t}^{(j)}) + g(\mathbf{t}^{(j)})} \qquad\qquad w_{(g)} = \sum_{j=1}^{w} \frac{g(\mathbf{t}^{(j)})}{f(\mathbf{t}^{(j)}) + g(\mathbf{t}^{(j)})}$$

(Numerically only one sum has to be evaluated because of w=$w_{(f)}$+$w_{(g)}$.) Next, ρ_0 is multiplied by the factor $(w_{(f)}/w_{(g)})^{1/n}$, where n is the dimension of $\mathbf{t}$. If the sampling

would be done with a huge number of points, the new ρ would already lead to a reference function with equal volume as the original function. However, for a bounded number of sampled points it is not unlikely, that the points are all sampled in regions where e.g. the reference function is zero, so that $w_{(f)}/w_{(g)}$ result in infinity. The reason therefore is, that the initial base volume of the reference function can be many orders of magnitude away from the base volume of the original function. To avoid a multiplication of the current radius with zero or infinity, the used value of $w_{(f)}/w_{(g)}$ is artificially limited upwards and downwards, and the short sampling process, together with the adjustment of ρ, is repeated as long as $w_{(f)}/w_{(g)}$ is not too far from the value one.

Third step: Final sampling run and calculation of the function volume

Closing, an accurate longer Gibbs-sampling run is performed on the sum-function $f+g$, using the resulting value of ρ from above for the reference sphere. This leads to a definite ratio of the point sums $w_{(f)}/w_{(g)}$ for the final choice of ρ, which is the conclusive ratio between the volume of the original function $\int_{vol} f(\mathbf{t})\,\mathrm{d}\mathbf{t}$ and that of the reference function $\int_{vol} g(\mathbf{t})\,\mathrm{d}\mathbf{t}$. The latter is given by the product of the constant function value within the hyper-sphere ϕ, which was fixed in the first step, and the volume of a hyper-sphere with dimension n and radius ρ (where n is the dimension of the function argument $\mathbf{t}$):

Equation 4.12:
$$\int_{vol} g(\mathbf{t})\,\mathrm{d}\mathbf{t} \;=\; \phi \cdot \frac{\pi^{\frac{n}{2}} \cdot r^n}{\Gamma\left(\frac{n}{2}+1\right)}$$

All needed values of the Gamma-function Γ result analytically from the following relations:

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}\ , \qquad \Gamma(1) = 1\ , \qquad \Gamma(x+1) = x \cdot \Gamma(x)$$

Finally, the volume of the original function f is deduced directly from the reference function volume and the point-sum ratio $w_{(f)}/w_{(g)}$:

Equation 4.13:
$$\int_{vol} f(\mathbf{t})\,\mathrm{d}\mathbf{t} \;=\; \frac{w_{(f)}}{w_{(g)}} \cdot \int_{vol} g(\mathbf{t})\,\mathrm{d}\mathbf{t} \;=\; \frac{w_{(f)}}{w_{(g)}} \cdot \phi \cdot \frac{\pi^{\frac{n}{2}} \cdot \rho^n}{\Gamma\left(\frac{n}{2}+1\right)}$$

The benefit of this Gibbs sampling based method is the fact, that parts of the function that add the same fraction to the function volume are also scanned by the same number of points on average, which leads to good statistics and enhances the accuracy. Additionally, one needs not to pay attention to a close bounding of the base space at which the function is defined, because zones with very small function values do not contribute in the sampling process. However, a serious drawback of the method are increasing convergence problems at higher dimensions, which are discussed in the next section.

4.4.3 Some remarks to convergence problems

Previous to discuss convergence problems, there is a clarification necessary: One could ask, why it is not possible to sample with uniformly distributed points, since Gibbs sampling suffers on convergence problems. This seems to be even more reasonable, as the volume integral would simply be the sum of the function values at all sampled points divided by the mean density of the point pattern (i.e. the number of points per coordinate-space volume). However, the following example demonstrates clearly, that this procedure is inadequate for multi-dimensional functions.
We look at a multi-dimensional Gaussian on an e.g. 100-dimensional space. Now we cut off the function at different radii by hyper spheres within the coordinate space. Comparing the volume of an inner hyper-sphere that covers the inner 40% of the function volume with that of an other sphere covering 80%, one finds, that the latter is about 2000 times higher than the former. (See the according expressions in the following paragraph.) Thus, if one would numerically integrate the Gaussian with uniform distributed points restricted on that 80% domain (restriction is necessary when using uniform distributed points), only one of 2000 sampled points would lie within the inner sphere at average, although it represents half of the function volume! In other words, one of 2000 points contributes to the sum of the function values as much as all the other points together. It is clear, that this will result in bad accuracy due to statistical deviations. In contrast, Gibbs sampling would produce an equal number of points within the two halves of the function volume at average, caused by the fact, that it produces points with a density proportional to the function value.
For completeness, see the analytic expressions for the function volume of an n-dimensional Gaussian, cut off at the radius r (which can be deduced by integrating the function-volume accretions of all hyper-spherical shells up to r):

$$1 - e^{-q} - e^{-q} \cdot \sum_{i=1}^{\frac{n}{2}-1} \frac{q^i}{\Gamma(i+1)} \quad \text{for even } n \qquad \frac{2}{\sqrt{\pi}} \cdot \int_0^{\sqrt{q}} e^{-z^2} \mathrm{d}z - e^{-q} \cdot \sum_{i=\frac{1}{2}}^{\frac{n}{2}-1} \frac{q^i}{\Gamma(i+1)} \quad \text{for odd } n$$

Where the shortcut $q=r^2/(2\sigma^2)$ is used, and the increment for i is 1 for both sums. The first term in the expression for odd n is the well-known error-function evaluated at $\sqrt{q}$. Additionally needed information on the Γ-function and on the volume of a hyper-sphere have been already given in the previous section.

So, as there are powerful reasons to use Gibbs-sampling, the convergence problems associated with this method will be discussed now. The essential problem when sampling the sum of original function and reference function, as described in section 4.4.1, is the following: Contrary to the low dimensional case (as suggested by Figure 4.7) the reference function may overlap with the original function only at a tiny fraction of its base volume. For sure, this could be avoided by enlarging the reference sphere as much as needed to cover the original function completely. Unfortunately, in the high dimensional case, that would make the base volume of the reference function by a huge factor larger than the base volume of the original function, except for functions with a very simple shaped base space. Subsequent, the large base volume of the reference function would result in the fact, that the Gibbs sampling process would need a huge number of iterations to jump to the overlapping region, when currently being outside, so that there is no convergence at a manageable

number of sampled points. This is even true if the constant function value of the reference function is set to that very small number, which makes its function volume equal to that of the original function.

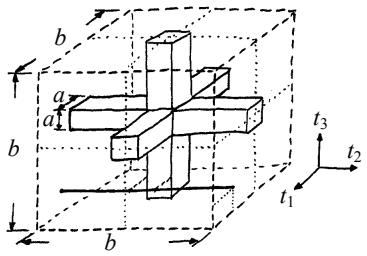

Figure 4.8: Illustration of the origin of the convergence problems. The cross in full lines indicates the space of the original function (constant value inside and zero outside), and the cube in dashed lines indicates the reference function (also with a constant value inside and zero outside). Note that this figure shows a three-dimensional argument space; there is no axis for the function values within this plot. The single line symbolises a one-dimensional cross section, used to draw the recent sampled point. The latter has to fall within an interval with length *a* at the centre of the cross section, to keep the chance to hit the cross with the next vertical cross section through the drawn point. Contrary to the three dimensional case, in a high dimensional case, even the next cross section will not hit the cross in general. One has to draw the points out of the centre interval nearly for all dimensions to hit the cross finally.

The described problem can be seen clearly by the following example. A *n*-dimensional function with constant value on a cross shaped base space, as shown by Figure 4.8 for the three-dimensional case, shall be integrated. The reference function is assumed to cover the original function completely. To simplify the considerations, a hyper-cube is used instead of a hyper-sphere. The thickness of the bars of the cross (denoted by *a*) shall be much smaller than the length, which is equal to the side length of the reference cube (denoted by *b*). Gibbs sampling draws its points out of one-dimensional cross sections through the sampled density, which is the sum function here. The cross sections are always located at the position of the last sampled point and step through all coordinate directions (see section 2.4.1). If the currently sampled point lies outside the original function (the cross), the next one-dimensional cross section (which would occur as e.g. horizontal line in Figure 4.8) will not hit the cross in general. To give the next cross section the possibility to hit the cross, the new point has to be drawn out of an interval with length *a* in the centre of the cross section. Since the cross section does not hit the cross, the probability density is constant along the whole cross section (determined by the reference function). Thus, the probability to draw a point from the needed interval would be a/b. Contrary to the low dimensional case as illustrated, in a high dimensional case the next cross section would not hit the cross in general again. Actually, one has to hit the centre interval blind (without attraction from the function value within the cross, which is usually much higher than the reference value) nearly for all coordinate directions, before drawing a point within the original function. Thus, the probability to hit the original function after stepping through all dimensions (one sampling cycle) is roughly $(a/b)^n$, independent of the chosen function value of the reference function. That means, that for e.g. $a/b=1/10$ and 30 dimensions, the sampling process would need the unacceptable number of about 10^{30} cycles to hit the original function.

Concluding this example shows clearly, that the reference function has to be defined very carefully. The chosen procedure described in the previous section would, in case of the recent example, work with a reference function (we remain thinking on a reference cube for simplicity, but actually it works with a sphere) that would have the

same function value as the original function and a comparable function volume (and also base volume), and would be placed around the centre of the cross. As the base volume of the original function is approximately $n \cdot b \cdot a^{n-1}$, the side length of the reference cube with equal volume would be $a \cdot (n \cdot b/a)^{1/n}$, which is $a \cdot 1.21$ for the numbers chosen above. Similar considerations than above give a probability to hit the function in one cycle of at least $(a/(a \cdot (n \cdot b/a)^{1/n}))^n = (a/b)/n = 1/300$, which is absolutely acceptable. It can also be shown that the probabilities to get into the branches outside the reference function and also back to the reference function outside the original function are also unproblematic.
Circumstances as illustrated by this example have been the motivation for establishing the special procedure of defining the reference function as described in section 4.4.2.

4.4.4 Some remarks to the performance

The accuracy of the integration method depends highly on the properties of the function. 'Compact' functions can be integrated easily, highly branching functions are challenging caused by convergence problems. Naturally, the problem increases with the dimension of the function. In principle, the accuracy can always be improved by using a higher number of sampled points. Unfortunately, if the convergence is really bad, the necessary number cannot be achieved in reality. In the following, the performance of the method is briefly characterised by three test functions.
A very compact and easy to integrate test function is an n-dimensional Gaussian (centred at an arbitrary point $\mathbf{t}^{(0)} = (t_1^{(0)},...,t_n^{(0)})$:

$$f(\mathbf{t}) = \frac{1}{\left(\sqrt{2\pi} \cdot \sigma\right)^n} \cdot \exp\left(-\frac{(t_1 - t_1^{(0)})^2 + ... + (t_n - t_n^{(0)})^2}{2 \cdot \sigma^2} \right)$$

The numerical procedure delivers the analytical value of one, up to a dimension of 100, with an accuracy of about 2% (standard deviation). This can be done with 2000 Gibbs cycles (for the final accurate run) or within about 2 minutes total calculation time, when running the Matlab program on a common personal computer. (One Gibbs cycle means, that for each dimension a position is drawn.)
A step more challenging for the procedure is the integration of a hyper-cube (constant value inside; zero outside), when using the hyper-spherical reference function, as usually done.

$$f(\mathbf{t}) = \left\{ \begin{array}{ll} f_0 & \text{for} \quad t^{(1)} \le t_i \le t^{(2)} \quad \text{for each dimension} \quad i = 1 \ldots n) \\ 0 & \text{else} \end{array} \right\}$$

In three dimensions a cube is not much less compact as a sphere, however, this is not true in the high dimensional case. There a hyper cube has already a 'branched' shape, which can be imagined, when looking at the fact that there is a factor of $\sqrt{50} \approx 7$ between the length of the diagonal and the side length for n=50 dimensions. In this 50-dimensional case the procedure delivers the analytical value of $f_0 \cdot (t^{(2)} - t^{(1)})^n$ with an accuracy of about 15% (standard deviation), using 5000 Gibbs cycles, which equates to a total runtime of about 10 minutes.

Finally, an example that is typical for the degree of difficulty that arises in the real application when calculating the prior-prediction, is the integration of the uniform sequence prior. For this example, the function is confined by a hyper-sphere with its centre at an arbitrary point with identical components $\mathbf{t}^{(c)} = (t^{(c)},...,t^{(c)})$ and an arbitrary radius r:

$$f(\mathbf{t}) = \left\{ \begin{array}{ll} 1 & \text{for} \quad t_1 < t_2 < \ \ldots \ < t_n \quad \text{and} \quad (t_1 - t^{(c)})^2 + \ \ldots \ + (t_n - t^{(c)})^2 \ \leq \ r \\ 0 & \text{else} \end{array} \right\}$$

The analytical value of the volume integral for this function is:

$$\frac{1}{n!} \cdot \frac{\pi^{\frac{n}{2}} \cdot r^n}{\Gamma\left(\frac{n}{2}+1\right)}$$

Where the second term is just the volume of the complete hyper-sphere (see the rear part of section 4.4.2), and the first term is the fraction that is cut out by the uniform prior constraints. (This was deduced in the rear part of section 4.3.1 within a cubic confinement, but the factor remains the same within a sphere. The reason therefore is the fact, that each single condition $t_i<t_{i+1}$ cuts both, the cube and the sphere into two symmetric parts, leading to an equal total fraction.) For 25 dimensions the function can be integrated with an accuracy of about 35%, when using 25000 Gibbs cycles, or a total runtime of about 30 minutes. So one can see clearly, that the latter function is already a challenge for the procedure; because the function is highly non-compact. Although the mathematical description of the constraints do not look complicated, one can imagine the non-compact shape of the function in the 25-dimensional case, when considering the fact, that the constraints cut out a fragment from the hyper-sphere with a volume fraction of only $6.4 \cdot 10^{-26}$ (1/25!), although it extends still from the centre to the surface of the sphere. It should be noted that the fineness of the structure leads to an additional problem in the numerical calculation on a discrete grid: The volume gets significantly different whether counting the surface points or not, except when using an extremely tiny division. Therefore, the prior for simple sequences (built-in in the developed program and used for this example) counts the surface points only fractionally to correct for the discrete grid.

The very different accuracies for these three examples can be explained by the very different degrees of overlap of original function and reference function, resulting in highly different convergence behaviours as well. The degree of overlap can be easily measured during the sampling of the sum function. For the uniform sequence prior (third example) the functions overlap only by about 0.1% (percentage of the volume of the sum function within the overlapping region), and the sampling stays at average for about 300 full Gibbs cycles at one function, before it changes to the other one. Naturally, this requires a huge number of cycles for good statistics.
The situation is already much better for the integration of the hyper-cube (second example), even though the number of dimensions is doubled: The degree of overlap is about 3% and there are just about 20 cycles needed at average to change between the functions.
Lastly, there is quite enough overlap for the integration of the Gaussian (first example; in this case the overlap can not be defined as simple as for the constant functions before), causing the sampling to change between the region inside and

outside the reference sphere more than once per cycle at average. This results in a very good convergence and correspondingly a high accuracy even for 100 dimensions.

The problems with the weak overlap of original and reference function suggested to use a reference function that can be adapted to the shape of the original function. Tests with a hyper-cuboid, whose side length are adapted adequately within the procedure, were performed. However, this effort does not result in a significant advantage, because the shapes of high dimensional functions are often too complex to be approximated by simple geometries. It turned out to be the best to stay with the hyper-spherical reference function, and find a good balance between a large overlap on one hand, and a not too large base volume of the reference function on the other hand. This is done by the procedure describe in section 4.4.2 and justified in section 4.4.3.

It should be noted, that there were additionally done some test to a complete different idea for evaluating volume integrals by Gibbs sampling: Since the reason that one can not simply add up the function values at the sampled points is the fact, that the point density is not constant, one could overcome this problem by using a measure for the local point density. A possible measure at any point could be its distance to the nearest neighbour. At average, the point density has to be proportional to $1/r_{next}^{n}$. Thus, the function volume is proportional to a value, resulting when summing the product $f \cdot r_{next}^{n}$ over all sampled points, because r_{next}^{n} compensates for the local point density. (The constant of proportionality depends only on the number of dimensions n, and has to be evaluated once for each n.) Although this method seems to be very simple and straightforward in principle, the tests did not show satisfying results, caused mainly by its very strong sensitivity to statistical fluctuations.

For sure, there are publications on integration of multi-dimensional functions using partly Markov chain Monte Carlo methods too. See e.g. OGATA (1989), CHEN and SCHMEISER (1996) or BAYARRI *et al.* (2006). In general two basic principles are mostly discussed, which are 'quasi-random integration' (see e.g. SOBOL, 1998 or TAKHTAMYSHEV *et al.*, 2007) and 'importance sampling' (see e.g. PETER-LEPAGE, 1978 or MOSKOWITZ and CAFLISCH, 1996). Quasi-random integration samples on a non-random grid with properties that enhance the integration result. Importance sampling uses a point distribution based on an 'importance function' which should approximate the shape of the original function and thus improve the result. There is a relation between the latter method and the method described above, as in some sense both methods base on the comparison of two functions. Although, the idea of importance sampling is to improve the non-Markov-Chain integration based on uniform distributed points by using more adequate point distributions. On the other hand, the method from above associates an absolute function volume with the number of Gibbs-sampled points just by a comparison with a reference function.
All in all, most of the methods from the literature are very sophisticated and not easy to use for non expert. Thus, the development of an integration method based just on the (slightly adapted) Gibbs-sapling procedure, which was already used for the actual sequencing calculations, seemed to be reasonable.

5 APPLYING 'ROBUST BAYESIAN ANALYSIS' TO IMPROVE BAYESIAN SEQUENCING

A fundamental problem within the procedure of Bayesian multi sample calibration (explained in chapter 2) is the transformation of the archaeological information to a prior function, which has to be a multi-dimensional probability density in the real age space (or in the parameter space, more generally). But usually, the archaeological facts do not determine the shape of this function in an unambiguous way. This means for example, that a given set of time relations deduced from a stratigraphy can be described by differently shaped prior functions. All possible different prior functions contain more or less quasi information, which is not based on the actual available information, but they do affect the resulting posterior density. Thus, the choice of a particular shaped prior function, which is required in Bayesian sequencing (as describes in chapter 2), inserts subjectivity into the method. The problem is explained in detail in section 3.1.
In the following, a modification of Bayesian sequencing will be developed, that may overcome, or more realistically reduce, this ambiguity problem. This effort is based on a principle within Bayesian statistics called 'robust Bayesian analysis', or more precise the achievement of 'global robustness' using multiple priors. The mathematical foundation is summarised e.g. by BERGER (1994) or RIOS INSUA and RUGGERI (2000), although there are ongoing developments up to now; see just for example SIVAGANESAN (1999), PEREZ *et al*. (2006), GRECO (2008) or O'NEILL (2009). It should also be mentioned that there is discussion on critical aspects of robust Bayesian analysis too (e.g. GELMAN, 2006), and 'objective Bayesian analysis' (using a single well defined prior function) has undoubted still its place (see e.g. BERGER, 2006). An overview over various different approaches to deal with the ambiguity problem in Bayesian statistics can be found by BERGER, 2000. Thus, since the theory of Bayesian statistics has become an unmanageable wide field for non-experts, the goal of this thesis is not a detailed theoretical approach. Rather the basic idea of robust Bayesian analysis is developed for its specific use within the application of multi-sample calibration of radiocarbon dates.

First, a guideline through chapter 5 is given, although there have to be used some terms that will be explained in detail later in the text below.
After illustrating the fundamental idea of robust Bayesian analysis and its most important fundamental difficulties in section 5.1, a chronology of developed actual realisations of the method is discussed in section 5.2. Although not all of the alternative procedures are relevant for practical use, their discussion is meaningful to clarify the relation of various mathematical approaches, including new ones as well as established ones.
Previous to the description of the different approaches, section 5.2.1 introduces a new way of plotting the so called 'highest-posterior-density ranges' simultaneously for all possible confidence levels. The use of these ranges is necessary, because some of the procedures discussed below do not result in probability densities, they are directly built by a unification of these 'highest-posterior-density ranges'.

Section 5.2.2 ('Approach I') describes the initial developments performed to realise robust Bayesian sequencing, which are based on the unification of highest posterior density intervals of a set of individual prior functions. To get rid of 'corrupt prior functions' (that are roughly speaking priors that are inconsistent with the data) the final result is deduced as the best possible of a series got by iterative elimination of prior functions from the set. Thereby the prior functions are excluded in order of their degree of disagreement with the data, and the confidence levels of the reduced intermediate results are lowered to consider the increasing probability of excluding the 'correct prior' too. Methods of this kind seemed to be the most direct realisations of robust analysis including a reliable treatment of corrupt prior functions. However it turned out, that the most generalised form of this methods is equivalent to analysing just the weighted sum of the various posterior marginals corresponding to the different priors.
Weighted sums of the posterior marginals can be calculated alternatively very convenient by using parametric prior functions, which is described in Section 5.2.3 ('Approach II'). Hereby the prior function shape is varied by defining the prior function depending on a set of free parameters, which is included in the Gibbs sampling process. This method is a specific relation of a commonly known principle called 'model averaging'. The underlying mathematics show that the method intrinsically weights the contribution from the various priors with their 'prior prediction', whose fundamental meaning has been already discussed.
Section 5.2.4 illustrates that the results of Approach II depend on the chosen parameter scale, which is in an analogues form also true for Approach I. Thus, in the second part of the section an idea of finding a 'balanced parameter scale' is developed, which provides an optimised scale that realises an evenly rating of the various prior shapes in some sense.
Section 5.2.5 highlights a very fundamental point, which was not clear when starting the investigations to this thesis, and it seems to be stressed first time explicitly, although the underlying mathematical relations are not complex: The result found by a calculation varying the prior function by including prior parameters within the Gibbs sampling process (Approach II), can be also found by using just a single specific prior function, which is an 'effective prior' describing the entity of all included prior shapes. Since it is also shown that Approach I and II become equivalent under specific conditions, one stays still with a single-prior solutions up to here. Thus the concluding decision was to turn away from weighting prior functions and develop an un-weighted unification of the various resulting highest posterior density ranges from the individual priors, where corrupt prior functions are just excluded if they fall below a threshold level of a well defined agreement measure (Approach III, section 5.2.6). The latter is the actual used method and is described in detail in section 5.3. Section 5.4 gives some important fundamental clarifications related to the developed realisation of robust analysis.
Finally, a completely different method was developed in this thesis (denoted 'overlap method'), which is described in section 5.5. It is a pragmatic approach that - roughly speaking -suppresses the sequencing effect within ranges where the single-sample likelihood functions overlap, because in these regions the result depends strongly on the chosen prior function. This method can be seen as an approximation for robust analysis, offering the advantage of much shorter run-time.

5.1 THE BASIC IDEA AND THE MAIN PROBLEMS

The basic idea of robust Bayesian analysis is to use a theoretically infinite set of prior functions, including all possible function shapes that are consistent with the available archaeological information. The Bayesian calculations are subsequently performed with each prior function individually, and the final result is the union of the individual results, exactly spoken the union of the highest-posterior-density ranges at an arbitrary confidence level (see the exact definition of these hpd-ranges in section 5.2.1). By this, one gets hpd-ranges for the sample ages that are valid for all possible prior functions and therefore independent of a subjective choice of a particular function.

For the simple example of two ages with known order the infinite set of various prior functions is symbolised in Figure 5.1: The probability density for wrong ordered ages is zero, the density for right ordered ages can have various shapes (see the explanations in section 3.1). An example that illustrates the unification of the hpd-ranges resulting from the set of various priors (at any confidence level) is given in section 5.2.1.

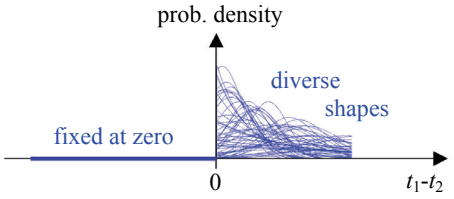

Figure 5.1: Symbolic illustration of an infinite set of prior functions with various possible shapes. All priors are consistent with the prior information that sample 1 is older than sample 2, i.e. t_1-$t_2 > 0$. (To simplify the plot, the actual two-dimensional prior functions are restricted to one-dimensional functions on the age difference.)

The described specific kind of robust analysis is based on BERGER (1994; section 1.3 and 4.1). There, the minimum and the maximum of a 'posterior quantity of interest' is calculated for a prior set, to get a robust conclusion. When dealing with limits of hpd-ranges as in this application, the use of the union of all hpd-ranges is a possible implementation of this method.

Although the method is very simple in principle, there are difficulties in the mathematical concept that have to be solved.
First, the strict application of the concept to use all priors that are consistent with the prior information leads to useless results. This is because generally one can always find extreme prior shapes that, although still consistent with the given information, can damage the result by producing posterior probabilities that are in disagreement with the measurements. These are priors that differ strongly from the unknown actual probability density for various real sample age combinations, resulting from the archaeological circumstances. For example, we think again on the simple case of two ages with known order. Let us assume, that the unknown true density for the age difference of the samples actually decreases exponentially with a mean value of e.g. 100 years, and it is zero for wrong ordered ages. Thus, the age difference of the real ages of two particular samples would occur between zero and some hundred years, resulting in correspondent measurements. However, the known prior information is only the order of the samples. Thus, the set of all possible priors would also include a prior, which e.g. gives a very high probability density for an age difference between

1000 and 1100 years and a very low density for all other age differences with correct ordered sign and a zero density for wrong ordered ages. Consequently, this prior will force the posterior density of one of the samples to lie far away from the measurement, although this prior is still in agreement with the used prior information concerning the temporal order exclusively. For sure, the mentioned prior is a highly artificial one, but even a prior that uses an exponential decreasing function, as assumed to be the correct one, will damage the result, if the slope is in strong disagreement with that of the actual density.
Concluding, as the true density of the real age combinations is not known, the only reliable way to identify such 'corrupt' prior functions, is to focus on their agreement with the measured data. If one assumes, that the radiocarbon ages are determined correctly, this is a reasonable criterion. The different approaches of applying robust analysis, as discussed in the following sections, will deal with corrupt prior functions in different ways, however, always based on the prior-data agreement.

A second serious problem for the kind of robust analysis as introduced above is the fact, that it is naturally not possible to consider an infinite number of different prior functions in the actual calculations. In principle there are two ways to deal with this problem: One way is to use a finite set of functions, that approximates the infinite set with adequate accuracy. Pragmatic approaches to define such finite prior sets are discussed in section 5.3.2. A second way is to use prior functions including parameters that can be varied directly within the sampling process. In that case one uses an infinite set of functions somehow. However, to simulate really all possible priors one would still need an infinite number of parameters too. Approach I and III from below use a discrete finite prior set, approach II uses the latter parametric set.

5.2 DIFFERENCES AND SIMILARITIES OF VARIOUS THINKABLE APPROACHES

Before starting this topic, it is necessary to have a look at the definition of hpd-ranges, which are essential for the framework of robust analysis. Additionally a generalised representation of hpd-ranges is introduced.

5.2.1 Highest posterior density ranges and 'hpd-ranges envelopes'

The highest posterior density range (hpd-range) of a marginal posterior distribution at a given confidence level κ is defined as illustrated by Figure 5.2. The range includes all parts of the distribution that exceed a particular probability density p^*. The latter has to be chosen at a level, so that the integral probability that is cut out from the distribution by the range, gives just the confidence level κ. Naturally, the so defined range needs not to be uninterrupted, but can split up in an arbitrary number of parts. It is obvious, that this is the most reasonable definition for a range that represents the given total probability κ.

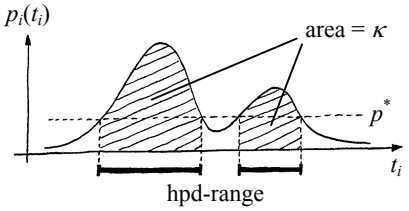

Figure 5.2: Definition of a highest posterior density range of a posterior marginal $p_i(t_i)$. The value p^* is set at a level, so that the probability represented by the hatched area is equal to the chosen confidence level κ.

As introduced in section 5.1, when performing robust analysis the hpd-ranges of various possible priors are unified. It is important to mind, that the meaning of a unified hpd-range is different to that of the single one. The single hpd-range is a range that includes the real value (sample age) with a probability of κ. In contrast, a unified hpd-range includes the real value with a probability of at least κ, but possibly with higher probability. Of course, the statement for the single range is only true, if the correct prior was used; the statement for the unified range from robust analysis is true, if the correct prior was among the used prior set. (The meaning of the term 'correct prior' is discussed in more detail in section 5.4.3).

A disadvantage of working with hpd-ranges is the restriction to a particular confidence level. However, it is not really necessary to focus on a particular value for the confidence level. It is also possible to give all hpd-ranges to any confidence level between one and zero simultaneously. The whole information can be simply expressed by a single curve that is the envelope of all hpd-ranges, where the latter are plotted at a position on a linear scale that corresponds to their confidenc level. Figure 5.3 illustrates the definition of such a 'hpd-ranges envelope'.
The numerical calculation of an envelope does not need explicit calculations of single hpd-ranges at particular confidence levels. The procedure is simple: One starts at the highest value of the considered probability density, and plots the first point of the future envelope at this age position and at a confidence level that is equal to the density value at this position. (The density value can be directly associated with a probability value, because in a discrete representation of a probability density, each point represents a well defined fraction of the total probability one.) Next, one continues with the density point next in height, but now the point of the envelope is plotted at a level equal to the sum of the recent and the previous processed density value. Continuing this procedure repeatedly, always using the sum of all previous density values, results finally in the hpd-range envelope, as shown in Figure 5.3.

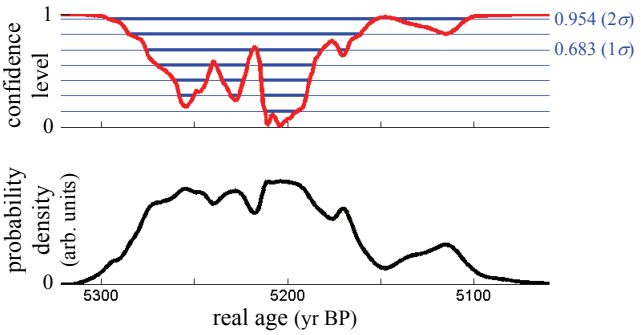

Figure 5.3: The 'hpd-ranges envelope' shows simultaneously the highest posterior density ranges to any confidence level. The curve (red) envelopes the hpd-ranges at all different confidence levels. For clarification hpd-ranges to some particular levels (thin blue lines) are also given (thick blue lines). The lower plot shows the original marginal posterior density.

It should be noted, that the definition of a hpd-ranges envelope, as well as the definition of the hpd-range itself, can be analogously used for any probability densities, not just for posterior marginals.

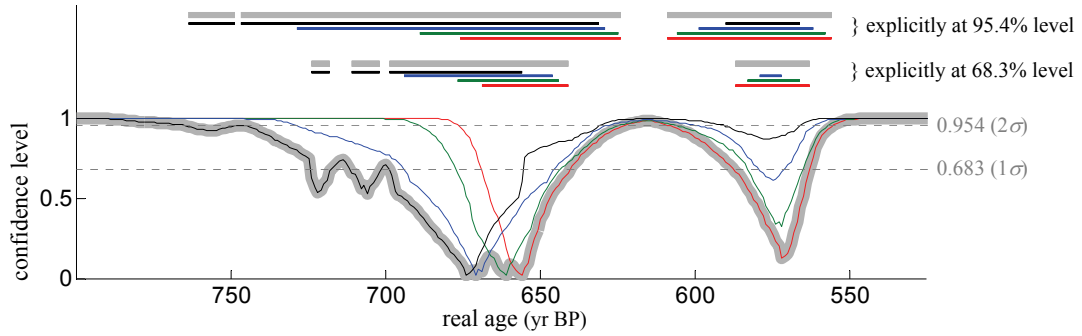

Figure 5.4: Unification of hpd-ranges simultaneously at any confidence level. The example shows (for a particular sample of a sequence) the hpd-ranges envelopes, calculated for four different posterior marginals (red, green, blue and black curves) based on for different priors. The hpd-ranges envelope for the unified hpd-ranges is again the envelope of the found set of curves, shown by the broad grey curve. For clarification, the unification of the hpd-ranges is given explicitly at two particular confidence levels (one and two sigma levels).

The union of hpd-ranges of a set of various marginal posteriors, which is evaluated in robust analysis, can also be performed simultaneously for any confidence level with the help of the hpd-ranges envelopes. The envelope for the unified ranges is constructed by using, at each position on the age axis, just the lowermost of all calculated different single hpd-ranges envelopes, as demonstrated in Figure 5.4.

5.2.2 Approach I: range unification by progressive elimination of priors

As introduced at the beginning of section 5.1, the basic principle of robust analysis is to unify the hpd-rages of the posterior marginals evaluated with various possible prior functions. It was discussed further, that this basic idea can not be performed directly, because corrupt prior shapes would damage the result. A thinkable approach to get rid of corrupt priors is discussed in this section.
The idea is to eliminate more and more priors from the set, starting with this, which is indicated as the most problematic by its model data agreement. The latter can be measured by any method discussed in section 4. The procedure is illustrated by Figure 5.5, keeping the example already used in the previous section. This example assumes a known temporal order (t_1 older than t_2) of two samples, measured at radiocarbon ages of x_1 = 730±55 BP and x_2 = 595±50 BP. To get a clear figure, a set of only four priors is used, including the uniform prior (const. for $t_1>t_2$ and zero else) and three exponential decreasing priors (exp $-(t_1-t_2)/\alpha$ for $t_1>t_2$ and zero else, with α of 40, 10 and 2.5 years). The procedure is shown for the posterior marginals of the first sample and would be the same for the other sample or any sample of a longer sequence.
One starts with the unification of the hpd-ranges envelopes, using initially the complete set of priors, given by the uppermost curve in Figure 5.5, which is identically with the broad grey curve of Figure 5.4 that shows its construction. Next, one removes the hpd-ranges envelope generated by the prior with the lowest

agreement (the red curve in Figure 5.4, which is generated by the fourth prior with α=2.5yr) and builds a unification of the remaining curves. Additionally the resulting curve is generally reduced on its confidence-level scale by a factor equal to the sum of all remaining agreement indices (after standardising the total sum of all indices to one). Here, the standardised agreement value of the rejected prior is 0.14, thus the curve is multiplied by 0.86 (the 'agreement factor' as introduced by Equation 4.7 in section 4.3 is used). The so reduced curve is shown as the second one within Figure 5.5. The idea of this reduction is, to consider the possibility that the rejected prior is quite correct, and his low agreement is not caused by inadequate modelling, but by e.g. statistical reasons. So the reduction is an estimate of the probability, that the removed prior is actually not corrupt and thus was removed unjustified. Next step the hpd-ranges envelope according to the prior with the second-lowest agreement (0.18) is removed from the set too (green curve in Figure 5.4, generated by the third prior with α=10yr), and the unification of the remaining hpd-ranges is multiplied by the sum of the remaining indices again (third curve in Figure 5.5; the sum is 0.68). This procedure is continued until all priors are removed, except that with the highest agreement.

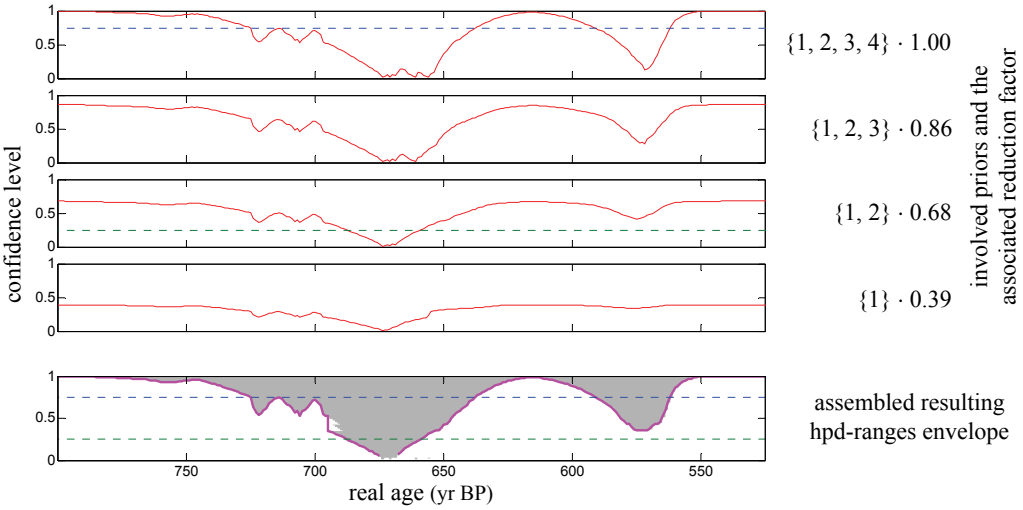

Figure 5.5: Hpd-ranges unification by progressive elimination of priors. The uppermost curve shows the unification of all hpd-ranges envelopes according to all priors within the set. In the following curves priors are successively removed from the set, starting with that with the lowest agreement with the data. The used 'agreement factor' results in values of 0.39, 0.29, 0.18 and 0.14 for the priors 1 to 4, when standardised to a total sum of one. Each curve is reduced in its confidence level by a factor equal to the sum of the agreements of the included priors. Finally, a resulting envelope is constructed by using at each confidence level the shortest hpd-range of all curves. For example, the resulting range at confidence level of 0.75 originates from the uppermost envelope (indicated by the dashed blue lines), and that at 0.25 originates from the third envelope (dashed green lines).

In a final step a resulting envelope is assembled out of all these curves by selecting, at any confidence level, the shortest (total length) hpd-range of all curves. It is permitted to take the shortest one, because any of these ranges provides the required confidence level. The composition of all shortest hpd-ranges is illustrated by the grey area within the last plot in Figure 5.5. For clarification the dashed blue and green lines illustrate, that, at a confidence level of 0.75, the shortest range originates from the first plot (with all priors included), and at 0.25, the shortest range originates from the third plot (two priors excluded). At the transitions from one to another curve the

so constructed envelope can be non-monotonic, as it can be seen twice at about 690 BP. The final hpd-range envelope (purple curve) may be artificially brought to a monotonic shape as illustrated.
In the presented example (which is close to real applications), even the lowest occurring agreement factor is not extremely bad. Therefore the difference between the initial unified hpd-ranges envelope and the resulting envelope from the procedure is not dramatic, but significant. However, it is obvious that the contribution of an extreme prior with a very low agreement factor is nearly fully suppressed by the method, because it is removed from the set without a significant reduction on the confidence-level scale for the remaining envelope.

The demonstrated method seems to be an adequate approach to realise the fundamental idea of the unification of the hpd-ranges of various possible priors, which includes additionally a mechanism to remove corrupt priors gradually. However, the method is somehow arbitrary and can be generalised as described in the following.
Instead of removing successively the contributions of the various priors, one can (theoretically) test all thinkable combinations of hpd-ranges generated by the individual priors, and this at arbitrary confidence levels, chosen individually for each prior. Denoting the individual chosen confidence level for the posterior marginal of the prior j with κ_j and the standardised agreement factor of prior j with η_j, a resulting combined hpd-range for an arbitrary combination is put at a resulting confidence level of $\sum_j \kappa_j \eta_j$. This is done in that way, because the probability, that the real age lies within the hpd-range of an individual prior, is (at least) the confidence level multiplied by the probability that the prior is the correct one, which is approximated by η_j. (See again the exact meaning of the term 'correct prior' in section 5.4.3). Thus, the probability, that the real age lies within the union of these ranges, is at least $\sum_j \kappa_j \eta_j$. Out of the various unified hpd-ranges, generated by the (theoretically infinite number of) combinations, the shortest is chosen at any confidence level. All these shortest ranges together build the final combined hpd-ranges envelope. It should be noted, that this procedure includes still the possibility, that contributions of individual priors are completely removed, which is realised by the case of $\kappa_j=0$.
An intermediate step of this generalisation, which can be still handled numerically, is implemented in the developed Matlab program.

A further generalisation can be performed by testing all combinations, not just by using the highest posterior density ranges, but all (theoretically) thinkable ranges that cover the total probabilities, which are chosen as confidence levels. Thus, additionally to the arbitrary choices of the individual confidence levels κ_j, each κ_j is realised by an infinite number of actual ranges. The rest of the procedure remains the same: The combined ranges are put at a resulting confidence level of $\sum_j \kappa_j \eta_j$, and the shortest range is chosen for each confidence level.
It is meaningful to analyse this very general procedure in detail: Since one can use any arbitrary range for each posterior marginal, it is obvious, that the shortest unified ranges will generally origin from completely overlapping individual ranges. This is clear, because for a just partially overlapping unified range one can extend all individual ranges to the unified extent, and thus increase the confidence level for the unified range without extending the unified range itself. So focusing on completely

overlapping ranges, the resulting confidence level for the unified range $\sum_j \kappa_j \eta_j$ can be expressed as below. The posterior marginal for the sample i, calculated with the prior j, is denoted as $p_{i,j}$, and κ_j is written more completely as $\kappa_{k(i),j}$, where $k(i)$ indicates one specific set of ranges, chosen for the various posterior marginals of the sample i originating from the different priors j:

$$\kappa_{k(i)} = \sum_j \kappa_{k(i),j} \cdot \eta_j = \sum_j \left(\eta_j \cdot \int_{ll(k(i))}^{ul(k(i))} p_{i,j} \, \mathrm{d}t_i \right) = \int_{ll(k(i))}^{ul(k(i))} \left(\sum_j \eta_j \cdot p_{i,j} \right) \mathrm{d}t_i$$

$\kappa_{k(i)}$ denotes the resulting confidence level for the unified range for a specific set of individual ranges k for the sample i. $ll(k(i))$ and $ul(k(i))$ are the chosen lower and upper limits of the individual ranges, which are the same for any prior j, because we consider completely overlapping ranges only. Notwithstanding the used notation with an integral over a continues range, discontinuous ranges are also permitted, and could easily be considered in the relations with a sum of integrals for the various parts. This generalisation would not alter the following conclusion.
Since the sum $\sum_j \eta_j p_{i,j}$ is the weighted sum over the posterior marginals, weighted with the agreement factors η_j, one can denote this weighted-mean posterior marginal with p_i, and gets:

$$\kappa_{k(i)} = \int_{ll(k(i))}^{ul(k(i))} p_i \, \mathrm{d}t_i = \kappa^*_{k(i)}$$

Where the resulting integral gives just the confidence level resulting directly from the mean marginal p_i, which is denoted with $\kappa^*_{k(i)}$. Considering the equivalence of $\kappa_{k(i)}$ and $\kappa^*_{k(i)}$ and the fact, that finally this combination of individual ranges $k(i)$ (equal ranges for all priors in case of complete overlap) is selected, for any resulting confidence level, that delivers the shortest unified range (equal to the individual ranges in case of complete overlap), the resulting ranges must be the hpd-ranges of p_i again, because these are, at any particular confidence level, the shortest possible ones.
Concluding it turns out, that in its most generalised form, the introduced method is equivalent to the use of hpd-ranges on the weighted sum of the posterior marginals. For sure, the latter procedure is much easier to perform than the initial idea. A kind of weighted summation can be done very directly, as demonstrated in the next section.

5.2.3 Approach II: free prior parameters within the Bayes model

A method that allows actually the handling of an infinite set of priors is discussed now. The obvious idea is, to describe the various priors with a single prior function that includes a set of free parameters. These parameters are then treated as additional dimensions within the Gibbs sampling, and thus various prior shapes are simulated simultaneously. So the prior function is defined one an $n+\eta$-dimensional space with n age coordinates and η free prior parameters. The Gibbs sampling process is performed on that $n+\eta$-dimensional space. For simplification the first n coordinates are denoted as sample age coordinates (t_1,...,t_n), but generally they can also include

non-sample age values (e.g. phase boundaries and so on), as described in section 2.6. Denoting the prior parameters with (α_1,...,α_η), the prior density can be written as:

$$a(t_1, \ldots, t_n, \alpha_1, \ldots, \alpha_\eta) = a(\mathbf{t}, \boldsymbol{\alpha})$$

With this prior the $n+\eta$-dimensional posterior function $p(\mathbf{t},\boldsymbol{\alpha})$ is calculated with the help of the Bayesian theorem:

Equation 5.1:
$$p(\mathbf{t}, \boldsymbol{\alpha}) = \frac{l(\mathbf{t}) \cdot a(\mathbf{t}, \boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t}, \boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t} \, \mathrm{d}\boldsymbol{\alpha}}$$

Again (as frequently above) the simplified notation without indicating the set of radiocarbon ages is used, because in practice the calculations are preformed exclusively for the particular determined set of radiocarbon ages. Finally, the resulting posterior density $p(\mathbf{t})$ is determined by integrating $p(\mathbf{t},\boldsymbol{\alpha})$ along all parameter dimensions, or projecting $p(\mathbf{t},\boldsymbol{\alpha})$ back to the sample-age sub-space in other words:

Equation 5.2:
$$p(\mathbf{t}) = \int_{vol} p(\mathbf{t}, \boldsymbol{\alpha}) \; \mathrm{d}\boldsymbol{\alpha}$$

Posterior marginals for the individual ages $p_i(t_i)$ can then be calculated based on $p(\mathbf{t})$, using the same equation as for the basic sequencing procedure (Equation 2.9), or directly based on $p(\mathbf{t},\boldsymbol{\alpha})$, which will practically be done within the sampling process, and leads to the same result. Additionally, $p(\mathbf{t},\boldsymbol{\alpha})$ offers the possibility to calculate marginal posterior densities for the individual prior parameters too. Further one can project $p(\mathbf{t},\boldsymbol{\alpha})$ to the parameter sub-space by:

$$p(\boldsymbol{\alpha}) = \int_{vol} p(\mathbf{t}, \boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t}$$

If the prior $a(\mathbf{t},\boldsymbol{\alpha})$ achieves the condition that $\int_{vol} a(\mathbf{t},\boldsymbol{\alpha}) \, \mathrm{d}\mathbf{t}$ is constant, which means equal weighting for the individual prior shapes (aside from the scaling problem, addressed below again), $p(\boldsymbol{\alpha})$ can be seen as the probability density for the various prior shapes according to $\boldsymbol{\alpha}$, based on their agreement with the measurements. This can be shown directly by expressing $p(\mathbf{t},\boldsymbol{\alpha})$ explicitly, leading to:

$$p(\boldsymbol{\alpha}) = \int_{vol} \left(\frac{l(\mathbf{t}) \cdot a(\mathbf{t}, \boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t}, \boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t} \, \mathrm{d}\boldsymbol{\alpha}} \right) \mathrm{d}\mathbf{t} \quad \propto \quad \int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t}, \boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t}$$

Where the last expression is (under the condition from above) proportional to the prior-prediction, in case of executing the Bayesian theorem for each prior individually, which is in fact a measure for the agreement, as discussed in section 4.2. However, the meaning of $p(\boldsymbol{\alpha})$ has always be seen in the light of a scaling problem discussed in section 5.2.4.

The latter considerations indicate, that there is an alterative view on the described procedure, which can clarify its meaning: Equation 5.2 can be seen as a weighted sum of posteriors within the normal space of sample ages, when the parameters are not treated as statistical variables for the Bayesian theorem. The posteriors on the

usual sample-age space, calculated with a particular parameter set $\boldsymbol{\alpha}$, shall be denoted as $p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha})$. The weighting factors that occur in the sum are the prior-predictions ($\int_{vol} l{\cdot}a\ \mathrm{d}\mathbf{t}$; see again section 4.2) calculated within the normal sample age space (see the prove subsequently), and thus, the method has capability to suppress corrupt prior shapes intrinsically.
In the sample age space the Bayesian theorem can be written as:

Equation 5.3:
$$p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) = \frac{l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})\ \mathrm{d}\mathbf{t}}$$

In difference to Equation 5.1 there is no integration over the prior parameters to calculate the prior-prediction in the sample-age space, the prior-prediction remains parameter dependent. Combining Equation 5.1 and Equation 5.3 one gets:

$$p(\mathbf{t},\boldsymbol{\alpha}) = p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) \cdot \frac{\int_{vol} l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})\ \mathrm{d}\mathbf{t}}{\int_{vol} l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})\ \mathrm{d}\mathbf{t}\,\mathrm{d}\boldsymbol{\alpha}} = p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) \cdot \frac{v^{(\mathbf{t})}(\boldsymbol{\alpha})}{v}$$

Where $v^{(\mathbf{t})}(\boldsymbol{\alpha})$ is the parameter dependent prior-prediction within the normal age space, and v is the prior-prediction in the combined age and parameter space, which is a constant independent of $\boldsymbol{\alpha}$. Putting latter relation into Equation 5.2, one gets an integration over the individual posteriors weighted with $v^{(\mathbf{t})}(\boldsymbol{\alpha})$, what had to be shown:

Equation 5.4:
$$p(\mathbf{t}) = \int_{vol} p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) \cdot \frac{v^{(\mathbf{t})}(\boldsymbol{\alpha})}{v}\ \mathrm{d}\boldsymbol{\alpha}$$

That means further, that also the posterior marginals $p_i(t_i)$ are equal to the weighted sum of the posterior marginals calculated for an individual prior, denoted with $p_i^{(\mathbf{t})}(t_i,\alpha)$:

Equation 5.5:
$$p_i(t_i) = \int_{vol} p_i^{(\mathbf{t})}(t_i,\boldsymbol{\alpha}) \cdot \frac{v^{(\mathbf{t})}(\boldsymbol{\alpha})}{v}\ \mathrm{d}\boldsymbol{\alpha}$$

This results directly by integrating both sides of Equation 5.4 over all age coordinates but t_i, according to the definition of posterior marginals (see Equation 2.9).
Of course, when interpreting the latter two equations, one has to be aware again, that the weighting depends on the arbitrary definable scaling of the parameters $\boldsymbol{\alpha}$, which will be discussed more detailed in section 5.2.4.
It should be further noted at this point, that the quantities analysed above (in practical use especially the marginals $p_i(t_i)$ and $p_i^{(\mathbf{t})}(t_i,\alpha)$), which were deduced by the use of the complete (standardised) Bayesian theorem, can be calculated with the reduced theorem (without using the standardisation integral, as done for the basic sequencing method) and standardised subsequently, without changing the relations from above.

Finally it should be mentioned that Equation 5.4 shows, that the method described in this section is a specific realisation of an already used principle, denoted as 'model

averaging' (see e.g. HOETING *et al*., 1999), which deals in general with the weighted averaging of resulting posteriors based on various models or priors.

5.2.4 The scaling problem and a well defined parameter scale

Both methods, approach I and approach II (which are equivalent in principle; see section 5.2.5) are affected by a scaling problem, which will be analysed here.

In approach I individual prior shapes are combined by the procedure described in section 5.2.2. The problem is, that a particular prior shape can be overrated to an arbitrary degree, if one would include a huge number of priors in the prior set, which are all very similar shaped. This is because the relative weightings η_j of all other priors would become marginal, compared to the sum of the weightings of that group of similar priors. A thinkable way to reduce this problem could be the definition of a measure for the difference in shape between various priors, which will be discussed below. However, it becomes evident, that one will not find a really objective way to get rid of this scaling problem.

For approach II the problem can be described more formally by assuming a change of the prior parameters $\boldsymbol{\alpha}$ to a different parameterisation $\boldsymbol{\alpha}^*$, which is a function of $\boldsymbol{\alpha}$**.** For example, a family of exponential decreasing priors for two samples with known temporal order can be expressed e.g. in the following different ways:

$$a(t_1,t_2,\alpha) = \alpha \cdot e^{-(t_1-t_2)\cdot\alpha} \qquad \text{or} \qquad a^*(t_1,t_2,\alpha^*) = \frac{1}{\alpha^*} \cdot e^{-(t_1-t_2)/\alpha^*}$$

Where the given exponential functions are valid for $t_1>t_2$, and the priors are zero else. Both, α and α^* are one-dimensional parameters between zero and infinity. So the parameterisation changes from α to $\alpha^*=1/\alpha$. It is helpful for the following consideration to behold the relation between the two prior families:

$$a^*(t_1,t_2,\alpha^*) = a(t_1,t_2,\alpha(\alpha^*))$$

In general, a change of the parameter scale changes the resulting posterior $p(\mathbf{t})$ in Equation 5.2. Putting $p(\mathbf{t},\boldsymbol{\alpha})$ from Equation 5.1 into Equation 5.2, and denoting the posterior calculated with the original parameterisation $p^{(\boldsymbol{\alpha})}(\mathbf{t})$, one gets:

$$p^{(\alpha)}(\mathbf{t}) = \int_{vol} \frac{l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})\ \mathrm{d}\mathbf{t}\,\mathrm{d}\boldsymbol{\alpha}}\ \mathrm{d}\boldsymbol{\alpha}$$

If the alternative parameter scale $\boldsymbol{\alpha}^*$ is used, the posterior $p^{(\boldsymbol{\alpha}^*)}(\mathbf{t})$ is:

$$p^{(\alpha^*)}(\mathbf{t}) = \int_{vol} \frac{l(\mathbf{t})\cdot a^*(\mathbf{t},\boldsymbol{\alpha}^*)}{\int_{vol} l(\mathbf{t})\cdot a^*(\mathbf{t},\boldsymbol{\alpha}^*)\ \mathrm{d}\mathbf{t}\,\mathrm{d}\boldsymbol{\alpha}^*}\ \mathrm{d}\boldsymbol{\alpha}^* =$$

$$= \int_{vol} \frac{l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha}(\boldsymbol{\alpha}^*))}{\int_{vol} l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha}(\boldsymbol{\alpha}^*))\ \mathrm{d}\mathbf{t}\,\mathrm{d}\boldsymbol{\alpha}^*}\ \mathrm{d}\boldsymbol{\alpha}^* = \int_{vol} \frac{l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t})\cdot a(\mathbf{t},\boldsymbol{\alpha})\cdot\left|\frac{\partial\alpha_i^*}{\partial\alpha_j}\right| \mathrm{d}\mathbf{t}\,\mathrm{d}\boldsymbol{\alpha}} \cdot \left|\frac{\partial\alpha_i^*}{\partial\alpha_j}\right| \mathrm{d}\boldsymbol{\alpha}$$

Where $|\partial\alpha^*_i/\partial\alpha_j|$ is the Jacobi determinant, needed for the transformation. Since the integrals in the denominator for both, $p^{(\boldsymbol{\alpha})}(\mathbf{t})$ and $p^{(\boldsymbol{\alpha}^*)}(\mathbf{t})$ are just fixed constants, the ratio between the two values can be given as:

Equation 5.6:
$$\frac{p^{(\alpha^*)}(\mathbf{t})}{p^{(\alpha)}(\mathbf{t})} \propto \frac{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha}) \cdot \left|\frac{\partial\alpha^*_i}{\partial\alpha_j}\right| \mathrm{d}\boldsymbol{\alpha}}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{\alpha}}$$

This shows, that the posterior density $p(\mathbf{t})$ is actually changed by the choice of the parameterisation, because the ratio is in general not constant over $\mathbf{t}$. This is because the variations of $|\partial\alpha^*_i/\partial\alpha_j|$ on the $\boldsymbol{\alpha}$ sub-space are in general passed through to the age-sub-space by the prior function, which depends on both, $\mathbf{t}$ and $\boldsymbol{\alpha}$.
Thus, the resulting posterior density can be altered arbitrarily by changing the parameterisation. The analogous diagnosis for approach I indicated the need of a measure for the difference in shape for the individual priors. In case of this approach one would need a criterion for a parameterisation, which changes the prior shape with a 'constant degree'. A possible way to perform that task will be discussed now for the case of a single prior parameter, without ignoring that this can not solve the scaling problem fundamentally. (Various generalisations for a multi-dimensional parameter space are thinkable, although their practical application may become elaborate.)

Assuming a slight change of the prior parameter denoted as $\delta\alpha$, and the corresponding change of the prior function $\delta a(\mathbf{t}, \alpha) = a(\mathbf{t}, \alpha+\delta\alpha) - a(\mathbf{t}, \alpha)$, a reasonable measure for the degree of change of the prior function could be:

$$\delta P_{\delta a}(\alpha) = \int_{vol} \left|\delta a(\mathbf{t},\alpha)\right| \mathrm{d}\mathbf{t}$$

Which is in some sense the total change of the probability density, when considering increases and decreases in the same way, or the hyper-volume enclosed between the two prior functions within the $\mathbf{t}$ sub-space in other words. Now, one claims that this probability change - which is expressed differentially by Equation 5.7 - shall be constant on the parameter scale.

Equation 5.7:
$$\frac{\mathrm{d}P_{\delta a}(\alpha)}{\mathrm{d}\alpha} = \int_{vol} \left|\frac{\partial a(\mathbf{t},\alpha)}{\partial\alpha}\right| \mathrm{d}\mathbf{t}$$

The consequence of this condition can be illustrated by the use of a family of exponential decreasing priors for two samples with known temporal order, which was discussed already at the beginning of this section. For simplification only a single temporal dimension is used, which is the age difference t^{Δ} between the two real sample ages. Using the straightforward parameterisation, the prior family is:

$$a(t^{\Delta},\alpha) \propto \begin{cases} \alpha \cdot e^{-t^{\Delta}\cdot\alpha} & \text{for} \quad t^{\Delta} > 0 \\ 0 & \text{else} \end{cases}$$

The proportional sign is used, because the prior is just standardised with respect to t^{Δ}, but it is not completely standardised as two-dimensional function on the (t^{Δ},α)-

space. Evaluating Equation 5.7 for this prior family, one gets by some basic conversions the relation below. (The absolute value within the integral can easily be considered, by splitting the integral in two parts below and above from $t^{\Delta}=1/\alpha$.)

$$\frac{\mathrm{d}P_{\delta a}(\alpha)}{\mathrm{d}\alpha} \propto \int_0^{\infty} \left| \frac{\partial(\alpha \cdot e^{-t^{\Delta}\cdot\alpha})}{\partial\alpha} \right| \mathrm{d}t^{\Delta} = \int_0^{\infty} \left| (1-\alpha\cdot t^{\Delta})\cdot e^{-t^{\Delta}\cdot\alpha} \right| \mathrm{d}t^{\Delta} = \frac{1}{\alpha}\cdot\frac{2}{e}$$

Consequently, one has to switch to an alternative parameter α^* that achieves the condition $\mathrm{d}\alpha/\alpha=\mathrm{d}\alpha^*$, which is realised by $\alpha=e^{\alpha^*}$. (Since α runs between zero and infinity, α^* runs between minus infinity and infinity.) So the prior family has to be formulated as follows:

$$a^*(t^{\Delta},\alpha^*) \propto \begin{cases} e^{\alpha^*}\cdot e^{-t^{\Delta}\cdot e^{\alpha^*}} & \text{for} \quad t^{\Delta}>0 \\ 0 & \text{else} \end{cases}$$

Evaluating Equation 5.7 again for this alternative parameterisation, results in a constant changing rate of the prior shape on the α^* scale as claimed:

$$\frac{\mathrm{d}P_{\delta a}(\alpha^*)}{\mathrm{d}\alpha^*} \propto \int_0^{\infty} \left| \frac{e^{\alpha^*}\cdot e^{-t^{\Delta}\cdot e^{\alpha^*}}}{\partial\alpha^*} \right| \mathrm{d}t^{\Delta} = \int_0^{\infty} \left| (1-e^{\alpha^*}\cdot t^{\Delta})\cdot e^{\alpha^*}\cdot e^{-t^{\Delta}\cdot e^{\alpha^*}} \right| \mathrm{d}t^{\Delta} = \frac{2}{e}$$

Concluding, it should be mentioned again, that the described method may avoid the overestimation of particular prior shapes to a certain degree. However, the used criterion is still arbitrary. Thus, the use of parametric prior families cause always arbitrary weightings of the prior functions. And this is also the case for the method introduced as approach I.

5.2.5 Equivalence of approach I and II and the characterisation of both by a resulting effective prior

At the end of section 5.2.2 it was shown, that approach I (range unification by progressive elimination of priors) in its most generalised form, gets equivalent to the use of hpd-ranges on the weighted sum of the posterior marginals. On the other hand, within section 5.2.3 the equivalence of approach II (free prior parameters) to a weighted sum of posteriors, with weighting factors that are the prior-predictions, was demonstrated too. That means, if one uses the prior-predictions as weighting factors within approach I, the result is the same as got by approach II (same hpd-ranges at any level), provided that the discrete priors within approach I are equidistantly extracted out of the parametric prior family of approach II. The latter condition provides the same effective prior-weighting for both methods. Since approach II can be performed numerically in a convenient way by a single Gibbs-sampling run, it seems to be the perfect realisation of the idea of robust analysis, which means trying all possible priors and unify their results, as introduced in section 5.1. Even more, since the method suppresses corrupt priors intrinsically.

Unfortunately this perception is not correct, because it turns out that the resulting posterior density can be also achieved by the use of a single particular prior function, what is shown just below. This is a very serious fact, because it discredits the main motivation of robust analysis to overcome the subjectivity, caused by the choice of a particular prior function.
Combining Equation 5.1 and Equation 5.2 from section 5.2.3 the resulting posterior density $p(\mathbf{t})$, found by integrating over all parameters, is:

$$p(\mathbf{t}) = \int_{vol} \frac{l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t}\, \mathrm{d}\boldsymbol{\alpha}} \; \mathrm{d}\boldsymbol{\alpha} = \frac{l(\mathbf{t}) \cdot \int_{vol} a(\mathbf{t},\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\alpha}}{\int_{vol} l(\mathbf{t}) \cdot (\int_{vol} a(\mathbf{t},\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t}}$$

Denoting $\int_{vol} a(\mathbf{t},\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\alpha}$, which is a particular function in $\mathbf{t}$ again, as effective prior function $a_{eff}(\mathbf{t})$, it is obvious, that the final posterior can be calculated by the help of this prior by a basic single application of the Bayes theorem again:

$$p(\mathbf{t}) = \frac{l(\mathbf{t}) \cdot a_{eff}(\mathbf{t})}{\int_{vol} l(\mathbf{t}) \cdot a_{eff}(\mathbf{t}) \; \mathrm{d}\mathbf{t}}$$

The shape of this effective prior depends on the chosen kind of parameterisation.

Concluding, the calculation with a parametric prior family (approach II) means actually just the use of an effective single prior, and this is also the fact for the method with progressive prior elimination (approach I), because both methods are equivalent in principle. Thus, both methods are unfortunately not realisations of the original idea of unifying the results from calculations using all possible different prior functions.
Nevertheless, it is possible to realise profit from the considerations up to here, because the method to find a parameterisation that defines a prior family with a balanced changing rate of the function shape (described in section 5.2.4) can be seen as a way to define a single prior function less subjective. That means, even if one wants to perform a common Bayesian analysis with a single prior, one can define a family of possible prior shapes, find a well balanced parameter scale and get the resulting prior function by integrating over the parameters.
Applying this procedure to the example discussed at the end of section 5.2.4, results in an effective prior as follows:

$$a^*_{eff}(t^{\Delta}) \propto \int_{-\infty}^{\infty} e^{\alpha^*} \cdot e^{-t^{\Delta} \cdot e^{\alpha^*}} \, \mathrm{d}\alpha^* = \frac{1}{t^{\Delta}}$$

Thus, a family of exponential decreasing priors for the age difference t^{Δ} results in a $1/t^{\Delta}$-shaped effective prior, if the balanced parameterisation α^* (as described above) is used.

5.2.6 Approach III: range unification using a threshold for prior elimination

Within approach I and II it turned out that the attempt to weight the various priors to suppress corrupt ones, made the methods equivalent to a simple calculation with a single prior. Thus, approach III will step back again to an un-weighted unification of

the highest-posterior-density-ranges according to the various priors. An un-weighted unification of hpd-ranges is illustrated within Figure 5.4. However, it is not possible to perform the calculations without providing a method to eliminate corrupt priors. Therefore a quality criterion for the priors has to be defined again, and all priors are discarded, if the defined criterion drops below a particular threshold level. The quality criterion can be based again on the model-data agreement, as in the methods above. A reliable definition for the particular used threshold level has to be found in a most objective way.

Since this approach is next to the original idea of robust analysis, it was chosen to be the actual used method; details are given below. Approach III, is absolutely independent of any scaling, because it performs a pure unification of the original individual hpd-ranges. Naturally, the definition of the quality criterion and its threshold level remain ambiguous. Thus, even this method is not objective in a strict sense any more. However, with reasonable definitions, one will be able to solve essential problems of Bayesian sequencing, as demonstrated in chapter 6. Approach III has to work with individual, discrete prior functions again, which is a difficulty, that has to be handled (see section 5.3.2).

5.3 THE ACTUAL RESULTING METHOD

As justified above, the actual chosen method to perform robust Bayesian analysis is approach III, the pure unification of the hpd-ranges of a set of discrete prior functions. The criterion and threshold level to discard corrupt prior functions has to be defined carefully, as it brings in unavoidable subjectivity. See the detailed definitions below.

5.3.1 The chosen mechanism to discard corrupt prior functions

The characterisation of the individual priors is based on the 'agreement factor J', defined by Equation 4.7 in section 4.3. However, as justified below, it is reasonable to use the relative form of the agreement factor J_{REL}, defined by Equation 4.8. The reference prior will generally be the uniform prior, which is the straightforward one-zero representation of the known constraints (Equation 3.2 or Equation 3.5 in section 3.2 give e.g. the uniform prior for a sequence). The reason why a relative agreement factor is used (as initially mentioned in the rear part of section 4.3.2) is the following: The agreement factor J describes quantitatively the quality of the individual model (prior function) in respect of its consistence with the measurements. However, in the usual archaeological application all individual shaped prior functions base on a common definition of archaeological constraints. Even though this constraints should represent the actual knowledge, they can still be more or less consistent with the measurements practically, caused by uncertainties in the stratigraphy or by defective radiocarbon ages of some samples. Inconsistencies of this kind affect all prior shapes commonly and would reduce all agreement factors, so that less (or even none) individual priors would exceed the threshold limit. What one wants to do contrary, is to discard priors that are in strong disagreement with the

measurement caused by their individual shape, regardless the total quality of the model. Therefore the reference to the uniform prior, that defines the given constraints in a basic form and is usually uncritical in respect of the agreement with the measurements, is used to monitor the total model quality.
The threshold level J^*_{REL} is set according to Equation 4.10, using a confidence level of P^*=0.9545 (2σ). The resulting values for J^*_{REL} depend on the dimension n of the model (number of samples and other parameters as e.g. boundaries) and are plotted in the form of $(J^*_{REL})^{1/n}$ within Figure 4.3 (the middle curve). (Actually Figure 4.3 shows the threshold for the total agreement index I^*_{Π}, which is numerically identical with J^*_{REL}.) The values have been evaluated numerically (see the relations in section 4.1.2) and are listed for various dimensions within the sequencing program. The justification for the use of this threshold level is given in section 4.3.2.
Within the developed Matlab program the relative agreement factor J_{REL} is evaluated individually for each prior function, by the help of the Gibbs-sampling based integration method described in section 4.4. It should be mentioned at this point, that in principle, there is an alternative way to evaluate the agreement factor J and J_{REL} respectively, which can be more convenient in some cases: If it is possible to define the prior set as a parametric prior family $a(\mathbf{t},\boldsymbol{\alpha})$, as done for approach II, one can use the following relation, which was explained already in section 5.2.3. The projection of the resulting posterior on the parameter space $p(\boldsymbol{\alpha})$ is proportional to the prior predictions for the individual priors, if the condition that $\int_{vol} a(\mathbf{t},\boldsymbol{\alpha})\, d\mathbf{t}$ is constant is provided. This relation can also be used for the priors restricted by the domain function $\lambda(\mathbf{t})$, which are used to evaluate J ore J_{REL}. Thus, one can get values proportional to the prior predictions needed for the evaluation of J or J_{REL} (see section 4.3.1) simultaneously for all individual priors, resulting from a single Gibbs sampling run using the parametric prior family. The fact, that one does not get the absolute values for J is irrelevant, when using J_{REL} anyway. However, one is forced to find a parameterisation of the prior set, which furthermore may extend the coordinate space by many additional dimensions and lead possibly to a complex posterior with low convergence. So up to now, the evaluation of J_{REL} is performed as stated at the beginning of this paragraph, except for cases where the dimension of the model is to high for the implemented integration procedure.
In the latter case (for a dimensionality exceeding about 25; compare section 4.4.4) the following tentative criterion is used so far. Instead of the prior prediction based agreement factor J_{REL}, the single sample agreement indices I_i as described in section 4.1.1, Equation 4.1 are used. As it turned out, there is no advantage to use the total agreement index $I_{\Pi}=\Pi I_i$, and thus the individual indices are used directly. Again the indices are related to that resulting from the uniform prior: $I_{i,REL}=I_i/I_{i,unif}$. The threshold level for each I_i is taken from Equation 4.3, evaluated at dimension n=1 and at a confidence level of P^*=0.9545 (2σ), resulting in a value of $\sqrt{2}/e^2$ (0.191); see the explanations at the end of section 4.1.2. Analogous to the situation for the agreement factor J the latter threshold level is directly used for the relative index $I_{i,REL}$ too (compare the considerations previous to Equation 4.10 in section 4.3.2). Finally, a prior function is discarded, if any of the $I_{i,REL}$ drops below the threshold level.

5.3.2 The pragmatic choice of finite sets of priors

The goal is, to approximate the set of different prior function shapes, which is of infinite number in general, by a finite set of priors, without altering the result significantly. There is no general recipe to do so, as the definition of the prior set depends on the actual given archaeological constraints, which can be of various kind.
However, there are some obvious rules that should be considered. First, as the limits of the resulting unified hpd-intervalls (highest posterior density intervals) will be determined by - in some sense - extreme prior shapes that push the marginal posterior density of an arbitrary sample strongly to younger or older ages, it is reliable to focus mainly on these functions to keep the number of priors manageable low. Again, there is no general rule to determine these prior functions, however, in practical applications the elementary structure of these priors is frequently obvious (see e.g. the application given in section 6.3).
In many cases, the priors include age differences, e.g. between phase boundaries limiting the positions of the included sample ages, as introduced in section 2.6.2. If there is no information about the probability density for the phase length, it is reliable to model it with exponential functions with various slopes, even more as these functions have low information content in the sense of the maximum entropy principle, explained in section 3.4.
Frequently, a huge amount of possible prior shapes is the result of the large number of possible combinations of various contributions. For example, if each length of all individual phases of a sequence of phases is modelled by exponential functions with various slopes, one gets in principle a huge number of different multi-dimensional priors, one for each combination of slopes (see again the application given in section 6.3). Fortunately, in many cases a lot of individual contributions are independent of each other, i.e. there are no marginals that are significantly affected simultaneously by these contributions. In that case, there is no need to include all different combinations within the prior set, it is sufficient to have priors that include each variant for each of the independent contributions. This can reduce the number of needed priors significantly.
One can see, that the approximation with a finite prior set is strongly arbitrary for sure. Although this is hurtful, the consequences are much less serious. This is because the definition of the prior set will not change the result significantly, so far the set is not significantly incomplete. To get a roughly complete set is eased by the fact, that one can include every prior which seems to miss, because corrupt prior are discarded automatically and the presence of redundant priors has no consequences within the used method.

5.4 SOME CLARIFICATIONS

5.4.1 Deviation from pure Bayesian statistics

In a strict sense, robust Bayesian analysis as described above deviates from the pure formalism of Bayesian statistics. This is because information from the measurements is used to discard (or weight) the various possible prior functions. This is a violation

of the pure concept, where the prior information (here the archaeological facts) and the measurements should be strictly independent and joined exclusively by the Bayesian theorem. However, this problem is in principle also given for any kind of model selection.

5.4.2 Remaining sources of unavoidable subjectivity

Although the goal of robust Bayesian analysis is the elimination of subjectivity, even so, the following sources of subjectivity remain within the method:
First, the choice of the particular criterion to discard corrupt priors is arbitrary. To use some measure of the prior-data agreement is reasonable, but there are different measures possible. To base the measure on the prior prediction may be the most objective choice, however, even this measure (described in section 4.3) includes additional subjectivity, due to the arbitrary definition of the domain function for standardisation. Furthermore, in the actual used procedure the measure is related to an arbitrarily chosen reference prior (see section 5.3.1). Additionally, one has to define a particular threshold level, which is arbitrary again, although based on careful considerations (section 4.3.2). Finally, there is the subjectivity caused be the arbitrary selection of the finite prior set.
This may sound discouraging, but fortunately the situation is not so bad actually: The sources of subjectivity related to the prior discarding process (all but the last one mentioned) may alter the limits of the resulting hpd-ranges. Although, of primary importance is just, that the prior, nearest to the 'correct one' (explanation in section 5.4.3), is not excluded from the set, because this condition is sufficient to have the real sample ages, with a probability of at least the specified confidence level, within the resulting unified hpd-ranges (see the initial part of section 5.2.1). Thus, subjectivities arising from prior discarding will not cause incorrect results. The question of subjectivity within the procedure to define the prior set is mainly related to the completeness of the set, and has been already discussed in the last paragraph of section 5.3.2. Generally, one can never fully exclude to miss priors or classes of priors that would influence the result, but even then the risk to get an incorrect result has been reduced in comparison to the usual method using a single particular prior only.

It should be noted at this point, that alternatively to the automated procedure of discarding corrupt priors as used here, there is a related procedure discussed in the literature, termed 'prior elicitation', which is the semi-manual identification and rejection of such priors, see e.g. BERGER (1994). However, in some sense this method focuses on a set of different priors associated with different model assumptions, which can be analysed concerning their reasonability. This is less adequate for a prior set as used here, where the meaning of the individual priors within the set is not known explicitly, as the priors are just possible functions, which are consistent with the archaeological constraints. Furthermore, semi-manual prior selection bears the risk of 'tuning' the result towards the expectations of the user, and thus introduces again an unwanted subjectivity.

5.4.3 Specifying the term 'correct prior function'

As mentioned above (see section 5.2.1 and 5.2.2) one aspect of robust analysis (in its form as used in this thesis) is the idea, that there is an ideal or correct prior, and that a close approximation to the latter should be included within the prior set. Thus, it is important to clarify, what is meant when speaking of the correct prior. The idea is, that the correct prior function is the unbiased functional representation of the constraints given by the archaeological facts. The meaning of 'an unbiased functional representation' can be illustrated clearly by having a preliminary look at an example that will be analysed in detail later (see section 6.2.2): An Early-Bronze-Age man (the famous Iceman found in the Oetztaler Alps in 1991) was found together with an axe carried by him. Both, the body and the wooden shaft of the axe have been radiocarbon dated. Regardless some details, this is a real world representation of two samples with known temporal order as frequently discussed previously: The axe shaft cannot be younger than the Iceman himself. Thus, the question is, what is the correct shape of the prior function describing this constraint within reasonable limits.

If we imagine, one would know the real ages (times of death) of a large number of Early-Bronze-Age men and the ages of the wooden shafts of the axes carried by them too. Then one could plot a point for each pair of ages within a two-dimensional coordinate system and fit the points with a corresponding probability density function. This density would represent an unbiased prior function, because its estimation of the probability of a particular age difference between man and axe is the 'right' one. For sure, in reality this density is not known, and therefore the use of a particular prior function usually biases the result. However, it should be noted, that the perception expressed by this example is not valid in general, because the actual probability for an event does not have to be based on a distribution of events (imaginary or not) in every case.

Finally it should be noted, that there is a paper by the author of this thesis together with others authors (WENINGER *et al.*, 2010), which provides additional clarifications to remarkable questions within the discussion section.

5.5 AN RELATED APPROXIMATION: THE 'OVERLAP METHOD'

When analysing various artificial and real-world examples for sequences of samples, or examples of similar kind, with robust Bayesian analysis (a selection of examples is shown in chapters 6), one can recognise a main characteristic, which achieves the expectancies as well: The main difference of the resulting marginal posteriors, compared with these using a particular single prior, is concentrated on the overlapping regions of the single-sample likelihood functions. Parts of the single-sample likelihood functions that are in clear disagreement with the constraints, are fully suppressed with both, the robust and the usual method, because they are processed with parts of the prior function, or the prior function set respectively, that are always zero for both methods. The difficulty arises for the overlapping regions, which allow resulting ages that are consistent and inconsistent with the given temporal order as well. There, the result highly depends on the used prior function

shape. Thus, the pragmatic and cautious approach is, to suspend the sequencing for these regions at all.
The exact rule that will realise this idea will be given below. Before that, the principle will be illustrated for the simple standard case of two samples with known temporal order (Figure 5.6). The effect of the method can be seen more clearly, if one approximates the single-sample likelihood functions to be constant on a particular interval and zero outside. This results in a two-dimensional likelihoods function $l(\boldsymbol{t})$, which is constant on a rectangular base space. The assumed prior information that sample 1 is older than sample 2 is realised by a uniform prior, which is constant left above the diagonal in the shown coordinate system (hatched) and zero right below. The likelihood function disagrees with the prior information, except for the age-range, where the single sample likelihoods overlap. There, the prior function $a(\boldsymbol{t})$ extends into the likelihood function $l(\boldsymbol{t})$ and cuts out a triangular posterior function. Projecting the latter to the axes, results in ramp-shaped posterior marginals (p_1, p_2). This particular shape of the marginals results from the use of the uniform prior, which is an arbitrary decision, as discussed in detail in section 3.1. To avoid sequencing within this overlap region completely, the part of the likelihood function that lies on the right-ordered side is mirrored to the wrong-ordered side (hatched in blue and denoted with m.p.). This part is added to the prior function, but explicitly these regions that lie within the original likelihood function, which is the case for the complete mirrored part for the recent example. The latter restriction prevents of artificially extending the likelihood function. The procedure results in posterior marginals p_1^* and p_2^*, which are constant within the age range where the single sample likelihoods overlap, and correctly-ordered ages are possible. On the other hand, the marginals stay zero for these ranges of the likelihoods that are inconsistent with right-ordered ages. That is exactly what one achieves.

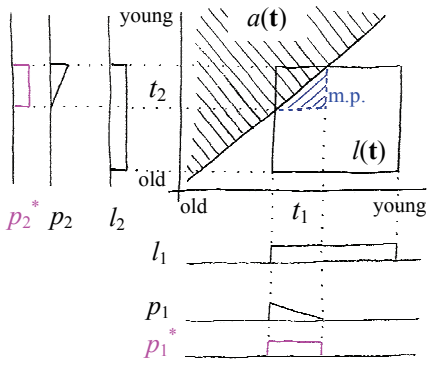

Figure 5.6: Illustration of the basic idea of the 'overlap method': The part of the likelihood function l(t) that is in agreement with the prior information a(t) (here a uniform prior carrying the information that sample 1 is older than sample 2) is mirrored to the disagreeing side. The mirrored part (m.p.) is hatched in blue. This recipe suppresses the sequencing for the age region where the single sample likelihood (l_1, l_2) overlap, without disturbing the sequencing for the remaining regions. The original ramp-shaped posterior marginals (p_1, p_2), which depend on the arbitrary choice of the particular prior shape, are replaced by constant marginals (p_1^*, p_2^*), which represent just the single-sample likelihood functions within the overlapping region.

This basic idea of the 'overlap method' can be generalised for multi-dimensional sequences and arbitrary shaped likelihood functions by the definition of a modified posterior function of the following kind:

$$p(t_1,\ldots,t_n) = \left\{ \begin{array}{l} l(t_1,\ldots,t_n) \quad \text{if} \quad (t_1 \geq t_2 \geq \ldots \geq t_n), \\ \min\left(l(t_1,\ldots,t_n), \left\{ \begin{array}{l} l(t_1,\ldots,t_i^*,\ldots,t_j^*,\ldots,t_n) \quad \text{with} \quad t_i^* = t_j \text{ and } t_j^* = t_i \\ \text{for all pairs } i,j \text{ (with } i<j) \text{ where } t_i < t_j \end{array} \right\} \right) \text{ else} \end{array} \right\}$$

Where the posterior function $p(t_1,...,t_n)$ is deduced from the likelihood function $l(t_1,...,t_n)$ directly by the help of the prior constraints, e.g. $t_1 \geq t_2 \geq ... \geq t_n$. Although this relation is based on the procedure explained just above for the two-dimensional case, it is not a straightforward generalisation of this procedure. The meaning of the generalised procedure can be described in the following way: Combinations of real ages that are wrong ordered with respect to the prior constraints shall be suppressed, but only if their order is 'significant in the light of the measurements'. A particular set of wrong ordered real ages is defined to be 'insignificant' in that sense, and thus will not be suppressed, if definitely all pairs of ages that are in wrong temporal relation can be interchanged, without resulting in disagreements with the measurements.
The procedure can furthermore be generalised for all kinds of priors that consist of various constraints describing older/younger (or more general larger/smaller) relations. A frequently occurring case would be a sequence of phases (including samples without time relation between each other). Therefore the recipe from above has just be generalise to:

$$p(t_1,\dots,t_n) = \left\{ \begin{array}{l} l(t_1,\dots,t_n) \quad \text{if} \quad (t_1,\dots,t_n) \text{ is in agreement with the constraints ,} \\ \min\left(l(t_1,\dots,t_n),\ \left\{ \begin{array}{l} l(t_1,\dots,t_i^*,\dots,t_j^*,\dots,t_n) \quad \text{with } t_i^* = t_j \text{ and } t_j^* = t_i \\ \text{for all pairs } i,j \text{ (with } i<j\text{) where } (t_i,t_j) \text{ is} \\ \text{inconsistent with the constraints} \end{array} \right\} \right) \text{ else} \end{array} \right\}$$

The implementation of the procedure into the Gibbs sampling process is somewhat challenging. There are some difficulties to transform the procedure on the level of one-dimensional cross sections, which are processed within the sampling process. The numerical realisation is not described here, however, it could be analysed within the program code directly; see the reference at the end of section 2.7.

In some sense, the 'overlap method' is the generalisation of 'conventional reasoning' towards the use of continuous and multi-dimensional probability densities. 'Conventional reasoning' is a very simple approach to multi-sample sequencing: One calculates the confidence intervals for all single sample calibrations (at e.g. 95.4% confidence level), and subsequently discards all ranges that are in disagreement with the prior constraints. Thus, the 'overlap method' is a very cautious method, probably leading to wider hpd-ranges as really necessary from the view-point of robustness.
Finally it should be mentioned clearly, that the 'overlap method' is just a pragmatic approximation that can roughly realise central features of the idea of robust analysis for the special case of temporal sequences. It bears the serious handicap, that the meaning of the resulting posterior marginals is not defined in an exact theoretical way any more. However, it is a method that can be executed needing much less run time than the actual robust analysis using multiple priors, described in the previous sections. For some examples within chapter 6 the performance of the 'overlap method' is analysed additionally to the actual method.

6 CHARACTERISING THE FEATURES OF ROBUST BAYESIAN SEQUENCING EXEMPLARILY

The consequences when using a single individual prior instead of including all possible function shapes, as done in robust analysis, can not be characterised in general for various specific applications. However, there are some typical differences that can be pinned down by some simple characteristic examples, as shown in this chapter.
For sure, the selection of examples can never give a complete characterisation of the differences between the common method and robust analysis. This is because on one hand only a few cases can be analysed, and on the other hand the used robust method is in some aspects still arbitrary, as discussed in section 5.4.2. Nevertheless, one can get an impression of the potential of robust analysis and of specific problems as well.

6.1 ILLUSTRATIVE ARTIFICIAL EXAMPLES

6.1.1 The conservation of the sequencing profit

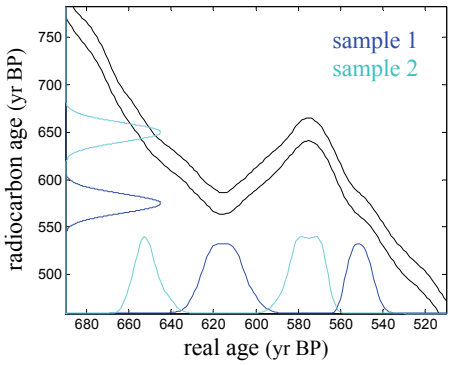

Figure 6.1: Assumed radiocarbon measurements for the discussed example, indicated as Gaussian distributions due to their measurement accuracies (with arbitrary units). The relevant part of the calibration curve is shown by its one-sigma accuracy-band. On the horizontal axis the resulting single sample calibrations (or likelihood functions) are shown (also with arbitrary units).

The goal of robust Bayesian sequencing is to avoid the calculation of inadequate short hpd-ranges, caused by the subjective choice of a single individual prior. However, the method would become useless, if that would lead to such large hpd-ranges so that the benefit of Bayesian sequencing would disappear widely. Therefore one has to test first of all, if the fundamental profit of Bayesian sequencing, that is described exemplarily in section 2.1, is conserved by robust analysis. For this a characteristic simple example (the example of section 2.1, slightly modified) is analysed with the robust method: There shall be two samples with radiocarbon age as shown in Figure 6.1 and the additional knowledge that sample 1 is older than sample 2. The wiggle in the calibration curve causes both single-sample likelihood functions to become double peaked. As demonstrated very detailed in section 2.1, Bayesian sequencing suppresses two of the four peaks of the single-sample

likelihood functions, so that the remaining two peaks are consistent with the given information. Figure 6.2 shows, that this fundamental behaviour can be conserved by robust analysis too and how this works in detail.
The robust sequencing was performed with a prior set consisting of exponentially increasing and decreasing individual priors $a(t_1,t_2)$ of the form

$$a(t_1,t_2) \quad \propto \quad \begin{cases} e^{-(t_1-t_2)/\alpha} & \text{for} \quad t_1-t_2>0 \\ 0 & \text{else} \end{cases} ,$$

where t_1 and t_2 are the real sample ages, and the parameter α steps through the values 2, 3, 5, 10, 20, 50, 1000000, -50, -20, -10, -5, -3, -2 to vary the prior shape from steeply decreasing, over uniform, to steeply increasing. The resulting marginal posterior densities are shown together with the corresponding single-sample likelihood functions for both samples by the two uppermost plots within Figure 6.2. In general, one can see that all marginal posteriors produced by the different individual priors keep the older peak of sample 1 and the younger peak of sample 2, and suppress the other two as required. Priors with strongly decreasing exponential functions try to pull the two remaining peaks closer together, and that with strongly increasing ones push them apart. If this behaviour exceeds a certain level, the (two-dimensional) posterior function becomes inconsistent with the particular measured radiocarbon ages (the two dimensional likelihood function), and the corresponding prior is discarded. (The discarded functions are plotted in grey; the accepted in red.) For this and all other examples in this section, the relative agreement factor J_{REL}, defined by Equation 4.8 in section 4.3.1, was used as criterion to discard priors that are in bad agreement with the measurements (see the individual values in the plot). The threshold level results from Equation 4.10 and depends on the number of dimension. For the given two-dimensional example the level is 0.091.
The two plots in the middle part of Figure 6.2 show the highest posterior density (hpd)-ranges envelopes (showing the hpd-ranges to any confidence level; see section 5.2.1 for explanation) of all accepted individual priors, and additionally the final unified hpd-ranges envelope (broad line in pale red). The plots illustrate the way how the hpd ranges are extended by robust analysis.
The lowermost plots give a comparison of the resulting hpd-ranges from robust sequencing (red) with the result from the simple uniform prior (blue). Additionally the resulting ranges generated by a totally constant prior (or in other words the ranges for the single-sample likelihood functions themselves) are shown. The plots show, that robust analysis preserves the benefit of Bayesian sequencing, as done by using the uniform prior only: In this example both methods suppress two of the four peaks of the marginal posteriors to get a result that is consistent with the prior information. For sure, robust analysis extends the hpd-ranges compared to using an individual prior only.

For this and again for the further examples too, the prior set is built by exponential functions. This choice is justified by two different reasons. First, as discussed in section 5.3.2, the extend of the resulting hpd-ranges of robust analysis is significantly influenced by priors that try to shift the single sample likelihoods on the age scale strongly. A set of exponential priors with various slopes includes strongly shifting function as well as more neutral ones. Naturally one could include various other function shapes in the set too, but there seem to be no significant improvements

compared with the used set. Further, the exponential (decreasing) function is in some sense a special one, because the maximum-entropy method shows that it has the lowest information content for non negative values with given prediction value (see a detailed discussed in section 3.4). Therefore it is reasonable to model an age difference of two samples with know temporal order with a decreasing exponential function. The fact that there is no prediction value for the age difference known is considered by the use of a whole set of functions with different slopes. To use increasing functions too, is a pragmatic extension for a better completeness of the prior set. It should be remembered at this point, that - in principle - it is allowed to use any function that is consistent with the prior information, because 'corrupt' functions will be discarded by the agreement criterion.

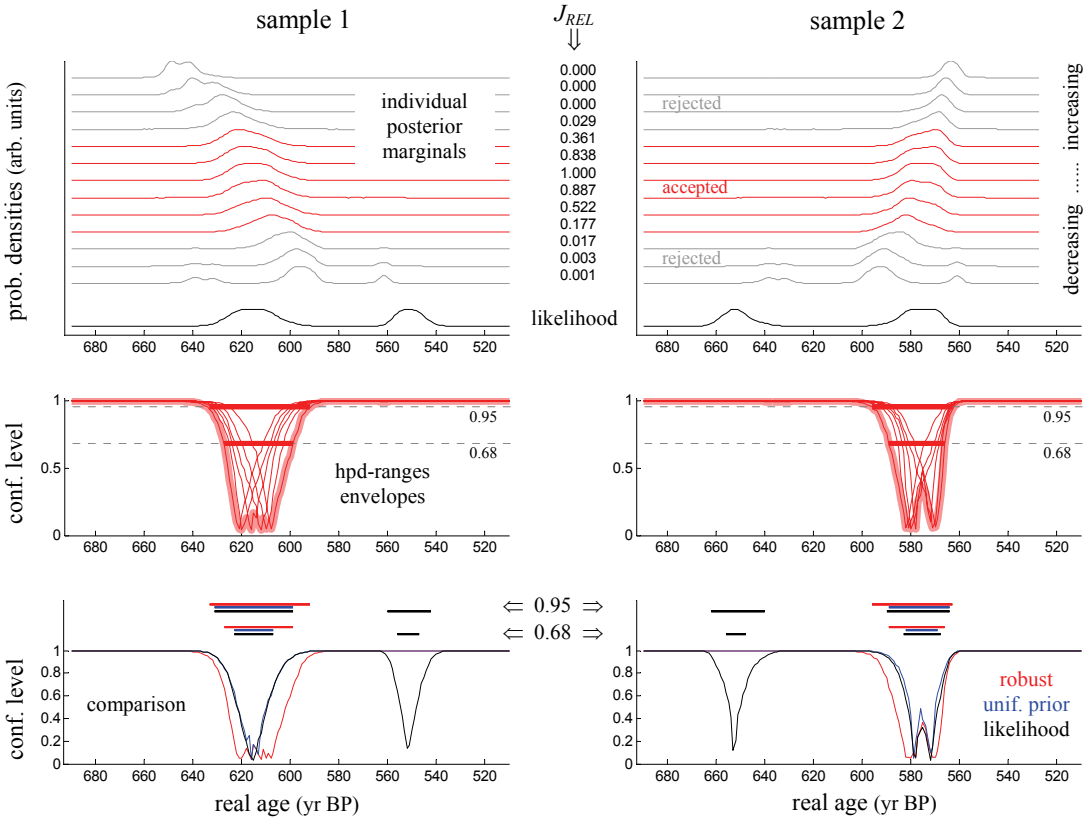

Figure 6.2: Results for the discussed example. The upper two plots show the marginal posterior probability densities to the different used prior function shapes; from strongly decreasing to strongly increasing functions from the bottom to the top. The hpd-ranges envelopes corresponding to the accepted priors are unified as shown in the two middle plots. The last two plots give a comparison of the resulting hpd-ranges from robust sequencing with the result from the simple uniform prior, and additionally with the ranges for the single-sample likelihood functions themselves. The example illustrates that robust analysis preserves the benefit of Bayesian sequencing.

6.1.2 The need of suppressing 'corrupt' priors

The reason to present this section's example is used to illustrate clearly, that it is absolutely necessary to discard corrupt prior functions, as described in section 5.3.1. For this we assume two radiocarbon dates with clearly separated likelihood functions

as shown in Figure 6.3. (To get an example with clearly understandable results, the radiocarbon dates are positioned at a section of the calibration curve that is relatively flat when seen on a large scale.) Additionally we assume again the given prior knowledge that sample 1 is older than sample 2. Thus, the likelihood functions are in perfect agreement with the given prior information and there is no need to perform Bayesian sequencing at all. However, although there is no need for sequencing, sequencing has to work in this case too for sure, which means the sequencing procedure should keep the single sample calibrations preferable unchanged, as they already agree with the prior information. So again, as in the example above, a prior set with decreasing and increasing exponential functions modelling the age difference is used (α steps through positive and negative values):

$$a(t_1,t_2) \quad \propto \quad \begin{cases} e^{-(t_1-t_2)/\alpha} & \text{for} \quad t_1-t_2>0 \\ 0 & \text{else} \end{cases}$$

The resulting posterior marginals, one set for each of the two samples, are shown in Figure 6.4. As already identifiable in the previous example, it can be seen here very clearly, that priors that are steeply decreasing exponential functions try to pull the marginals close together towards a similar age, and oppositional, steeply increasing ones push them far apart on the age scale. Thus the marginals of these priors lie completely away from the single sample calibrations. It is clearly evident for this example that the latter are unwanted results that have to be discarded, considering the fact that the likelihood functions are without modifications completely consistent with the given prior information.

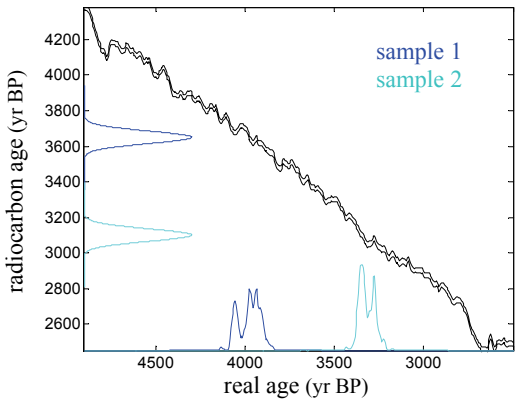

Figure 6.3: Assumed radiocarbon measurements for the example discussed within this section (Gaussian distributions due to the measurement accuracies). The calibration curve is given by its one-sigma accuracy-band. On the real-age axis the resulting single sample calibrations (or likelihood functions) are shown. (Distributions in arbitrary units)

As already mentioned in the previous section, the actual procedure to discard corrupt priors is based on the relative agreement factor J_{REL}, which is a Bayes factor based agreement measure (see sections 5.3.1 and 4.3). For the current example the particular values for J_{REL} corresponding to the different individual priors are given in Figure 6.4. In the two-dimensional case the threshold for J_{REL} is 0.091 (Equation 4.10). Discarded results below this level are plotted in grey colour within the figure. The pair of marginals at a level of J_{REL}=1 results from the uniform prior. The probability densities just above these pair correspond to priors with slightly increasing probabilities along increasing sample-age differences. Initially they show values for J_{REL} that exceed the level of one, which means these priors are more likely than the uniform prior in the light of their agreement with the data. This can be

understood when considering the fact that the single sample likelihoods are clearly separated. Therefore, a prior that shows a higher difference between the probabilities of separated and of overlapping single sample likelihoods than the uniform prior, represents the better model considering the data. When the slope of the increasing exponential functions rises further, the posterior is more and more shifted out of the likelihood and the agreement factor decreases.
Naturally, there is no perfect general procedure to distinguish between 'correct' and 'corrupt' priors. Thus it is not avoidable that there will remain priors that extend the resulting age ranges (hpd-ranges) in some cases more than necessary to achieve robustness. In the recent example this may be the case for the pair of marginals from the steepest increasing prior function that is still accepted (the uppermost marginals in red colour within Figure 6.4). These differ from the single sample likelihoods significantly. However, this is the price for avoiding an incorrect shortening of the resulting age ranges caused by an arbitrary choice of a single prior. On the other hand, there may be found improved procedures to discarded corrupt priors in further investigations that could reduce this problem.

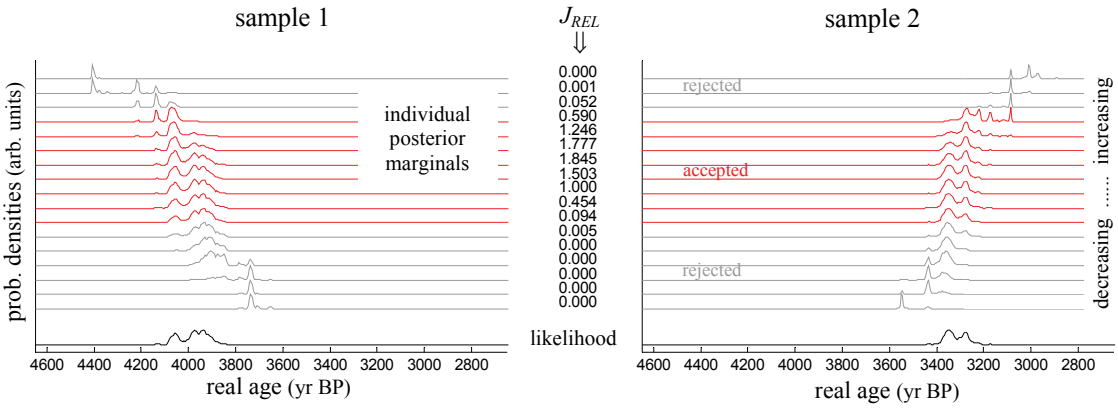

Figure 6.4: Marginal posterior probability densities related to the different used prior function shapes, varying from strongly decreasing exponential functions at the bottom to strongly increasing ones at the top. The results with a relative agreement factor below the threshold level (which is 0.091 in the two-dimensional case) are discarded (grey). The posterior marginals associated with the value 1 for the relative agreement factor result from the uniform prior (α=∞). The black curves give the single sample calibrations (i.e. likelihood functions) for comparison.

It is reasonable to analyse briefly the characteristics of exponential prior functions at this point, as the prior sets used in this thesis are mainly based on exponential functions in general. The effect of an exponential prior function can be typically seen when considering a simple one-dimensional Gaussian likelihood function $l(t)$ (generated when assuming a linear calibration function) which is multiplied by an exponential prior function $a(t)$ (which would represent a prior knowledge that older ages are more likely than younger) to get the posterior $p(t)$:

$$p(t) \quad \propto \quad l(t) \cdot a(t) \quad \propto \quad e^{-t^2/(2\sigma^2)} \; . \; e^{-(t+d)/\alpha}$$

The fact that the Gaussian is centred artificially at t=0 does not disturb the universality of this considerations, because the parameter d allows for placing the exponential function relatively to this position fully free. The right side from the equation above can be transformed elementarily to:

$$e^{-(t+\frac{\sigma^2}{\alpha})^2/(2\sigma^2)} \cdot e^{-\sigma^2/(2\alpha^2)} \cdot e^{-d/\alpha} \quad \propto \quad e^{-(t+\frac{\sigma^2}{\alpha})^2/(2\sigma^2)}$$

This shows, that the resulting posterior function remains Gaussian keeping the same σ in that simple case, but the Gaussian is shifted by the distance σ^2/α on the age axis. Furthermore, the shift is independent from the position of the exponential function in relation to the likelihood function. The expression for the age shift shows clearly, that the posterior can be shifted away from the likelihood function to any extend, if the parameter α is defined very small to make the exponential prior function very steep. Thus, for very small α the priors become corrupt and have to be discarded.

6.1.3 Dealing with an asymmetric 'statistical pressure'

A very fundamental problem of Bayesian sequencing, which could even be denoted as an artefact, is illustrated by the example within this section. Frequently there are sequences of phases that contain various numbers of samples modelled. It turns out, that the calculated posteriors for the boundaries between the phases are strongly biased by the number of samples within the phases. A phase boundary between a phase containing many samples and another phase with just few, tends to be shifted towards the latter phase (a significant overlap of the likelihood functions provided). Therefore one can speak of 'statistical pressure'. For sure, one could try to justify this behaviour by the idea, that the number of samples is an indicator for the phase length. However, in general the chosen number of samples within a phase is totally arbitrary and should not influence the result. Therefore it is commonly accepted for cases of this kind to use a prior function that includes factors that suppresses the influence of the sample numbers. The problem of statistical pressure was theoretically discussed in section 3.3, and the mentioned prior function including these additional factors (and a further factor also described in section 3.3) called 'uniform overall-span prior' is given by Equation 3.4.
To get an illustrative example showing the discussed behaviour very clearly, one assumes five measurements with equivalent radiocarbon ages, just falling on a plateau of the calibration curve, which is idealised in its shape to get well understandable results (see Figure 6.5). One sample is supposed to lie within an older phase and the remaining four within a younger phase.

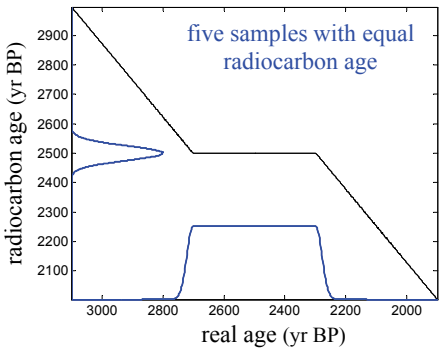

Figure 6.5: Measured values and single-sample likelihood functions of the example explaining the problem with the 'statistical pressure'. Five samples, one within an older phase and four with a younger phase, are assumed to lie on an (idealised) plateau and to show all similar measurements and consequently similar likelihood functions.

Using this example, various variants of sequencing methods shall be compared. Beside modelling with the simple uniform prior and robust analysis on the other hand, there are the following methods demonstrated additionally: The use of the 'uniform overall-span prior' mentioned above, two different calculations using free prior parameters described in section 5.2.3, and finally the 'overlap method', which was introduced in section 5.5 and can be seen as an approximation of robust analysis. The results are given in two different ways, once by the resulting ranges at a confidence level of 95.4% (2σ) shown in Figure 6.6, and second by the resulting ranges envelopes shown by Figure 6.7. (In case of this example there are symmetric ranges instead of hpd-ranges used; see the explanations at the rear part of the section.) In both figures, only on sample of the four within the young phase is given exemplarily, because these four samples show (aside of statistical variations) identical results. Beside the results for the samples, the posterior marginal of the boundary between the two phases is shown too (the middle graph in both figures; the model includes two outer boundaries too, which are not shown).
In the following the prior functions for the different methods are describe briefly. Therefore the real ages of the samples ($t_{...}$) and the boundaries ($b_{...}$) will be denoted as shown below:

b_3	... beginning of the older phase
$t_{O,1}$	... single sample within the older phase *O*
b_2	... boundary between older and younger phase
$t_{Y,1}, t_{Y,2}, t_{Y,3}, t_{Y,4}$	... four samples within the younger phase *Y*
b_1	... end of the younger phase

Note, that different to the idealised example in section 3.3, which was used to discuss the prior marginals of the 'uniform overall-span prior' and was of similar structure, the outer boundaries are now not fixed any more. For a more compact notation a characteristic function $\chi(...)$ will be used, that is defined to be one if the relation (...) is true and zero otherwise.

The ***uniform prior*** has the following simple structure:

$$a(\mathbf{b},\mathbf{t}) \;\propto\; \chi\left(b_3 > \{t_{O,1}\} > b_2 > \{ t_{Y,1}, t_{Y,2}, t_{Y,3}, t_{Y,4} \} > b_1\right)$$

The ***uniform overall-span prior*** adds correction factors to the uniform prior as mentioned above:

$$a(\mathbf{b},\mathbf{t}) \;\propto\; \chi\left(b_3 > \{t_{O,1}\} > b_2 > \{ t_{Y,1}, t_{Y,2}, t_{Y,3}, t_{Y,4} \} > b_1\right) \;\cdot\; (b_3-b_2)^{-1} \cdot (b_2-b_1)^{-4} \;\cdot\; (b_3-b_1)^{-1}$$

Where the last factor (in general $(b_m-b_1)^{-(m-2)}$) provides a constant prior marginal for the overall-span (i.e. for $b_{m(=3)}-b_1$); see details in section 3.3 (Equation 3.4).

The calculation is further performed with two different ***priors with free prior parameters***. Each prior is defined with two free parameters that characterise the prior information in respect to the length of the two phases. In the first case the lengths of the phases are characterised by two parameters (α_O, α_Y) that are the expectation values of exponential functions modelling the length of the two phases, and in the second case the reciprocal expectation values (or the decay constants in other words) of the

exponential functions are used as free parameters (α'_O, α'_Y). The two prior functions are formally

$$a(\mathbf{b},\mathbf{t},\boldsymbol{\alpha}) \propto \chi\left(b_3 > \{t_{O,1}\} > b_2 > \{t_{Y,1}, t_{Y,2}, t_{Y,3}, t_{Y,4}\} > b_1\right) \cdot (1/\alpha_O^2) \cdot \exp(-(b_3-b_2)/\alpha_O) \cdot \cdot (1/\alpha_Y^5) \cdot \exp(-(b_2-b_1)/\alpha_Y)$$

or

$$a(\mathbf{b},\mathbf{t},\boldsymbol{\alpha}') \propto \chi\left(b_3 > \{t_{O,1}\} > b_2 > \{t_{Y,1}, t_{Y,2}, t_{Y,3}, t_{Y,4}\} > b_1\right) \cdot \alpha'^2_O \cdot \exp(-(b_3-b_2)\cdot\alpha'_O) \cdot \cdot \alpha'^5_Y \cdot \exp(-(b_2-b_1)\cdot\alpha'_Y)$$

respectively.

The free parameters α_O, α_Y or α'_O, α'_Y respectively are treated as additional variables within the Gibbs sampling procedure. The definition range for α_O and α_Y was chosen from 10 to 1000 yr, where the latter creates more or less a uniform prior. The ranges for α'_O and α'_Y where chosen from 0 to 0.1 yr^{-1}, where the latter equates to the 10 yr from above and the former creates the uniform prior. The factors previous to the exponential functions within the relations above, are in some sense standardisation constants. The used exponents are equal to the number of samples within the corresponding phase plus one. It can be shown by an elementary integration of the priors over all age and boundary coordinates, that the use of this factors makes the hyper-volume independent of the particular value of the parameters. E.g. for the first type of parameterisation the integral is:

$$\int_{-\infty}^{+\infty} \mathrm{d}b_1 \int_{b_1}^{+\infty} \mathrm{d}b_2 \, \frac{1}{\alpha_Y^5} \cdot e^{-(b_2-b_1)/\alpha_Y} \int_{b_1}^{b_2} \mathrm{d}t_{Y,1} \cdots \int_{b_1}^{b_2} \mathrm{d}t_{Y,4} \int_{b_2}^{+\infty} \mathrm{d}b_3 \, \frac{1}{\alpha_O^2} \cdot e^{-(b_3-b_2)/\alpha_O} \int_{b_2}^{b_3} \mathrm{d}t_{O,1} = 4!\cdot 1!\cdot \int_{-\infty}^{+\infty} \mathrm{d}b_1 \, 1$$

The remaining infinite term can be seen as constant in this context; see similar results in sections 3.2 and 3.3. It is obvious that this standardisation can be generalised for an arbitrary number of phases and samples.
For sure, the parametric prior function do not have to be standardised (or more exact spoken independent of the particular parameter values) necessarily, but only in this case the weighting within the prior set is done with the prior predictions of the individual prior shapes (this is explained within section 5.2.3); and thus the method is closest to robust analysis, where the procedure to discard priors is based on the prior prediction too.

For ***robust analysis*** the used prior set is similar to the parametric priors above:

$$a(\mathbf{b},\mathbf{t})_{\boldsymbol{\alpha}} \propto \chi\left(b_3 > \{t_{O,1}\} > b_2 > \{t_{Y,1}, t_{Y,2}, t_{Y,3}, t_{Y,4}\} > b_1\right) \cdot \exp(-(b_3-b_2)/\alpha_O) \cdot \cdot \exp(-(b_2-b_1)/\alpha_Y)$$

Different to the parametric priors, in case of robust analysis there are no standardisation factors within the priors needed, as all different prior shapes are calculated individually, and prior standardisation has no meaning within an individual sampling process. (The procedure to calculated the agreement factor includes an intrinsic standardisation procedure; section 4.3.1.) Within the calculation each of the two parameters undertook the values 10, 20, 50, 200 and 100000 yr, leading to 25 different priors including all possible combinations of the two parameters.

Finally, for the **'overlap method'** the posterior function is directly deduced from the prior constraints without using an explicit prior function. The constraints are still:

$$b_3 > \{t_{O,1}\} > b_2 > \{ t_{Y,1}, t_{Y,2}, t_{Y,3}, t_{Y,4} \} > b_1$$

The procedure, how the constraints are embedded within the calculation is explained in section 5.5.

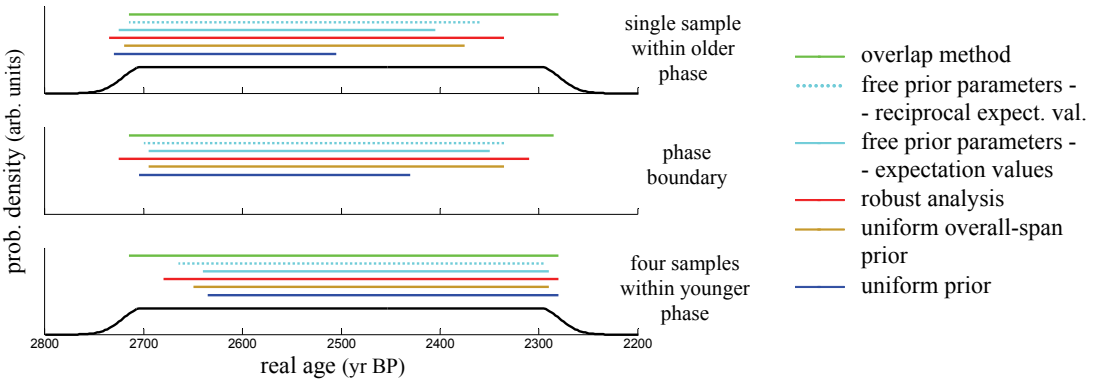

Figure 6.6: Ranges at a 95.4% confidence level (symmetric ranges instead of hpd-ranges; see text) resulting from the various methods listed in the legend. The black curves show the original likelihood functions. The similar results for the four samples within the younger phase are shown as one.

The different degrees of the individual procedures to resist the asymmetric statistical pressure can be best seen when looking at the results for the boundary between the two phases (the middle graph in Figure 6.6 and Figure 6.7). As the example is totally symmetric except the different numbers of samples within the two phases, the result for the boundary should be ideally symmetric. It is further expected, that the resulting hpd-range for high confidence levels should tend to cover the full range of the likelihood plateau, because all likelihood functions are equal, and thus, expecting a small duration of the whole sequence, all samples ages could lie nearly anywhere within this range.
The uniform prior (blue line or curve) is far away from this expectations. Robust sequencing (red) shows a roughly symmetric result and ranges that are close to the latter expectation. With the use of the 'uniform overall-span prior' (ochre) one can get results that are close to these of robust analysis in case of this example. The two variants using free prior parameters (turquoise) show similar results as the 'uniform overall-span prior' too. The 'overlap method' (green) shows the best result for this example in the light of the mentioned expectations.
It is reasonable in case of this example, that the quality of the result got with the 'uniform overall-span prior' is not much less than this of robust analysis, because the prior is designed to solve just the specific problem treated by this example. However, this is not the case in general, as demonstrated later.
Although the calculations with the free prior parameters produce good results, one can see a significant difference between the two different scaling variants that is completely artificial. As theoretically discussed in section 5.2.5, this results from the fact, that the method using free prior parameters can be reduced to a calculation with a single effective prior function, which depends on the used parameterisation. This

scaling problem is further discussed in section 5.2.4, together with a theoretical idea to find a somehow optimised scale. However, the idea has not been adapted for general practical use, because it could not solve the problem fundamentally.
The reason that the 'overlap method' delivers the best result is the fact, that the complete equivalence of all single-sample likelihood functions in case of this example benefits the method.

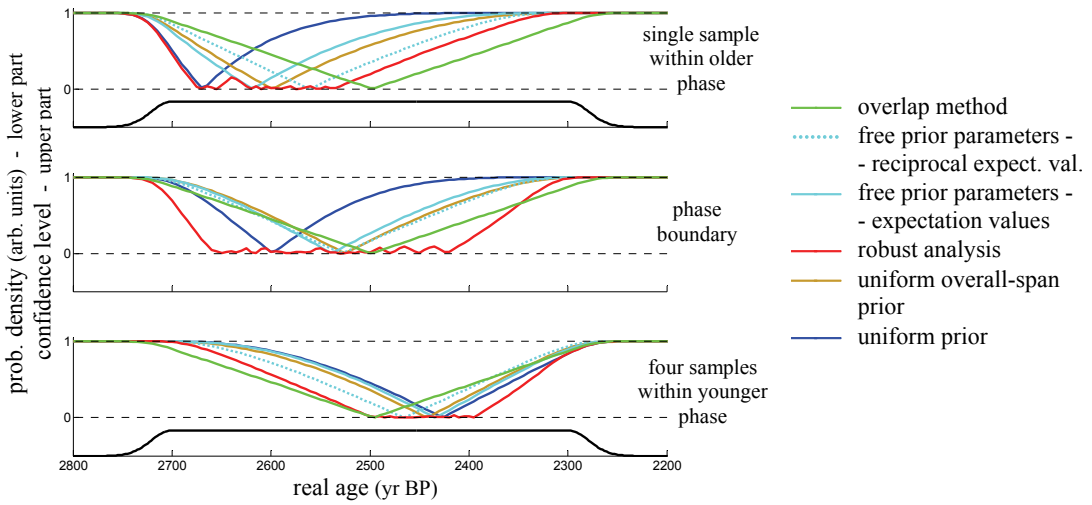

Figure 6.7: Ranges envelopes (based on symmetric ranges instead of hpd-ranges; see text) resulting from the various methods listed in the legend. The black curves show the single-sample calibrations (likelihood functions). The similar results for the four samples within the younger phase are shown as one.

It should be mentioned that for this example there where no increasing exponential functions used within the prior set for robust analysis to get a direct comparability with the method using free prior parameters. In the latter case increasing exponential functions can not be used, if one wants to work with a standardised prior set, what is reasonable as justified above.

As noted above, there are symmetric intervals instead of highest posterior density ranges used in this examples. They are defined as continuous intervals covering a particular total probability of the posterior marginals in a way, that the excluded probability at the younger side is equal to this at the older side. The reason for this modification is caused by the artificial flat structure of the single-sample likelihood functions influencing the structure of the posterior marginals as well. As hpd-ranges become ambiguous on flat plateaus, their use would cause strong statistical fluctuations within the results and violate the clearness of the conclusions.

6.1.4 Dealing with the 'spread out' artefact

The example demonstrated in this section deals with a further fundamental problem of Bayesian sequencing that was already introduced within section 3.2. The same (artificial) measurements and also the same idealised calibration curve that where used in the previous section are reused unchanged for the current example (see

Figure 6.5). The prior knowledge assumed now is a defined temporal order of all individual samples as a sequence ($t_1 < t_2 < ... < t_5$).
The same methods as above shall be analysed here; the analogues prior functions (denoting the real sample ages $t_1, t_2, ... t_5$, from young to old) are listed in the following. Explanations that have been given already in the previous section, and which can analogously be transferred to the recent case, are not repeated.

The ***uniform prior*** is:

$$a(\mathbf{t}) \propto \chi(t_5 > t_4 > ... > t_1)$$

The ***uniform span prior*** for a simple sequence is (which is the equivalent to the 'uniform overall-span prior' for a sequence of phases):

$$a(\mathbf{t}) \propto \chi(t_5 > t_4 > ... > t_1) \cdot (t_5-t_1)^{-3}$$

In general the additional factor is $(t_n-t_1)^{-(n-2)}$ (n is the number of samples); see details in section3.3. (Equation 3.3).

The two different ***priors with free prior parameters*** are defined, analogous to the previous section. Here one assumes a prior information about the total phase length, characterised again by an exponential function using the expectation value α or the decay constant α' as single free parameter, leading to

$$a(\mathbf{t},\alpha) \propto \chi(t_5 > t_4 > ... > t_1) \cdot (1/\alpha^4) \cdot \exp(-(t_5-t_1)/\alpha) \cdot$$

or

$$a(\mathbf{t},\alpha') \propto \chi(t_5 > t_4 > ... > t_1) \cdot \alpha'^4 \cdot \exp(-(t_5-t_1)\cdot\alpha') \cdot$$

respectively.

The exponent in the 'standardisation term' is the number of samples minus one. This can be easily justified by integration over all coordinates. E.g. for the first type of parameterisation the integral is:

$$\int_{-\infty}^{+\infty} dt_1 \int_{t_1}^{+\infty} dt_2 \int_{t_2}^{+\infty} dt_3 \int_{t_3}^{+\infty} dt_4 \int_{t_4}^{+\infty} dt_5 \, \frac{1}{\alpha^4} \cdot e^{-(t_5-t_1)/\alpha} = \int_{-\infty}^{+\infty} dt_1 \, 1$$

Again, for ***robust analysis*** a prior set similar to the parametric priors above is used:

$$a(\mathbf{t})_\alpha \propto \chi(t_5 > t_4 > ... > t_1) \cdot \exp(-(t_5-t_1)/\alpha)$$

The executed values for the parameter α are: 10, 20, 50, 100, 200, 500 and 100000 yr. (Again no increasing functions are used for better comparability with the free-parameter methods.)

The **'overlap method'** is again based directly on the prior constraints, which are:

$$t_5 > t_4 > ... > t_1$$

Using these priors and methods the resulting ranges (again the symmetric ranges are used as justified in the previous example) show clearly that the uniform prior strings

the ranges of the individual samples over the whole plateau, so that the youngest sample is positioned at the younger end of the plateau, and the oldest at the opposite end (see Figure 6.8). So roughly spoken, the method suggests that the real ages are always spread out over the full range of the plateau. It is evident that this is not correct, because it is possible (and not unlikely) that the total time span of the sequence is not similar to the duration of the plateau. A total span of the sequence that is small compared to the length of the plateau, would make ages at any position along the plateau equally likely, and this for each of the samples. All methods, apart from using the uniform prior, get more or less close to be in agreement with the latter fact by covering the whole plateau (see Figure 6.8). Robust analysis shows the best result aside from the overlap method, which is again favoured by the symmetric structure of the example.

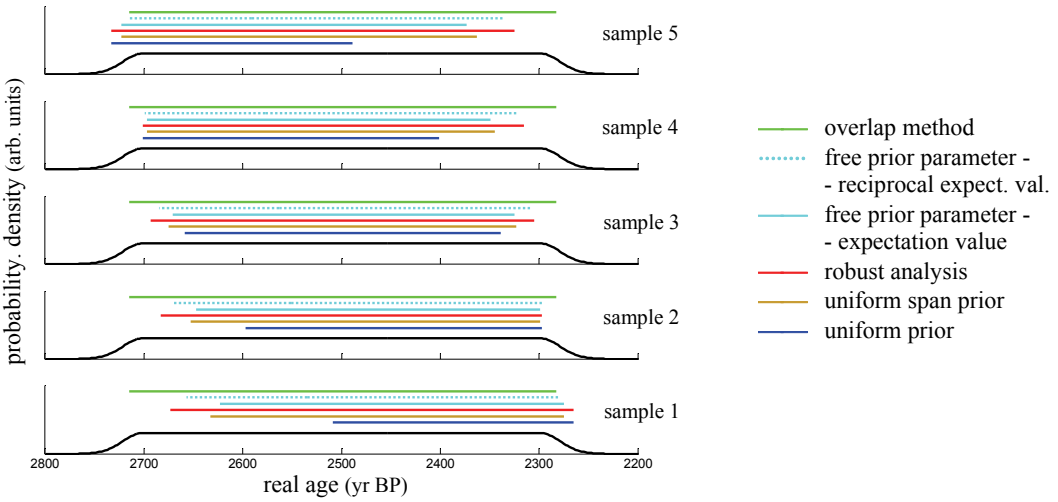

Figure 6.8: Ranges at a 95.4% confidence level (symmetric ranges instead of hpd-ranges; see text) resulting from the various methods listed in the legend. The black curves show the single-sample calibrations (likelihood functions). Samples 1 to 5 are assumed to be temporal ordered from young to old.

At this point the used simple kind of parameterisation of the prior set for robust analysis, where just the prediction for the total span of the sequence is varied by the use of a single parameter, should be analysed roughly. A more general prior set could be defined by assembling four exponential functions, each characterising the time span between two neighbouring sample ages using an individual parameter:

$$a(\mathbf{t})_{\boldsymbol{\alpha}} \;\propto\; \chi\,(t_5 > t_4 > \ldots > t_1) \;\cdot\; \exp\,(-(t_5-t_4)/\alpha_{5,4}) \cdot \exp\,(-(t_4-t_3)/\alpha_{4,3}) \cdot \ldots \cdot \exp\,(-(t_2-t_1)/\alpha_{2,1})$$

If one simplifies this prior set by varying the four parameters simultaneously instead of each one individually, (which is for sure a restriction of the set) the prior set becomes identical with the actually used one, because of:

$$\chi\,(t_5 > \ldots > t_1) \;\cdot\; \exp\,(-(t_5-t_4)/\alpha) \cdot \ldots \cdot \exp\,(-(t_2-t_1)/\alpha) \;=\; \chi\,(t_5 > \ldots > t_1) \;\cdot\; \exp\,(-(t_5-t_1)/\alpha)$$

Regardless of the possibility to write the prior function (for any value of α) in this simple mathematical form, any age difference between neighbouring samples shows still a marginal that is proportional to $\exp(-(t_{i+1}-t_i)/\alpha)$, what is desired. (The method

to calculate prior marginals is described in section 3.2.) For the current example this simple prior set is yet sufficient to solve the 'spread out' problem.

Both, the current example and this of the previous section show that typical problems occurring from the use of a single particular prior can be solved well with the method of robust analysis. In case of these simple examples the commonly used corrected priors achieve also results that are not far from these of robust analysis. The methods using free prior parameters differ significantly from each other depending on the chosen parameter scale. As they are furthermore theoretically equivalent to the use of a single particular prior yet (see section 5.2.5), they will not be discussed in the following example any more.

6.1.5 Comparison with the non-Bayesian 'conventional reasoning'

Conventional reasoning (already mentioned in section 5.5) is a very straightforward non-Bayesian approach to sequencing: The resulting ranges are simply deduced from the confidence ranges of the single sample calibrations (e.g. at 95.4% confidence level; defined analogous to hpd-ranges) by discarding all parts that are in disagreement with the prior constraints. That means, if e.g. the range of a sample A that is know to be younger than a second sample B exceeds the range of the latter at the older side, this part is cut off. It is easy to imagine that considerations of this kind can be generalised for whole sequences.

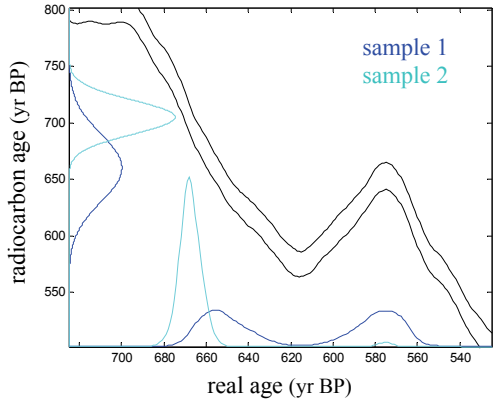

Figure 6.9: Assumed radiocarbon measurement and corresponding single-sample likelihood functions to demonstrate the fundamental difference of 'conventional reasoning' and Bayesian methods. (Again the prominent wiggle around 600 BP is used.)

Looking at the last two examples, where the methods are tested whether they are able to produce resulting ranges that cover the full single sample calibrations, which for sure would be the case with conventional reasoning, one could get the impression that robust Bayesian analysis, and the overlap method as well, are just complicated ways to realise conventional reasoning. Actually there is a fundamental difference between these two methods: Bayesian sequencing changes the probability density when transforming the likelihood function to the posterior function, according to the prior probability density. This can gradually reduce or enlarge parts of the likelihood as well. Conventional reasoning is always based on the unchanged single sample likelihoods.
It should be mentioned at this point, that different to the examples of the two previous sections, the usual hpd-ranges are used again for the current example and

for all further examples in this thesis too. The example given now shows one particular illustrative case to point out the difference between conventional reasoning and Bayesian methods. Assuming measurements as given in Figure 6.9 and once more the prior information that sample 1 is older than sample 2, one finds the results, shown in Figure 6.10, for various Bayesian methods and conventional reasoning. All Bayesian methods include the younger peaks of both samples, because the probability that the age of sample 2 falls in region of the younger peaks, which is very small in the non-modelled single sample calibration, is enlarged by the prior information. It is clear that conventional reasoning can not include this range (at the currently used 95.4% confidence level), as the range is already excluded by the single sample calibration. (Naturally, if one increases the confidence level more and more, the younger peaks will finally be included with conventional reasoning too.)

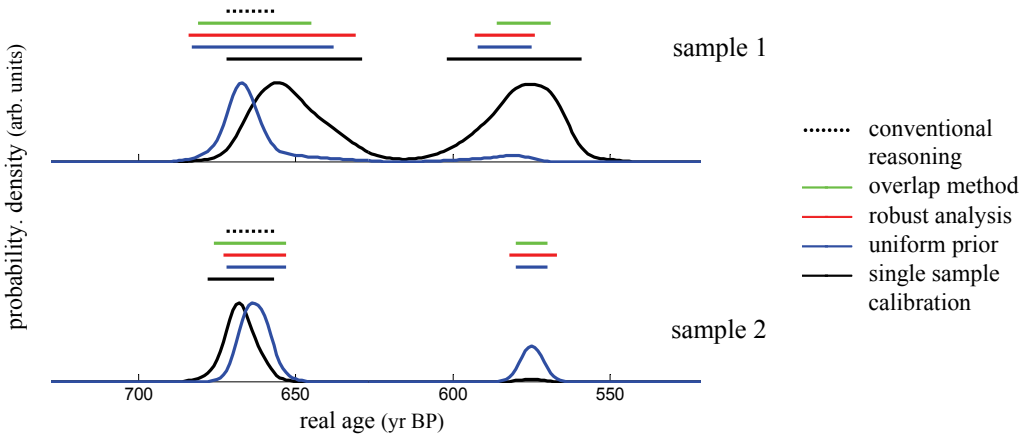

Figure 6.10: The fundamental difference of 'conventional reasoning' and the Bayesian methods. The probability densities in black are the single sample calibrations. The Bayesian methods enlarge the younger peak of sample 2 drastically (typically demonstrated by the marginal posterior densities calculated with the uniform prior; in blue). Therefore the 95.4% hpd-ranges for the three different Bayesian methods demonstrated (see the legend), cover the younger region for sample 2, and subsequent also for sample 1. As the 95.4% range of the single sample calibration of sample 2 does not cover the region of the younger peak, this region is lost with conventional reasoning (at a 95.4% confidence level).

For sure the example given here is just an arbitrary case, but one can imagine that this kind of differences will also occur in more complex examples in analogous ways.

6.2 TWO EXAMPLES CLOSE TO REAL APPLICATIONS

It should be noted initially, that in the following a BC-age scales will be used for the real sample ages in general, as the subsequent examples are more realistic. The BP-scale was advantageous, because it was well compatible with the mathematical description of the ages. However, in practical use the BC/AD scale is common.

6.2.1 Sequencing within the Hallstatt period

This example is a realistic simulation of the problem that is denoted as 'spread out' artefact in section 6.1.4., although still artificial, since there are no real measurements used.

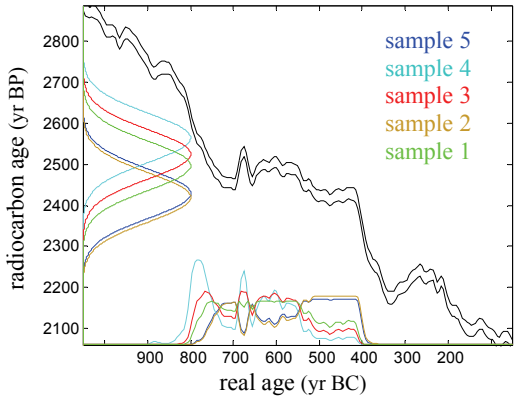

Figure 6.11: Radiocarbon ages and single-sample likelihood functions of a sequence of five samples. The example simulates radiocarbon measurements of samples that all originate from the early part of the plateau region, with real ages between 760 BC and 690 BC. The temporal order of the samples is assumed to be known, showing a sequence with sample 1 youngest and sample 5 oldest.

As commonly known, there is an age period that is a great challenge for radiocarbon dating, which is the range between 800 BC and 400 BC, because the calibration curve is roughly flat within this region (see Figure 6.11). In Europe, this part of the calibration curve is frequently called the Hallstatt plateau, because it covers roughly the time period of the Hallstatt culture.

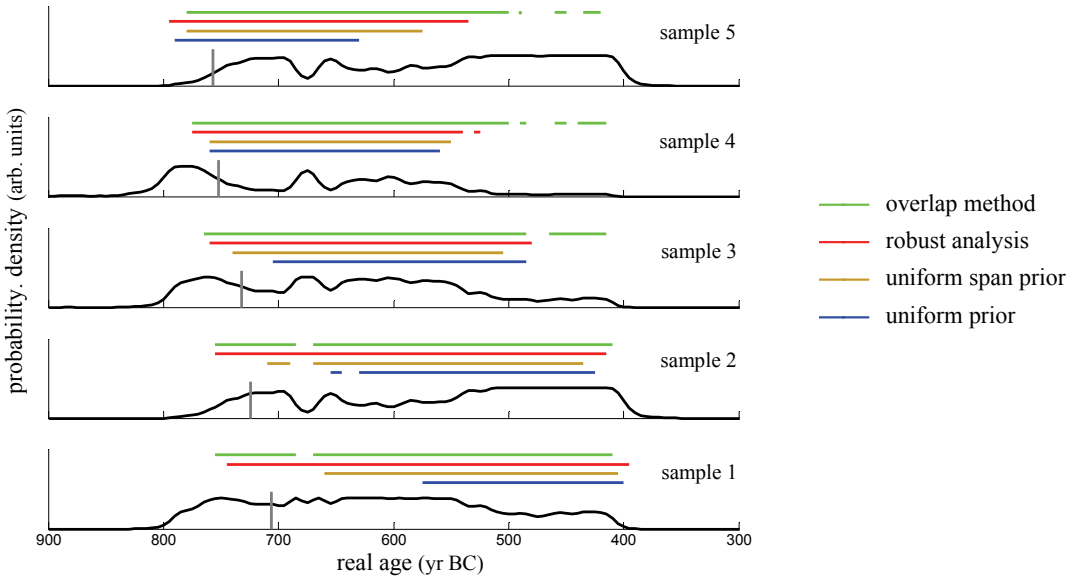

Figure 6.12: Resulting hpd-ranges at a 95.4% confidence level from four different methods as listed in the legend. The probability densities in black are the single-sample likelihood functions of the five samples. The grey bars show the real ages of the samples, on which the simulation is based.

Let us assume that there is a Hallstatt-culture village excavated and one finds organic material that can be associated with different rebuilding phases of a building or of an

other structure within the village, so that the samples can be ordered as a sequence. In total the building shall have been used between 760 BC and 690 BC and there where five samples excavated with the following unknown real ages: 706 BC, 724 BC, 732 BC, 752 BC, 757 BC (these values where found by drawing five values from a uniform distribution over the time range mentioned above, which were subsequently and ordered). Assuming an accuracy of the radiocarbon measurements taken from these samples of 60 yr, the measured radiocarbon ages according to the sequence of samples could for example be the following: 2496 BP, 2419 BP, 2524 BP, 2564 BP, 2428 BP (these values scatter statistically around the former, based on a Gaussian distribution with σ=60 yr). The radiocarbon ages and the according single sample calibrations are shown in Figure 6.11.
The used priors are equal to those used within the example in section 6.1.4, but now the prior set for robust analysis includes increasing exponential functions too, since the methods with free prior parameters are not considered furthermore, and the possibility of direct comparison of robust analysis with them was the reason to skip increasing functions. The used values for the parameter α are now: 5, 10, 20, 50, 100, 200, 500, 1000000, -500, -200, -100, -50, -20, -10 and -5 yr. (The last three values result in priors that are identified as corrupt and suppressed by the method).

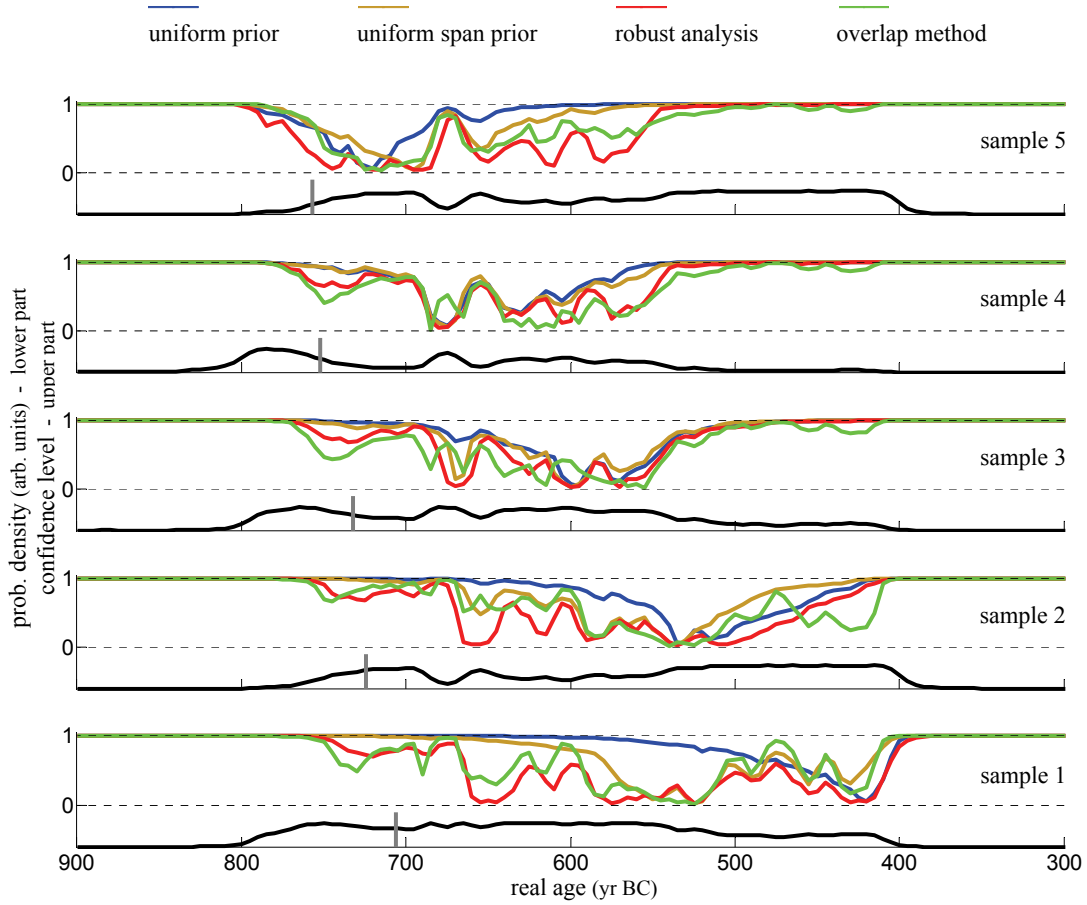

Figure 6.13: Resulting hpd-ranges envelopes (showing the hpd-ranges to any confidence level) from four different methods as listed in the legend. The probability densities in black are the single-sample likelihood functions of the five samples. The grey bars show the real ages of the samples, on which the simulation is based.

The results from robust analysis and additionally from the overlap method are given in comparison with results produced by the use of the uniform prior and the uniform span prior by both, Figure 6.12 that shows the hpd-ranges at 95.4%, and by Figure 6.13 that shows the hpd-ranges envelopes. The real sample ages on which the simulation is based are indicated by the grey bars. The problem called 'spread out' in the artificial example above, can be seen in this more realistic example as well, most clearly at the results of the uniform prior: Although the real sample ages lie all within a relative short time period at the beginning of the plateau region of the calibration curve, the resulting from Bayesian sequencing are spread out from young to old over the whole plateau region. The results for e.g. the youngest sample is clearly inconsistent with its real age, not only for the uniform prior but also for the uniform span prior. Contrary, the results of robust Bayesian sequencing, and as well this of the overlap method, are in full agreement with the actual set of real sample ages.

Additionally an earlier discussion of the described example can be found by STEIER *et al.* (2001).

6.2.2 The Iceman and his axe

As widely known (and already mentioned in the section 5.4.3), a very well preserved body of an Early Bonze Age man was found in the Alps (Ötztaler Alpen) in the year 1991. The 'Iceman' (also called 'Ötzi' according to the place of his finding) was released by a melting ice shield within a small basin. In the surrounding of the body numerous artefacts were found, and many of them could be identified as parts of his equipment. The body itself was radiocarbon dated by the accelerator mass spectroscopy laboratories in Oxford (HEDGES *et al.*, 1996) and Zürich (BONANI *et al.*, 1994) using bone and tissue samples. A combined value based on these measurements (taken from KUTSCHERA and ROM, 2000) is associated with the time of death of the Iceman in the following. The fact that the radiocarbon age of bone material (collagen) can be somewhat higher than that of tissue (which corresponds more or less directly to the time of death), caused by a slow reformation of the bones within the living body, is neglected. Many samples from the equipment of the Iceman were dated here at the Vienna Environmental Research Accelerator (ROM *et al.*, 1999; KUTSCHERA and MÜLLER, 2003). For the current example the radiocarbon age of the wooden shaft of an axe used by the Iceman is analysed together with the age of the man itself. See the measurements and the corresponding single-sample likelihood functions in Figure 6.14.

It is obvious that the real age of the wooden axe shaft can not be younger than the man, since the wood had to be cut before the Iceman died. The most simplest way to formulate this fact is the **uniform prior**:

$$a(\mathbf{t}) \;\propto\; \chi\,(t_{axe} > t_{man})$$

For sure, more detailed investigations could offer additional information that could make a more complex prior shape reasonable: For example, if the actually dated piece of wood would originate from inner tree rings, it would show an older

radiocarbon age than this related to the time when the wood was cut. This information could arise from detailed analysis of the sample. There could also exist archaeological evidences, that axes of this kind are usually made from inner parts of a trunk or even from old wood that was cut longer ago. This facts would make very short time differences between the two dates unlikely, what could be considered within the prior function. However, there is no definite information of that kind available for the current example.

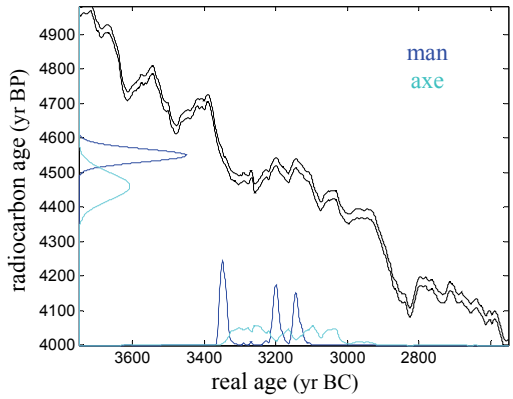

Figure 6.14: Determined radiocarbon ages of the Iceman (4550±19 yr BP) and of the wooden shaft of his axe (4460±40 yr BP), together with the corresponding single-sample likelihood functions. The calibration curve is shown by its one-sigma accuracy band.

Since the current case can be seen as a sequence of just two samples, one can ask for the uniform span prior. It turns out, that in case of only two samples the uniform span prior is identical with the uniform prior, because the additional prior factor becomes:

$$(t_n\text{-}t_1)^{-(n-2)} \;=\; (t_{axe}\text{-}t_{man})^{-(2-2)} \;=\; 1$$

That means, in this case the uniform prior offers already a constant prior probability for any age difference with correct temporal order. Thus the uniform prior is a reasonable choice when using just a single prior, because an equal probability for any age difference is a meaningful assumption, when the only available information is the temporal order of the samples. (The fact that very high age differences are unlikely for sure, has no meaning, because the prior shape is only relevant in the region of the likelihood function.) However, as discussed frequently in this thesis, the choice of a particular prior function remains arbitrary. One could as well prefer a $1/(t_{axe}\text{-}t_{man})$ prior, which would be suggested by the considerations at the rear part of section 5.2.5, where this prior results as effective prior of a parametric set of exponential priors with 'balanced' parameter scale. (Exponential priors are distinguished on their part by the maximum entropy method; see section 3.4.)

An alternative prior that is actually shown above is based on an extension of the model using two other boundaries (b_{old}, b_{young}) and applying a **'limiting uniform span prior'** for the span between these two boundaries.

$$a(\mathbf{t}) \;\propto\; \chi\,(b_{old} > t_{axe} > t_{man} > b_{young}) \;\cdot\; (b_{old} - b_{young})^{-2}$$

This prior is considered, because the use of boundaries to realise a uniform span prior for a sequence is a frequent method, applied for example within the OxCal program.

The used **prior set for robust Bayesian analysis** consists again of decreasing and increasing exponential functions as in previous examples:

$$a(\mathbf{t})_\alpha \;\propto\; \chi\,(t_{axe} > t_{man}) \;\cdot\; \exp\,(-(t_{axe}-t_{man})/\alpha)$$

And finally the **overlap method** bases directly on the prior constraint that is simply:

$$t_{axe} > t_{man}$$

Figure 6.15 and Figure 6.16 show the results in the two different representation already known from above. Additionally Figure 6.17 shows all individual marginal posteriors for robust sequencing, both the accepted and the discarded. Contrary to previous examples, only increasing exponential functions are discarded here, because even very steep decreasing functions, which suggest small time differences between the two ages, are in agreement with the data, due to the high degree of overlap of the two single sample likelihoods.
...

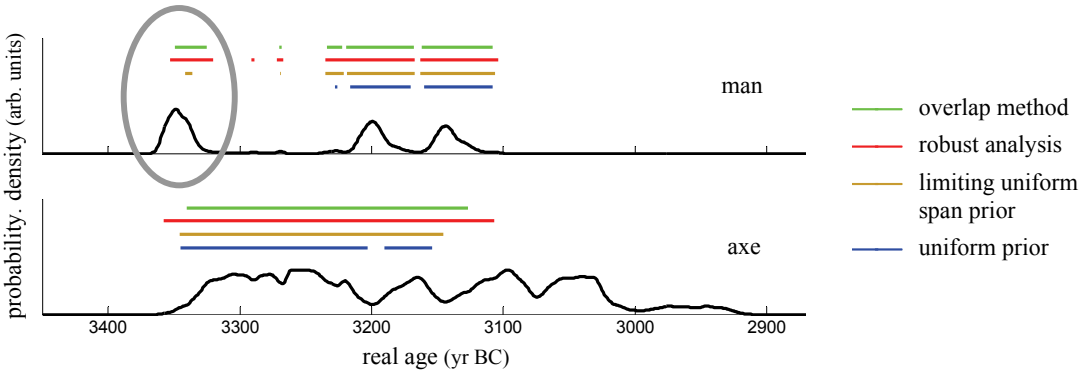

Figure 6.15: Resulting hpd-ranges at a 95.4% confidence level for the Iceman example. The probability densities in black are the single-sample likelihood functions. The differences between the methods can be seen most significantly at the marked peak.

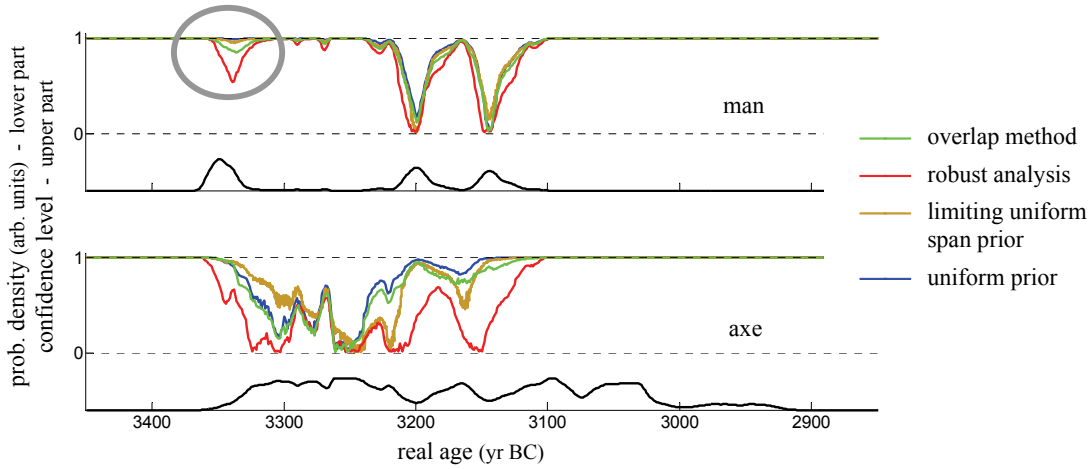

Figure 6.16: Resulting hpd-ranges envelopes (showing the hpd-ranges at any confidence level) for the Iceman example. The probability densities in black are the single-sample likelihood functions. The differences between the methods can be seen most significantly at the marked peak.

It turns out that for the Iceman's age the youngest peak of the single sample calibration (see the grey cycle within Figure 6.15) is clearly included within the 95.4% hpd-range by robust analysis, what is not the case for the uniform prior. This is a serious discrepancy, when keeping in mind that the uniform prior is identical with the uniform span prior in that case, and thus it is a reliable choice. So the Bayesian method in its common form, suppresses this peak, because it ignores the possibility that short age differences between axe and man could be much more likely than longer. This is very well considered by robust Bayesian analysis, and thus the peak is included. Using the 'limited uniform span prior' is more or less an artificial procedure to favour short time differenced between the ages of axe and man. It includes just a bit of the discussed peak within the 95.4% hpd-range. When looking at the hpd-ranges envelops (Figure 6.16), it can be seen that the robust sequencing includes the peak clearly. To include the peak has to be claimed in case of this example, because the single-sample likelihood functions overlap still considerably at the region of the peak, and as the age difference between axe and man could be just a view years, both axe and man could well be from this time.

So both, the current and the previous example show, that the use of a particular prior, although representing a reasonable model, can lead to incorrect restrictions of the possible ranges for the real sample ages. Robust sequencing seems to be a possible way to overcome this problem in many cases.

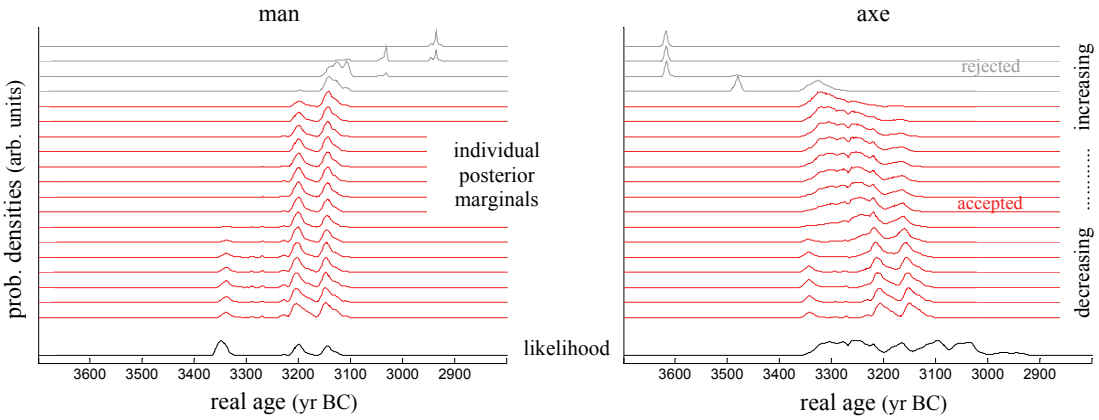

Figure 6.17: Collection of all individual posterior marginals corresponding to the different prior functions within the prior set used for robust sequencing. The priors vary from strongly decreasing exponential functions at the bottom to strongly increasing ones at the top. The marginals at the top plotted in grey correspond to very steep increasing exponential prior functions, and thus are strongly biased towards regions outside the main parts of the single sample likelihoods. However, the method detects the disagreement of these prior function with the measurements and discards them.

Naturally, the given examples can just illustrate the characteristics of robust Bayesian analysis in some aspects. There can never be a prove that the method (that suffers still on some unsolved question; see section 5.4.2) works properly in any case.

6.3 A LARGE REAL-WORLD SEQUENCE: THE AEGINA KOLONNA SITE

In this final section an archaeological site on an Aegean Island is analysed conventionally and with robust analysis. The excavation contributes to a research program of the Austrian Academy of Science called 'Synchronisation of Civilisations in the Eastern Mediterranean in the 2nd Millennium BC - SCIEM 2000' (BIETAK, 2000 and 2003; BIETAK and CZERNY, 2007). Bayesian sequencing is a very important tool to find synchronous or even absolute time scales for different cultures, and the investigations within this thesis to improve the sequencing technique were motivated amongst others by the mentioned program.

The small Island Aegina, which is located about 30 km South-West from Athens in the Aegean Sea, bears the Aegina Kollonna excavation site, which shows the remains of an important Bronze Age settlement. The Kolonna site can be chronologically linked by ceramic findings to distant regions from mainland Greece to Crete. Information to the site can be found by GAUß and SMETANA (2007).

6.3.1 Stratigraphic knowledge and radiocarbon measurements

No.	^{14}C-age (yr BP)		Laboratory identifier of the sample(s)
s_{02}	3727.9	± 13.5	VERA-2687, -2680, -2681, -2679, -2682, -2683, HV-5841, -5840, VRI-0395
s_{03}	3759	± 35	VERA-4641
s_{04}	3698	± 33	VERA-2688
s_{06}	3704	± 36	VERA-2692
s_{08}	3800	± 44	VERA-4640
s_{09}	3809	± 32	VERA-4639
s_{10}	3646	± 32	VERA-4638
s_{11}	3740	± 36	VERA-4281
s_{12}	3711	± 34	VERA-4282
s_{13}	3780	± 37	VERA-4283
s_{15}	3643	± 30	VERA-4637
s_{16}	3628	± 30	VERA-4636
s_{17}	3724	± 39	VERA-4280
s_{18}	3718	± 38	VERA-4279
s_{19}	3694	± 35	VERA-2687
s_{21}	3544	± 37	VERA-4634
s_{22}	3522	± 38	VERA-4278
s_{23}	3513.1	± 12.8	VERA-4038, -4576, -4575, -4578, -4579, -4580, -4276, -4275
s_{24}	3458	± 39	VERA-4577
s_{26}	3462.7	± 21.7	VERA-4571, -4574, -4573
s_{27}	3407	± 38	VERA-4572
s_{28}	3428	± 36	VERA-4570
s_{30}	3333	± 29	VERA-4633
s_{31}	3356	± 36	VERA-4632
s_{32}	3349	± 36	VERA-4631
s_{34}	3313	± 48	VERA-4630
s_{37}	3021.2	± 20.1	VERA-4284, -4582, -4285

Table 6.1: List of the used radiocarbon dates given with their 1σ measurement accuracy. s_x are the assigned parameter numbers used in the Bayesian model. The radiocarbon dates at s_{02}, s_{23}, s_{26} and s_{37} are weighted means of the listed individual samples, which are stratigraphically associated with particular layers of short duration. The measurements and the original sequencing of the site were performed by WILD *et al.* (2010).

The sequencing analysed here is mainly based on stratigraphies and related radiocarbon measurements obtained from recent excavations (GAUß and SMETANA, 2007; WILD *et al.*, 2010). Large coherent stratigraphies could be found, which are prominently based on a sequence of floors of a large building that was repeatedly reconstructed during many centuries. Additionally the chronology is characterises by ceramic phases, which connect separated stratigraphies. Figure 6.18 shows the stratigraphic situation already in this simplified way, which defines the model used for the Bayesian sequencing.

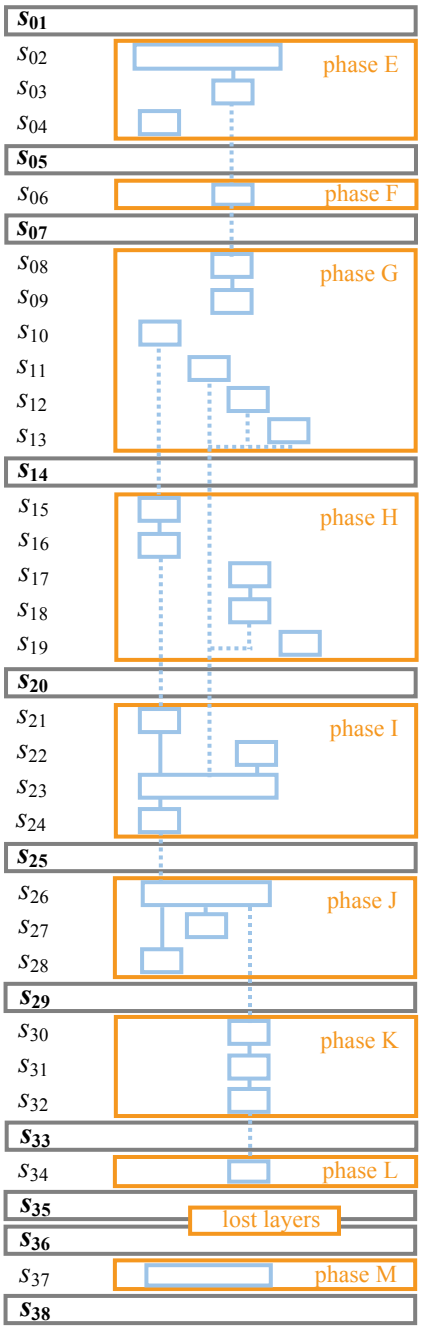

Figure 6.18: Simplified illustration of the stratigraphic situation of the Aegina Kolonna site. The orange rectangles show ceramic phases (identified by different ceramic styles). The grey rectangles symbolise the boundaries between the phases. The samples are drawn in blue, where the wider rectangles symbolise groups of samples with a single common real age, e.g. samples within a particular fire destruction layer. The vertical blue lines connecting individual samples or sample groups symbolise time sequences that can be directly deduced from a particular coherent stratigraphy (full and doted lines have different impact on the Bayesian model explained below). The denotation of the samples and boundaries shows already the parameters used in the model. Note that there is a Hiatus between the phases L and M, caused by destruction of layers by subsequent construction activity.

The ceramic phases (E to M), the sequence is divided in, are symbolised by the orange rectangles. They are separated within the model by phase boundaries symbolised by grey rectangles. The blue rectangles symbolise individual samples or groups of samples, which originate from a thin clearly defined layer and are associated with a common real age (expanded rectangles). For example the layer within phase E (s_{02}) defines a fire destruction layer containing six individual radiocarbon samples. The blue vertical lines connecting individual samples or sample groups symbolise time sequences that can be directly deduced from a particular coherent stratigraphy (see the difference between full and dotted lines in the next section).
Table 6.1 shows the used individual measured radiocarbon dates, already assigned to the parameter number s_x of the Bayesian model, which are also listed in Figure 6.18. ('*s*' is chosen to be consistent with basic descriptions in section 2.6.) It should be mentioned, that the given list shows just the selection of radiocarbon dates used finally in WILD *et al.*, 2010. The process of discarding improper samples shall not be discussed here. In recent times the question how to treat outliers has been analysed very seriously; see e.g. BRONK RAMSEY, 2009*b*. However, systematic outlier analysis is beyond the scope of this thesis.

6.3.2 Particular model definitions

The basic constraints defining the Bayesian model can be directly deduced from Figure 6.18. The uniform prior based on these constraints is shown below, allready in Matlab notation as used for the progam input (see section 2.4.2 for explanation of the Matlab syntax and section 2.6 for the principles to create models with phase boundaries):

```
Prior  =   (S(01) > S(02))                    * ...
                  (S(02)  >  S(03))           * ...
                             (S(03) > S(05)) * ...
           (S(01) > S(04)) * (S(04) > S(05)) * ...
           (S(05) > S(06)) * (S(06) > S(07)) * ...
           (S(07) > S(08))                    * ...
                  (S(08)  >  S(09))           * ...
                             (S(09) > S(14)) * ...
           (S(07) > S(10)) * (S(10) > S(14)) * ...
           (S(07) > S(11)) * (S(11) > S(14)) * ...
           (S(07) > S(12)) * (S(12) > S(14)) * ...
           (S(07) > S(13)) * (S(13) > S(14)) * ...
           (S(14) > S(15))                    * ...
                  (S(15)  >  S(16))           * ...
                             (S(16) > S(20)) * ...
           (S(14) > S(17))                    * ...
                  (S(17)  >  S(18))           * ...
                             (S(18) > S(20)) * ...
           (S(14) > S(19)) * (S(19) > S(20)) * ...
           (S(20) > S(21))                    * ...
                  (S(21)  >  S(23))           * ...
           (S(20) > S(22))                    * ...
                  (S(22)  >  S(23))           * ...
                  (S(23)  >  S(24))           * ...
                             (S(24) > S(25)) * ...
           (S(25) > S(26))                    * ...
                  (S(26)  >  S(27))           * ...
                             (S(27) > S(29)) * ...
                  (S(26)  >  S(28))           * ...
                             (S(28) > S(29)) * ...
           (S(29) > S(30))                    * ...
                  (S(30)  >  S(31))           * ...
                  (S(31)  >  S(32))           * ...
                             (S(32) > S(33)) * ...
           (S(33) > S(34)) * (S(34) > S(35)) * ...
           (S(35)          >          S(36)) * ...
           (S(36) > S(37)) * (S(37) > S(38));
```

This prior function reflects the sequence of phases and boundaries, as well as the additional information from the direct stratigraphic relations, shown by the blue vertical lines in Figure 6.18. For the latter, just the relations indicated by the full lines have to be included explicitly within the expression; the relations symbolised by dotted lines are already considered just by the phase structure.
To get the expression for the uniform overall-span prior (see section 3.3 for explanation) the expression for the uniform prior is divided by the following term:

```
(S(01) - S(05)) ^3  * ...
(S(05) - S(07)) ^1  * ...
(S(07) - S(14)) ^6  * ...
(S(14) - S(20)) ^5  * ...
(S(20) - S(25)) ^4  * ...
(S(25) - S(29)) ^3  * ...
(S(29) - S(33)) ^3  * ...
(S(33) - S(35)) ^1  * ...
(S(36) - S(38)) ^1  * ...
(S(01) - S(38)) ^9;
```

The factors express the phase lengths to the power of the number of the samples within the individual phases. The last factor is the overall length to the power of the number of boundaries reduced by 2. (When using this prior actually within the program, there is an additional term that provides divisions by zero when two boundaries are set on the same age within the sampling process.) It should be mentioned that the definition of the uniform overall-span prior in this way differs slightly from the pure concept, because it ignores the fact that there are additional time relations between samples within the same phase.
The overlap method (see section 5.5) which is shown for this example too, is based directly on the basic constraints as already defined for the uniform prior.
For robust analysis the length of each phase is modelled with exponentially decreasing or increasing probabilities with varying slopes. The reason to focus especially on exponential functions is discussed in the rear part of section 6.1.1. To get a prior set of that kind, the uniform prior is multiplied by the following expression:

```
exp (-(S(01) - S(05)) / Y(Z, 1))  *  ...
exp (-(S(05) - S(07)) / Y(Z, 2))  *  ...
exp (-(S(07) - S(14)) / Y(Z, 3))  *  ...
exp (-(S(14) - S(20)) / Y(Z, 4))  *  ...
exp (-(S(20) - S(25)) / Y(Z, 5))  *  ...
exp (-(S(25) - S(29)) / Y(Z, 6))  *  ...
exp (-(S(29) - S(33)) / Y(Z, 7))  *  ...
exp (-(S(33) - S(35)) / Y(Z, 8))  *  ...
exp (-(S(35) - S(36)) / Y(Z, 9))  *  ...
exp (-(S(36) - S(38)) / Y(Z,10));
```

Where each row of the matrix `Y` contains a particular combination of individual slopes of the exponential contributions, modelling the lengths of the ten different phases. Within the calculation process the index `Z` steps through all rows of the matrix. The particular used matrix `Y` is the following:

using:

```
U = 1000000;
S1 = 20;   L1 = -20;
S2 = 10;   L2 = -10;
S3 =  5;   L3 =  -5;
```

Y = [

```
 U  U  U  U  U  U  U  U  U  U
S1  U  U  U  U  U  U  U  U  U
 U S1  U  U  U  U  U  U  U  U
 U  U S1  U  U  U  U  U  U  U
 U  U  U S1  U  U  U  U  U  U
 U  U  U  U S1  U  U  U  U  U
 U  U  U  U  U S1  U  U  U  U
 U  U  U  U  U  U S1  U  U  U
 U  U  U  U  U  U  U S1  U  U
 U  U  U  U  U  U  U  U S1  U
 U  U  U  U  U  U  U  U  U S1
L1  U  U  U  U  U  U  U  U  U
 U L1  U  U  U  U  U  U  U  U
 U  U L1  U  U  U  U  U  U  U
 U  U  U L1  U  U  U  U  U  U
 U  U  U  U L1  U  U  U  U  U
 U  U  U  U  U L1  U  U  U  U
 U  U  U  U  U  U L1  U  U  U
 U  U  U  U  U  U  U L1  U  U
 U  U  U  U  U  U  U  U L1  U
 U  U  U  U  U  U  U  U  U L1
S1 L1  U  U  U  U  U  U  U  U
 U S1 L1  U  U  U  U  U  U  U
 U  U S1 L1  U  U  U  U  U  U
 U  U  U S1 L1  U  U  U  U  U
 U  U  U  U S1 L1  U  U  U  U
 U  U  U  U  U S1 L1  U  U  U
```

(... continued aside)

```
 U  U  U  U  U  U S1 L1  U  U
 U  U  U  U  U  U  U S1 L1  U
 U  U  U  U  U  U  U  U S1 L1
L1 S1  U  U  U  U  U  U  U  U
 U L1 S1  U  U  U  U  U  U  U
 U  U L1 S1  U  U  U  U  U  U
 U  U  U L1 S1  U  U  U  U  U
 U  U  U  U L1 S1  U  U  U  U
 U  U  U  U  U L1 S1  U  U  U
 U  U  U  U  U  U L1 S1  U  U
 U  U  U  U  U  U  U L1 S1  U
 U  U  U  U  U  U  U  U L1 S1
L1 L1 S1 S1 S1 S1  U  U  U  U
L1 L1 L1 S1 S1 S1 S1  U  U  U
L1 L1 L1 L1 S1 S1 S1 S1  U  U
 U L1 L1 L1 L1 S1 S1 S1 S1  U
 U  U L1 L1 L1 L1 S1 S1 S1 S1
 U  U  U L1 L1 L1 L1 S1 S1 S1
 U  U  U  U L1 L1 L1 L1 S1 S1
S1 S1 L1 L1 L1 L1  U  U  U  U
S1 S1 S1 L1 L1 L1 L1  U  U  U
S1 S1 S1 S1 L1 L1 L1 L1  U  U
 U S1 S1 S1 S1 L1 L1 L1 L1  U
 U  U S1 S1 S1 S1 L1 L1 L1 L1
 U  U  U S1 S1 S1 S1 L1 L1 L1
 U  U  U  U S1 S1 S1 S1 L1 L1
```

(... continued aside)

```
S2  U  U  U  U  U  U  U  U  U
 U S2  U  U  U  U  U  U  U  U
 U  U S2  U  U  U  U  U  U  U
 U  U  U S2  U  U  U  U  U  U
 U  U  U  U S2  U  U  U  U  U
 U  U  U  U  U S2  U  U  U  U
 U  U  U  U  U  U S2  U  U  U
 U  U  U  U  U  U  U S2  U  U
 U  U  U  U  U  U  U  U S2  U
 U  U  U  U  U  U  U  U  U S2
L2  U  U  U  U  U  U  U  U  U
 U L2  U  U  U  U  U  U  U  U
 U  U L2  U  U  U  U  U  U  U
 U  U  U L2  U  U  U  U  U  U
 U  U  U  U L2  U  U  U  U  U
```

... the matrix is continued for S2/L2 and S3/L3 similarly to the part for S1/L1

Each row of the matrix defines an individual multi-dimensional prior with a particular combination of slopes for the exponential-function factors describing the lengths of the individual phases. The constants `U`, `S1`, `L1`, `S2`, `L2`, `S3`, `L3` define the different used slopes, where '`S`' (short) denotes decreasing and '`L`' (long) increasing functions, '`U`' (uniform) denotes a constant factor (using an approximate realisation of e^0). The shown definition of the prior set contains 158 different prior shapes, since the uniform overall-span prior defined above, has been also added to the set.

As already discussed generally in section 5.3.2, the used prior set is always a more or less incomplete approximation of the theoretical set with its infinite number of possible prior shapes. However, one can see by some tests, that the used set is an acceptable approximation. As the sequence of phases and boundaries defines the main structure of the prior knowledge, it is reliable to focus on the variation of the lengths of the phases. The adequate choice of the different used slopes for the exponential functions is based on the two following aspects: Once one can see for the discussed case, that exponential-function factors with low slopes, produced by components of `Y` that have considerably higher absolute values than 20 yr, act already similar to the uniform prior (for both, decreasing and increasing functions). On the other hand there is an evidence that the prior set ranges sufficiently towards high slopes, characterising priors with a high potential to shift the posterior function strongly: This is the fact that a significant number of priors is discarded because of their disagreement with the measurements. In the recent example 58 of the 158 priors are discarded (based on the method used for applications with high dimensionality, as explained in the last paragraph of section 5.3.1). Finally, one has to combine different possible slopes associated to the individual phases. It is obvious that there is no way to realise all of the 5^{10} possible combinations. Fortunately, there is no need to consider combinations of slopes of distant phases, because they act approximately independent. For example, it is sufficient that each of the two outer phases is associates with each slope once; there is no need to realise all 25 combinations. Of

course one has to realise various combination of slopes for phases that lie chronologically close together. Primarily it is important to realise 'extreme' combinations, as e.g. the neighbouring of a phase with a decreasing and a phase with an increasing exponential-function factor, as e.g. in the 22th and the following rows of Y.

6.3.3 Results of the Bayesian sequencing

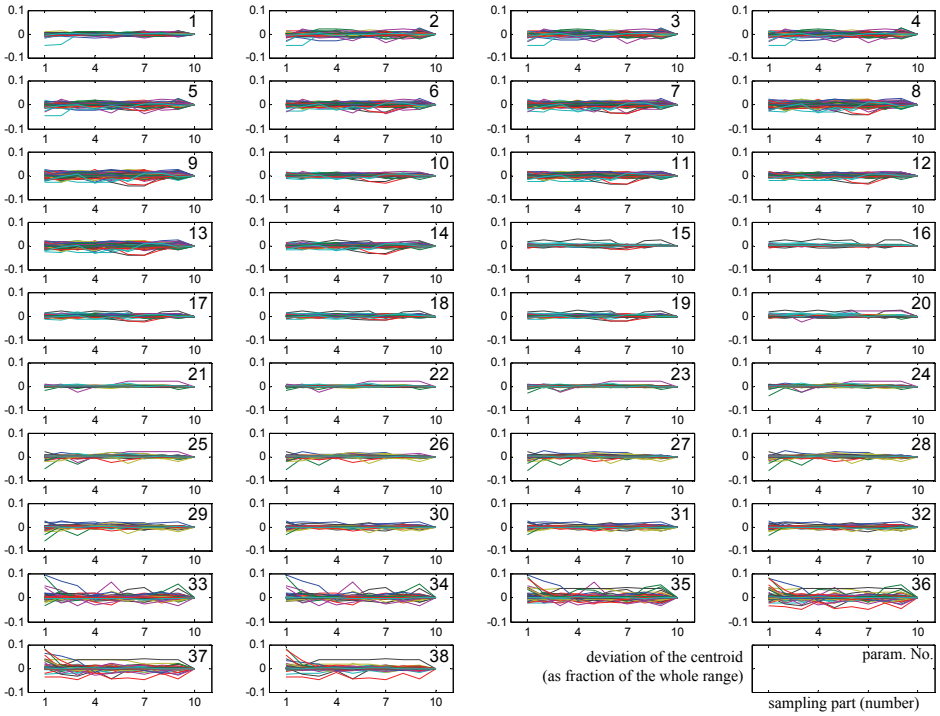

Figure 6.19: Illustration of the convergence behaviour for the robust Bayesian analysis of the Aegina Kolonna sequence. For the posterior marginals for each of the 38 model parameters the deviations (as fraction of the whole age range) of the centroids of the distributions are plotted by a polygon; referring to the value for the last of the ten partitions of the sampling process. For each parameter there are 158 polygons, each related to a different prior function plotted one upon each other. Good convergence is characterised by a narrow band around zero deviation, built by these polygons. Since the whole age range of the calculation was 1700 yr, the relative deviation of 0.1 corresponds to 170 yr.

Before discussing the sequencing results, it should be mentioned that for sequences of this size, containing 38 model parameters (sample ages and boundaries), the convergence of the calculation is not a matter of course any more. As mentioned in section 2.5.2 the calculation program developed for this thesis was not optimised for best convergence, just the basic Gibbs sampling mode is used. Thus one has to check for convergence carefully. Figure 6.19 gives an illustrative visualisation of the convergence behaviour for the robust analysis run, which is the most critical, because the need of many partial calculation using various priors does not allow excessive long runtimes. For each of the 38 parameters the movements of the centroids of their

posterior marginals during the sampling process (which is divided into ten parts) are shown, and this for all calculations performed with any of the various prior functions used within the prior set (polygons in various colours). (The deviation is always defined relative to the position of the centroid after the last calculation part.) Figure 6.19 shows, that for the posterior marginals of most of the model parameters, and also in case of almost all prior functions used, the fluctuation have amplitudes of only few percent of the overall time range used for the calculation. To get an acceptable convergence, each sampling procedure for the various priors was started at a starting point deduced from the posterior marginals calculated previously with the uniform prior. However, the plots for the parameters 33 to 38 show that the convergence is not perfect, the results have not jet reached the final level of equilibrium for some prior function shapes (for some of the different polygons).
Altogether the level of convergence achieved within the used runtime is sufficient. The whole calculation for robust Bayesian analysis lasted about 180 hours or about one hour for each prior function on a common personal computer. However, this could be reduced by orders of magnitudes by using better sampling methods, and additionally by optimising the program code with respect to runtime; but this is outside the focus of this thesis.

Let us look now actually on the sequencing results: Figure 6.20 shows the highest posterior density ranges at 95.4% confidence level, both for the samples and the phase boundaries. For all samples the single-sample likelihood function (single sample calibration) is additionally shown for comparison. The hpd-ranges given are calculated with the uniform prior (blue), with the uniform overall-span prior (ochre), with robust analysis (red) and additionally with the overlap method as an approximation of the latter (green).
All in all one can see that robust Bayesian analysis and the overlap method extend the hpd ranges drastically compared with the ranges produced by the uniform prior or the uniform overall-span prior. For example, if one looks at the sample with parameter number s_{08}, there are much older ages possible within the 95% hpd-range with robust analysis than with the conventional uniform span prior. Figure 6.21 shows additionally the hpd-ranges envelopes for this specific sample, where one can see again the enlargement of the ranges to any confidence level. By the way, another fact can be seen in case of this sample and generally: There is a significant difference between robust analysis and the overlap method. This is not surprising since the two methods have jet similarities in there goals but realise these in very different ways (see sections 5.1 and 5.5). Thus, the overlap method is only a very rough approximation of robust analysis.
Although robust Bayesian analysis extend the hpd-ranges significantly, the fundamental profit of Bayesian sequencing remains clearly even in this real-world application, and not just in a very simple case as shown in section 6.1.1. This can be seen by the fact that there are significant parts of the single sample calibrations (likelihood functions) discarded, as e.g. for the discussed sample s_{08} and many others (see again Figure 6.20). However, the price that has to be paid for a secure result is the loss of the drastic restriction of the posterior ranges to very small time spans, as resulting from the conventional method. Of course, this small time ranges are the result of the arbitrary choice of a particular model, and thus artificial (see section 3.1).

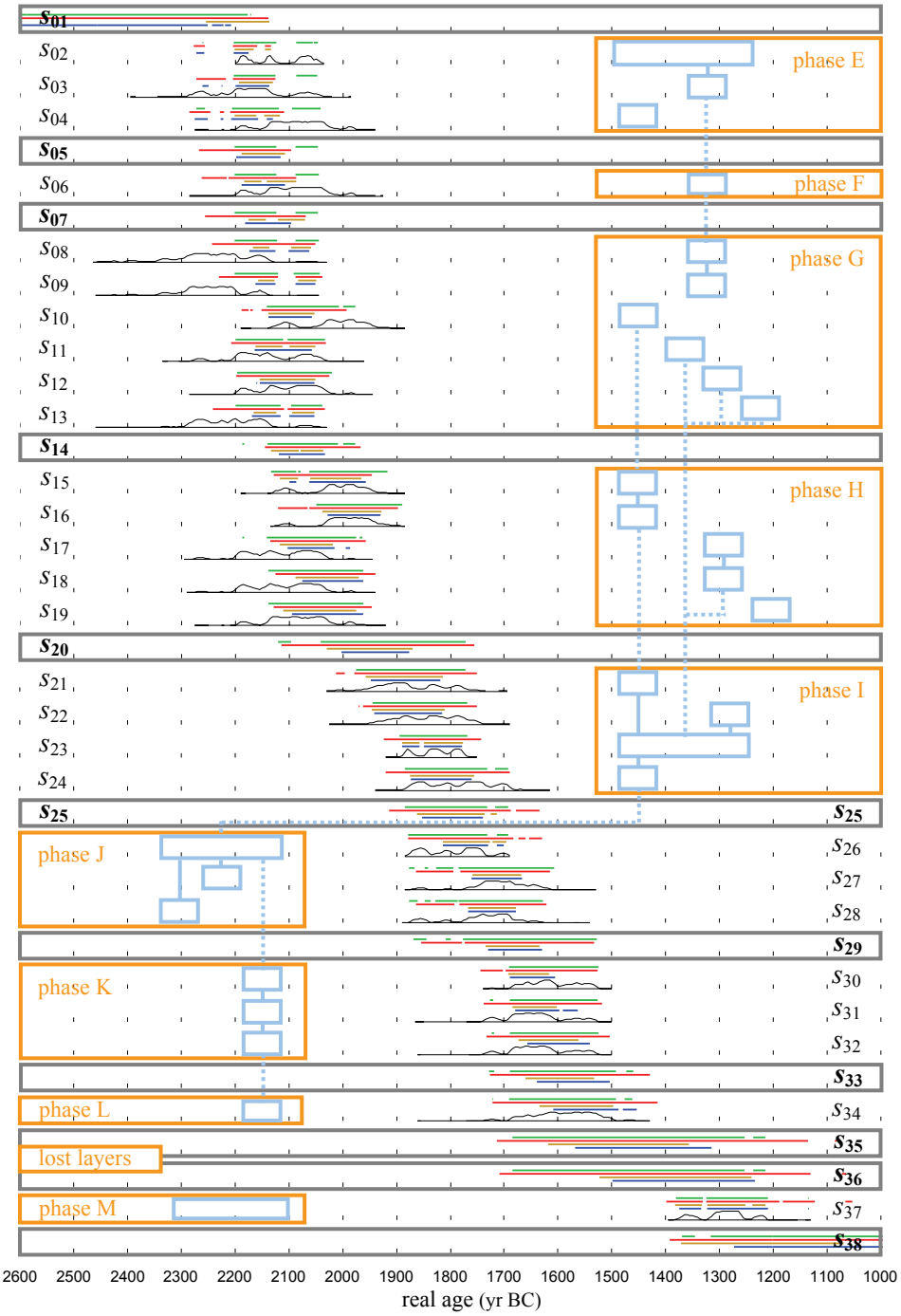

Figure 6.20: Sequencing results of the Aegina Kolonna site. The grey rectangles indicate boundaries between the ceramic phases (orange). Additional relations are shown in blue. The black curves give the single-sample likelihood functions of the samples. The coloured lines show the hpd-ranges at 95.4% confidence level, resulting from the following methods: using the uniform prior (blue); using the uniform overall-span prior (ochre); robust Bayesian analysis (red); overlap method (green).

On the other hand, one must not forget that the used prior set includes extreme prior function shapes (very steep increasing and decreasing probabilities for the phase lengths), because they are only restricted by the basic constraints. For sure, if one

would try to include really all available relevant archaeological information, maybe one could find additional restrictions for the shape of the prior functions (e.g. that the probabilities for the lengths of the phases have to decrease beyond e.g. 100 yr) that would exclude some priors, and thus shorten the resulting time ranges. It is clear, that the definition of declarations of that kind would be a challenging project, and will often not be possible in an objective way. In general one can understand a fundamental characteristic of robust analysis: If the prior knowledge is specified weakly, the method will utilise this high degree of freedom, and the posterior ranges will become very wide. Thus, one is forced to specify the prior knowledge as complete and accurate as possible to restrict the possible shapes of the prior function, leading to shorter ranges.

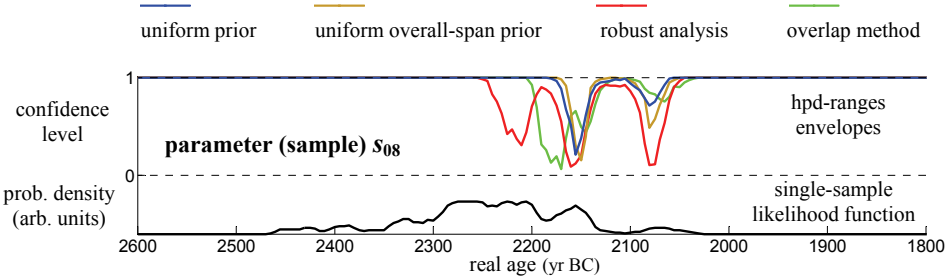

Figure 6.21: Exemplarily, the hpd-ranges envelopes are shown together with the single-sample likelihood function for the sample with parameter number s_{08} . Typically for the sequence in general, robust analysis shows significantly larger age ranges than the conventional method using the uniform overall span prior. The 'overlap method' approximates the robust analysis just very roughly.

Furthermore the sequence shows, that the difference in the length of the hpd-ranges between robust analysis and the conventional method (using the uniform overall-span prior) is most significant in case of the phase boundaries. This is caused by the fact, that model parameters as the phase boundaries are not so closely related to a likelihood function as the samples are, and thus they are more influenced by the prior shape. That means, the resulting posterior ranges are especially for phase boundaries highly dependent on the chosen prior shape, when working with a particular single prior.
To get an impression of the individual shapes of the resulting marginal posterior functions corresponding to different prior function used within robust analysis, Figure 6.22 shows a selection of posterior marginals for the repeatedly mentioned sample s_{08}. For clarity not all 158 marginals are shown; just one of each group of nearly equivalent marginals is shown. (These groups consist of priors that differ in a way, which does not affect sample s_{08}). When viewing this figure, one has to keep in mind that the posterior marginal corresponding to rejected priors need not to be bad in case of the shown sample. The disagreement with the measurements can origin from another sample, because any disagreement results in a rejection of the corresponding prior function as a whole.

In this thesis the sequence of Aegina Kolonna is used just to demonstrate the differences between robust analysis and conventional sequencing on a real word example. Naturally there arise interesting archaeological results from this sequence, as e.g. a contribution to the fixation of an absolute chronology of the Aegean Bronze

Age phases, which are discussed by WILD *et al.*, 2010 in detail. It would not be serious to discuss the influence of robust Bayesian analysis on these archaeological interpretations without completing the prior knowledge by accurate investigations of archaeologists as mentioned above. However, the result from robust analysis as given here, which is based just on the most evident prior knowledge, would not alter the given interpretations in principle, although the precision of the conclusions would be significantly lower.

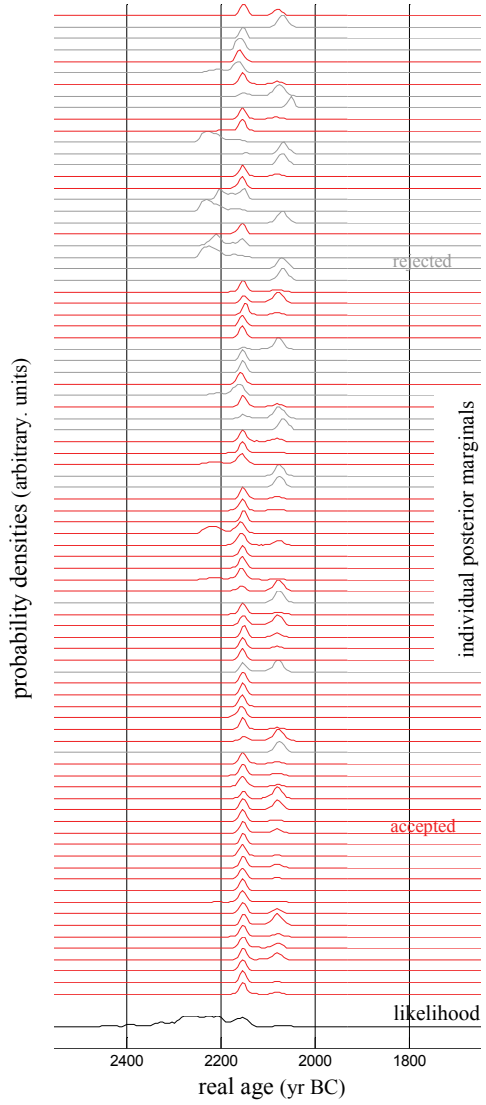

Figure 6.22: Illustration of the various posterior marginals corresponding to the various functions within the prior set used for robust Bayesian analysis. Again the example is given for sample s_{08}. Note, that just a selection of the 158 different marginals is shown; redundant functions are skipped. The prior functions leading to the marginals plotted in grey are rejected due to disagreements with the measurements. (Note that the disagreement concerns the prior as a whole and needs not to be visible in the marginals of this plot, that show only these of sample s_{08}).

7 CONCLUSION

The main task of this thesis was to study whether it is possible to eliminate subjectivity in the Bayesian sequencing procedure of radiocarbon dates. The subjectivity occurs from the fact that usually the available prior information can not be unambiguously be transformed to a particular prior function. There are various function shapes possible that lead consequently to different resulting posterior functions too. Each particular prior function carries more information than actually available and thus, sequencing with a particular prior function restricts the result erroneously.
The main approach given here was to perform the Bayesian sequencing with the whole entity of all possible prior functions, which is a specific application of robust Bayesian analysis, a known method in Bayesian statistics. All in all, the performed investigations indicate that the use of robust analysis for the Bayesian sequencing of radiocarbon dates is worth to be considered as an alternative to the common single prior method.

However, the analysis of the actual possible realisations of this improved sequencing method identified some points which leave possibilities for improvement.
It turned out that there are prior functions that destroy a reliable result, although they are consistent with the available prior information. Thus it is necessary to find a method to discard such corrupt functions from the prior set when performing robust Bayesian analysis. The used criterion to do this is based on the agreement of the particular prior with the measured data. An agreement measure has do be defined, as well as a reliable threshold level. This can be realised in a reliable way, although it causes the method to be not perfectly unambiguous any more, contrarily to the theoretical idea.
A second principle problem is the fact, that a theoretically infinite number of various possible prior functions has to be realised actually by a finite set of functions. An alternative concept is the use of a parametric prior function, where the parameters alter the shape of the prior function and are treated as additional model parameters within the sampling process. This approach allows the realisation of an infinite number of different shaped prior functions, although, even there not all possible shapes can be modelled. Unfortunately it turns out, that this method - which is a methodical realisation of the so called 'model averaging' - is equivalent to the use of a particular prior function again. The latter can be seen as effective prior that is realised by the use of the parametric prior function. Thus, to keep the fundamental idea of robust analysis, it is necessary to work with a finite set of discrete prior functions, which can just be an approximation for the infinite number of possible function shapes for sure. Fortunately, for many usual applications it is possible to estimate a reliable prior set by some basic considerations. However, this process has to be done manually jet, which is not a finally satisfying solution.
Concluding, there are serious questions which make further investigation necessary, but it seems possible to enhance robust Bayesian sequencing to a completed and easy applicable method. On the other hand, also the usefulness of related methods as the

parametric-prior method just mentioned above, or the 'overlap method' also introduced earlier, should be investigated further, even though they differ fundamentally from robust analysis performed with a set of discrete priors.

All in all it has been illustrated, that the fact that robust Bayesian analysis considers all possible various prior shapes that are consistent with the prior information, instead of using just one particular function, although the most reasonable one, generally reduces the precision of the result significantly, or enlarges the resulting highest posterior density ranges in other words. This is the price one has to pay for producing results with a preferably high reliability that are not affected by the arbitrary choice of a particular prior function. So robust Bayesian analysis forces the user to look carefully for all available prior information, because any additional information included into the model will restrict the set of possible prior functions, and thus, increase the precision of the result.
However, at the end it remains a subjective decision whether to use robust Bayesian analysis or to stay by a carefully chosen particular prior. It is the decision whether to focus on a high reliability or on a high precision, although paid by a higher risk of failure.

REFERENCES

ADAM G and HITTMAIR O. (1992). *Wärmetheorie.* Vieweg, Braunschweig. (ISBN 3-528-33311-1)

BAYES T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London* 53, 370-481.

BAYARRI MJ, CASTELLANOS ME and MORALES J. (2006). MCMC methods to approximate conditional predictive distributions. *Computational Statistics & Data Analysis* 51, 621-640.

BERGER JO. (1994). An overview of robust Bayesian analysis. *Test* 3(1), 5-124.

BERGER JO. (2000). Bayesian Analysis: A Look at Today and Thoughts of Tomorrow. *Journal of the American Statistical Association* 95(452), 1269-1276.

BERGER JO. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis* 1(3), 385-402.

BERGER JO and PERICCHI LR. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association* 91(433), 109-122.

BIETAK M (Ed.). (2000). *The Synchronisation of Civilisations in the Eastern Mediterranean in the Second Millennium B.C., Proceedings of an International Symposium at Schloß Haindorf, 15th-17th of November 1996 and at the Austrian Academy, Vienna, 11th-12th of May 1998*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

BIETAK M (Ed.). (2003). *The Synchronisation of Civilisations in the Eastern Mediterranean in the Second Millennium B.C. II, Proceedings of the SCIEM 2000 - EuroConference, Haindorf, 2nd-7th of May 2001*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

BIETAK M and CZERNY E (Eds.). (2007). *The Synchronisation of Civilisations in the Eastern Mediterranean in the Second Millennium B.C. III, Proceedings of the SCIEM 2000 - 2nd EuroConference, Vienna, 28th of May - 1st of June 2003*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

BLAAUW M, BAKKER R, CHRISTEN JA, HALL VA and VAN DER PLICHT J. (2007). A Bayesian Framework for Age Modeling of Radiocarbon-Dated Peat Deposits: Case Studies from the Netherlands. *Radiocarbon* 49, 357-367.

BONANI G, IVY S, HAJDAS I, NIKLAUS TR and SUTER M. (1994). AMS ^{14}C age determinations of tissue, bone and grass samples from the Ötztal Ice Man. *Radiocarbon* 36(2), 247-250.

BOWMAN S. (1990). *Radiocarbon Dating - (Interpreting the past)*. British Museum Publications, London. (ISBN 0-7141-2047-2)

BRONK RAMSEY C. (1995). Radiocarbon calibration and analysis of stratigraphy: the OxCal program. *Radiocarbon* 37(2), 425-430.

BRONK RAMSEY C. (2000). Comment on 'The use of Bayesian statistics for ^{14}C dates of chronologically ordered samples: a critical analysis'. *Radiocarbon* 42(2), 199-202.

BRONK RAMSEY C. (2001*a*). Development of the radiocarbon calibration program. *Radiocarbon* 43(2A), 355-363.

BRONK RAMSEY C. (2008). Deposition models for chronological records. *Quaternary Science Reviews* 27, 42-60.

BRONK RAMSEY C. (2009*a*). Bayesian analysis of radiocarbon dates. *Radiocarbon* 51(1), 337-360.

BRONK RAMSEY C. (2009*b*). Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon* 51(3), 1023-1045.

BRONK RAMSEY C, VAN DER PLICHT J and WENINGER B. (2001*b*). 'Wiggle Matching' Radiocarbon Dates. *Radiocarbon* 43(2A), 381-389.

BRONK RAMSEY C, DEE MW, ROWLAND JM, HIGHAM TFG, HARRIS SA, BROCK F, QUILES A, WILD EM, MARCUS ES and SHORTLAND AJ. (2010). Radiocarbon-Based Chronology for Dynastic Egypt. *Science* 328, 1554-1557.

BUCK CE, KENWORTHY JB, LITTON CD and SMITH AFM. (1991). Combining archaeological and radiocarbon information: a Bayesian approach to calibration. *Antiquity* 65, 808-821.

BUCK CE, LITTON CD and SMITH AFM. (1992). Calibration of radiocarbon results pertaining to related archaeological results. *Journal of Archaeological Science* 19, 497-512.

BUCK CE, CAVANAGH WG and LITTON CD. (1996). *Bayesian Approach to Interpreting Archaeological Data*. John Wiley & Sons Ltd, Chichester, England. (ISBN 0-471-96197-3)

CHEN M-H and SCHMEISER BW. (1996). General Hit-and-Run Monte Carlo sampling for evaluating multidimensional integrals. *Operation Research Letters* 19, 161-169.

CHRISTEN JA. (1994). Summarizing a set of radiocarbon calibrations: a robust approach. *Applied Statistics* 43(3), 489-503.

DEHLING H and VAN DER PLICHT J. (1993). Statistical problems in calibrating radiocarbon dates. *Radiocarbon* 35(1), 239-244.

GALIMBERTI M, BRONK RAMSEY C and MANNING SW. (2004). Wiggle-match dating of tree-ring sequences. *Radiocarbon* 46(2), 917-924.

GARCIA-DONATO G, and CHEN M-H. (2005). Calibrating Bayes factor under prior predictive distributions. *Statistica Sinica* 15, 359-380.

GAUß W and SMETANA R. (2007). Early and Middle Bronze Age Stratigraphy and Pottery from Aegina Kolonna. In: BIETAK M and CZERNY E (Eds.), *The Synchronisation of Civilisations in the Eastern Mediterranean in the Second Millennium B.C. III, Proceedings of the SCIEM 2000 - 2nd EuroConference, Vienna, 28th of May - 1st of June 2003*, 451-472. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

GELMAN A. (2006). The boxer, the Wrestler, and the Coin Flip: A Paradox of Robust Bayesian Inference and Belief Functions. *American Statistician* 60(2), 146-150.

GILKS WR, RICHARDSION S and SPIEGELHALTER DJ (editors). (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London. (ISBN 0-412-05551-1)

GODWIN H. (1962). Half-life of Radiocarbon. *Nature* 195, 984.

GRECO L, RACUGNO W and VENTURA L. (2008). Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference* 138, 1258-1270.

GUILDERSON TP, REIMER PJ and BROWN TA. (2005). The Boon and Bane of Radiocarbon Dating. *Science* 307, 362-364.

HEDGES REM, HOUSLEY RA, BRONK CR, VAN KLINKEN GJ. (1992). Radiocarbon dates from the Oxford AMS system: Archaeometry datelist 15. *Archaeometry* 34(2), 337-357.

HOETING JA, MADIGAN D, RAFTERY AE and VOLINSKY CT. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science* 14(4), 382-401.

JEFFREYS H. (1961). *Theory of Probability*. Oxford University Press, Amen House, London.

KASS RE and WASSERMAN L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* 91(435), 1343-1370.

KRAUSE A. (1994). *Computerintesive statistische Methoden: Gibbs sampling in Regressionsmodellen*. Gustav Fischer Verlag, Stuttgart. (ISBN 3-437-50372-3)

KUTSCHERA W and ROM W. (2000). Ötzi, the prehistoric Iceman. *Nuclear Instruments and Methods in Physics Research B* 164-165, 12-22.

KUTSCHERA W and MÜLLER W. (2003). "Isotope language" of the Alpine Iceman investigated with AMS and MS. *Nuclear Instruments and Methods in Physics Research B* 204, 705-719.

LEVINE RA, ZHAOXIA J, HANLEY WG and NITAO JJ. (2005). Implementing componentwise Hastings algorithms. *Computational Statistics & Data Analysis* 48, 363-389.

LIBBY WF. (1952). *Radiocarbon Dating*. The University of Chicago Press, Chicaco.

MANNING SW, BRONK RAMSEY C, KUTSCHERA W, HIGHAM T, KROMER B, STEIER P and WILD EM. (2006). Chronology for the Aegean Late Bronze Age 1700-1400 B.C. *Science* 312, 565-569.

MOOK WG and VAN DER PLICHT J. (1999). Reporting 14C activities and conentrations. *Radiocarbon* 41(3), 227-239.

MOSKOWITZ B and CAFLISCH RE. (1996). Smoothness and Dimension Reduction in Quasi-Monte Carlo Methods. *Mathematical and Computer Modelling* 23(8/9), 37-54.

NICHOLLS G and JONES M. (2001). Radiocarbon dating with temporal order constraints. *Applied Statistics* 50(4), 503-521.

OGATA Y. (1989). A Monte Carlo methode for high dimensional integration. *Numerische Mathematik* 55, 137-157.

O'NEILL. (2009). Importance sampling for Bayesian sensitivity analysis. *International Journal of Approximate Reasoning* 50, 270-278.

PEREZ CJ, MARTIN J, and RUFO MJ. (2006). MCMC-based local parametric sensitivity estimations. *Computational Statistics & Data Analysis* 51, 823-835.

PETER LEPAGE G. (1978). A New Algorithm for Adaptive Multidimensional Integration. *Journal of Computational Physics* 27, 192-203.

PRESS WH, TEUKLOSKY SA, VETTERLING WT and FLANNERY BP. (1992). *Numerical Recipes in C, The Art of Scientific Computing,* 2nd edition. Cambridge University Press, New York. (ISBN 0-521-43108-5)

REIMER PJ, BAILLIE MGL, BARD E, BAYLISS A, BECK JW, BERTRAND CJH, BLACKWELL PG, BUCK CE, BURR GS, CUTLER KB, DAMON PE, EDWARDS RL, FAIRBANKS RG, FRIEDRICH M, GUILDERSON TP, HOGG AG, HUGHEN KA, KROMER B, MCCORMAC G, MANNING S, RAMSEY C BRONK, REIMER RW, REMMELE S, SOUTHON JR, STUIVER M, TALAMO S, TAYLOR FW, VAN DER PLICHT J, WEYHENMEYER CE. (2004). IntCal04 Terrestrial Radiocarbon Age Calibration, 0-26 cal kyr BP. *Radiocarbon* 46(3), 1029-1058.

REIMER PJ, BAILLIE MGL, BARD E, BAYLISS A, BECK JW, BLACKWELL PG, BRONK RAMSEY C, BUCK CE, BURR GS, EDWARDS RL, FRIEDRICH M, GROOTES PM, GUILDERSON TP, HAJDAS I, HAETON TJ, HOGG AG, HUGHEN KA, KAISER KF, KROMER B, MCCORMAC FG, MANNING SW, REIMER RW, RICHARDS DA, SOUTHON JR, TALAMO S, TURNEY CSM, VAN DER PLICHT J, WEYHENMEYER CE. (2009). IntCal09 and Marine09 Radiocarbon Age Calibration Curves, 0-50,000 years cal BP. *Radiocarbon* 51(4), 1111-1150.

RIOS INSUA D, RUGGERI F. (2000). *Robust Bayesian Analysis*. Springer, New York. (ISBN 978-0-387-98866-5)

ROM W, GOLSER R, KUTSCHERA W, PRILLER A, STEIER P and WILD EM. (1999). AMS ^{14}C Dating of equipment from the Iceman and of spruce logs from the prehistoric salt mines of Hallstatt. *Radiocarbon* 41(2), 183-197.

SIVAGANESAN S. (1999). A likelihood based robust Bayesian summary. *Statistics & Probability Letters* 43, 5-12.

SIVIA DS. (1996). *Data Analysis - A Bayesian Tutorial*. Oxford University Press Inc., New York. (ISBN 0-19-851762-9)

SOBOL IM. (1998). On quasi-Monte Carlo integrations. *Mathematics and Computers in Simulation* 47, 103-112.

STEIER P and ROM W. (2000). The use of Bayesian statistics for ^{14}C dates of chronologically ordered samples: a critical analysis. *Radiocarbon* 42(2), 183-198.

STEIER P, ROM W and PUCHEGGER S. (2001). New methods and critical aspects in Bayesian mathematics for ^{14}C calibration. *Radiocarbon* 43(2A), 373-380.

TAKHTAMYSHEV G, VANDEWOESTYNE B and RONALD C. (2007). Quasi-random integration in high dimensions. *Mathematics and Computers in Simulation* 73, 309-319.

WASSERMAN L. (1996). The conflict between improper priors and robustness. *Journal of Statistical Planning and Inference* 52, 1-15.

WENINGER F, STEIER P, KUTSCHERA W, and WILD EM. (2006). The Principle of the Bayesian Method. *Egypt and the Levant* 16, 317-324.

WENINGER F, STEIER P, KUTSCHERA W, and WILD EM. (2010). Robust Bayesian analysis, an attempt to improve Bayesian sequencing. *Radiocarbon* 52(2-3), 962-983.

WILD EM, GAUß W, FORSTENPOINTNER G, LINDBLOM M, SMETANA R, STEIER P, THANHEISER U and WENINGER F. (2010). ^{14}C dating of the Early to late Bronze Age stratigraphic sequence of Aegina Kolonna, Greece. *Nuclear Instruments and Methods in Physics Research B* 268, 1013-1021.

ACKNOWLEDGEMENTS

First of all a want to thank my advisor emer.O.Univ.-Prof. Dr. Walter Kutschera and my co-advisor Ass.-Prof. Mag. Dr. Peter Steier for their careful and intensive support of my work. There was so much to learn from Univ.-Prof. Walter Kutschera, both about the detailed principles of radiocarbon dating, as well as about many specific archaeological applications of the method. Ass.-Prof. Peter Steier introduced me to the mathematical foundation of this work and supported extensively the mathematical investigations by elaborate and inspiring discussions. I want to thank both advisors not just for their competent supervision, but also for their friendliness and for so many very pleasant conversations.
Further I want to thank emer.O.Univ.-Prof. Dr. Walter Kutschera and also Ao.Univ.-Prof. Mag. Dr. Eva Maria Wild for giving me the opportunity to participate in important radiocarbon dating projects, and for discussing so many interesting methodological and archaeological aspects.

Thanks to the research project of the Austrian Academy of Science 'SCIEM2000 - Synchronisation of Civilisations in the Eastern Mediterranean in the Second Millennium B.C.' with its speakers emer.O.Univ.-Prof. Dr. Manfred Bietak and emer.O.Univ.-Prof. Dr. Walter Kutschera and its project managers Dr. Angela Schwab and Dagmar Melman, MA, MSc for the opportunity to participate in the project and for founding my works, and also for the possibility to visit ongoing excavations in Egypt. Particular thanks to Mag. Dr. Walter Gauß, Mag. Dr. Felix Höflmayer, Mag. Dr. Bettina Bader and Mag. Barbara Jettmar for many interesting discussions and/or cooperations.

Further thanks to Ao.Univ.-Prof. Dipl.-Ing. Dr. Robin Golser, the present speaker of the isotope-research group at the faculty for physics at which the works for this thesis were performed, and also to Univ.-Prof. i.R. Dr. Peter Hille, my former advisor of my diploma thesis, who introduced me to the research group. Particular thanks to Dipl.-Ing. Dr. Oliver Forstner for answering many computer-system related questions and to Helga Vincro for here patient help in administrative issues.

In general, I want to thank actually all people at the isotope research group, at the closely related nuclear physics group, and in the SCIEM2000 project for receiving me so kindly. Special thanks to my roommates and colleges just working on, or having finished their PhD or diploma thesis for many nice and pleasant conversations - not just on physics - that I enjoyed so much; particularly to (listed without academic degree) Matthias Auer, Kerstin Rumpelmayr, Leonard Michlmayr, Roswitha Avalos Ortiz, Petra Milota, Kathrin Buczak, Daniel Imrich, Jenny Feige, Jakob Liebl, Franz Dellinger, Lea Reichart, without excluding these not mentioned.

I thank my dear parents for supporting my diploma studies in the eighties and thus enabling this thesis too; and for so much more.
Finally I thank my wife Maria for her love, most important in my life.

CURRICULUM VITAE

Franz Weninger

born:	on October 20th, 1962, in Neunkirchen (Lower Austria)
son of:	Elfriede und Franz Weninger
married with:	Ing. Mag. Maria Fellner, since July 1997
address:	Goethestraße 1/1/9, A-2620 Neunkirchen, and Steingasse 18/2/6, A-1030 Vienna
1969 - 1973:	attendance at elementary school
1973 - 1977:	attendance at the grammar school (lower stage) at Neunkirchen
1977 - 1982:	attendance at the higher school for technical education at Wiener Neustadt (Lower Austria), electrotechnical branch
1982 - 1983:	military service
1983 - 1985:	developing power-electronic circuits at the company 'Schrack Elektronik AG' in Vienna
1985 - 1992:	physics studies at the University of Vienna, diploma thesis in the field of thermoluminescence dating
1992 - 1993:	working in the field of software development at 'Siemens AG' in Vienna
1993 - 1997:	working in the field of medical physics with special focus on dosimetry in radio-diagnostics at the department for bio-medical engineering and physics at the University of Vienna
1998 - 2002:	working as teacher for physics at the higher school for technical education at Wiener Neustadt
2002 - 2003:	participation at the 'Med-Austron' project - a planed novel radiation device for cancer treatment - at the company 'Forschungs- und Technologietransfer GmbH' in Wiener Neustadt
2004:	search for an interesting topic and preliminary investigations for my PhD thesis
2005 - 2011:	working on my PhD thesis; the investigations have been linked with a participation at the research project of the Austrian Academy of Science 'SCIEM2000 - Synchronisation of Civilisations in the Eastern Mediterranean in the Second Millennium B.C.', until the end of the project in February 2011

Vienna, June 2011